CONSTRAINED VARIATIONAL POLICY OPTIMIZATION FOR HYBRID ACTION SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Constrained reinforcement learning in hybrid action spaces must coordinate discrete modes and mode-conditioned continuous controls under long-horizon cost constraints. Existing constrained hybrid-action methods typically optimize factorized policies directly in parameter space, whereas Constrained Variational Policy Optimization (CVPO) separates nonparametric improvement from parametric projection in its original continuous-action formulation. We introduce Hybrid Constrained Variational Policy Optimization (H-CVPO), extending this variational framework to hybrid action spaces. Under boundedness and strict feasibility, its constrained E-step has zero duality gap and, when the dual optimum is attained at positive temperature, yields a closed-form joint target distribution over hybrid actions through exponential reward–cost reweighting. Using the exact hybrid KL decomposition, we formulate a factorized M-step that fits both policy components to this target under separate discrete and conditional-continuous trust regions. For exact E/M updates, we prove fixed-reference surrogate improvement and conditional reward–cost bounds with a sufficient feasibility-margin condition. For approximate updates, we establish the existence of a dual minimizer under continuity on a compact domain and use residual-based admissibility to determine whether to project the target. Ablations identify cost-aware reweighting as an important contributor to constraint control. Across four tasks and three cost limits, H-CVPO achieves favorable reward–cost trade-offs compared with hybrid-action baselines.

1 INTRODUCTION

Many real-world control systems are inherently hybrid: each decision couples a discrete operating mode with a mode-conditioned continuous control action. Hybrid electric vehicles illustrate this structure: a multi-mode powertrain selects an electric or engine-assisted mode and continuously allocates power between the engine and electric machines (Zhuang et al., 2017; Fan et al., 2021). Similar structures arise in power systems, where unit commitment determines generator on/off states and economic dispatch sets continuous generation levels (Montero et al., 2022; Swarup & Simi, 2006); in steel-plant oxygen supply, where production scheduling combines equipment start-up and shutdown decisions with continuous oxygen output (Li et al., 2023); in autonomous driving, where decisions to keep or change lanes are coupled with continuous longitudinal acceleration (Lin et al., 2024); and in robotic assembly, where selecting between coarse and fine motion control can be coupled with continuous velocity commands (Neunert et al., 2020). Such problems are neither purely discrete nor purely continuous and require coordinated optimization of mode selection and mode-conditioned control under operational constraints.

Constrained reinforcement learning (CRL) is commonly formalized as a constrained Markov decision process (CMDP), which maximizes long-horizon performance subject to cumulative cost limits (Altman, 1999). Representative methods enforce these constraints through primal–dual optimization (Tessler et al., 2019), trust-region constrained updates (Achiam et al., 2017), projection-based methods (Yang et al., 2020), or Lyapunov-based approaches (Chow et al., 2019). These methods address constrained policy optimization but do not explicitly account for factorized, mode-conditioned hybrid actions or coordinate their discrete and continuous updates. A complementary perspective on policy optimization comes from reinforcement learning as probabilistic inference (Levine, 2018). Maximum a Posteriori Policy Optimization (MPO) constructs an improved nonparametric action distribution and projects it onto a parametric policy through an expectation–maximization (EM) pro-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

cedure (Abdolmaleki et al., 2018). Constrained Variational Policy Optimization (CVPO) incorporates constraints through a nonparametric E-step and a KL-controlled M-step, connecting constrained policy optimization and probabilistic inference (Liu et al., 2022).

Extending this variational framework to hybrid control, however, requires a joint nonparametric target over discrete modes and mode-conditioned continuous controls. Projecting this target onto a factorized policy requires divergence controls that account for the dependence between mode selection and mode-conditioned continuous control. Existing work also addresses constrained hybrid control directly. In particular, CHPO formulates constrained parameterized-action Markov decision processes and optimizes factorized hybrid policies through feasibility-dependent objective switching (Zhou et al., 2025). Its updates operate in parametric policy space rather than through a joint nonparametric target, and it does not provide separate divergence controls for discrete-mode and continuous-action updates.

Building on the constrained variational framework, we develop Hybrid Constrained Variational Policy Optimization (H-CVPO) for hybrid action spaces. Its constrained E-step constructs a joint nonparametric target, and its M-step projects this target onto a factorized policy under separate discrete and conditional-continuous KL constraints.

Our contributions are summarized as follows:

- We formulate constrained variational policy improvement for hybrid action spaces, extending the continuous-action formulation to a joint nonparametric target that couples discrete mode selection with mode-conditioned continuous control.
- We derive the resulting hybrid E/M update: under stated assumptions, the constrained E-step has strong duality and, for attained positive-temperature dual optima, a closed-form reward–cost-reweighted target, while an exact hybrid KL decomposition yields a factorized M-step with separate discrete and conditional-continuous trust regions.
- We establish fixed-reference surrogate improvement for exact E/M updates and conditional reward–cost bounds with a sufficient feasibility margin. For approximate updates, we prove existence of a dual minimizer under continuity on a compact domain and introduce residual-based target-admissibility conditions.
- We compare H-CVPO with hybrid-action baselines across four tasks and three cost limits under matched interaction budgets, and use component-wise ablations to examine the effects of cost-aware reweighting, separate hybrid KL controls, and target-admissibility filtering on reward and cost.

2 Related Work

H-CVPO builds on variational policy optimization to address constrained reinforcement learning in hybrid action spaces.

Hybrid-action reinforcement learning. Hybrid-action methods couple a discrete choice with associated continuous parameters. Q-PAMDP learns action-selection and parameter-selection policies (Masson et al., 2016); PADDPG embeds parameterized actions in deterministic actor–critic learning (Hausknecht & Stone, 2016); P-DQN combines value-based discrete selection with continuous parameter optimization (Xiong et al., 2018); and HPPO separates discrete and continuous policy branches (Fan et al., 2019). Hybrid SAC further extends entropy-regularized actor–critic learning to hybrid actions (Delalleau et al., 2019). These methods do not explicitly impose cumulative cost constraints.

Constrained reinforcement learning. Within the CMDP framework, RCPO and PID Lagrangian methods handle constraints through multiplier updates (Tessler et al., 2019; Stooke et al., 2020), whereas CPO and PCPO use trust-region and projection-based updates (Achiam et al., 2017; Yang et al., 2020). Their formulations do not explicitly address the coupling between discrete mode selection and mode-conditioned continuous control.

Variational policy optimization. Probabilistic-inference formulations derive policy-improvement updates from inference objectives (Levine, 2018). MPO alternates nonparametric action-distribution

improvement with parametric policy projection (Abdolmaleki et al., 2018). CVPO incorporates cumulative constraints into this E/M procedure through a constraint-aware E-step and a KL-controlled M-step (Liu et al., 2022). CVPO does not explicitly formulate a joint hybrid-action target or its projection onto a factorized policy with separate discrete and conditional-continuous KL constraints.

Constrained hybrid-action reinforcement learning. CHPO directly optimizes a factorized hybrid policy in parametric space and, under the assumptions of its analysis, derives finite-iteration bounds on the policy-objective gap and the excess of expected cost over its threshold (Zhou et al., 2025). Its update remains within the parametric policy space and does not introduce a joint nonparametric target followed by policy projection. Following CHPO (Zhou et al., 2025), we use HPPO-Lag and PDQN-Rco as empirical baselines, augmenting HPPO and P-DQN with Lagrangian and reward-constrained updates, respectively; we make no additional convergence or feasibility claim for these baselines. H-CVPO instead performs joint nonparametric improvement before factorized policy projection.

3 PRELIMINARIES

3.1 CMDP FORMULATION

We consider a discounted CMDP (Altman, 1999), $\mathfrak{M} = (\mathcal{S}, \mathcal{A}, P, r, c, \rho_0, \gamma)$, with cumulative cost limit d. Here, $\mathcal{S}$ and $\mathcal{A}$ are standard Borel state and action spaces, P is a stochastic transition kernel from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$, ρ_0 is the initial-state distribution, and $\gamma \in (0, 1)$ is the discount factor. The measurable reward and cost functions $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ satisfy $|r(s, a)| \le R_{\max}$ and $0 \le c(s, a) \le C_{\max}$ for finite $R_{\max}, C_{\max}$. We restrict attention to stationary randomized Markov policies π, each a stochastic kernel from $\mathcal{S}$ to $\mathcal{A}$.

Under π, a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ is generated by $s_0 \sim \rho_0$, $a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. The expected discounted reward and cost returns are

$$J_r(\pi) = \mathbb{E}_{\tau\sim(\rho_0,P,\pi)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right], \qquad J_c(\pi) = \mathbb{E}_{\tau\sim(\rho_0,P,\pi)}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right]. \tag{1}$$

The constrained policy-optimization problem is

$$\max_{\pi}\ J_r(\pi) \qquad \text{s.t.} \qquad J_c(\pi) \le d. \tag{2}$$

A policy is feasible if $J_c(\pi) \le d$. We next specialize $\mathcal{A}$ to a hybrid action space.

3.2 HYBRID ACTION SPACE AND POLICY REPRESENTATION

For hybrid control, the generic CMDP action space is instantiated as $\mathcal{A} = \mathcal{H}$. A hybrid action $h = (m, a_c)$ consists of a discrete mode m and a mode-conditioned continuous action a_c. Let $\mathcal{M} = \{1, \ldots, M\}$ be the finite mode set and let $\mathcal{A}_c^m \subseteq \mathbb{R}^{p_m}$ be the Borel continuous action domain for mode m, with mode-dependent dimension $p_m \in \mathbb{N}_0$ and positive Lebesgue measure when $p_m > 0$. For $p_m = 0$, we set $\mathcal{A}_c^m = \mathbb{R}^0$, a singleton, and interpret da_c as its unit mass, thereby covering modes without a free continuous parameter. The hybrid action space is the disjoint union

$$\mathcal{H} = \bigsqcup_{m\in\mathcal{M}} (\{m\} \times \mathcal{A}_c^m). \tag{3}$$

For any nonnegative measurable function $\varphi : \mathcal{H} \to [0, \infty]$, define the sum–integral reference measure μ by

$$\int_{\mathcal{H}} \varphi(h)\, \mu(dh) = \sum_{m\in\mathcal{M}} \int_{\mathcal{A}_c^m} \varphi(m, a_c)\, da_c. \tag{4}$$

Thus, μ combines summation over modes with Lebesgue integration on each continuous action space, allowing its dimension and domain to depend on the mode. We restrict the hybrid action distributions considered below to those that are absolutely continuous with respect to μ. Their densities, and hence all KL divergences, are defined relative to μ.

Accordingly, a stochastic hybrid policy is represented by its μ-density

$$\pi_\theta(m, a_c \mid s) = \pi_{\theta_m}(m \mid s)\pi_{\theta_c}(a_c \mid s, m), \tag{5}$$

where $\theta = (\theta_m, \theta_c)$ collects the discrete- and continuous-policy parameters. For any measurable function $f : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$ that is integrable under $\pi_\theta(\cdot \mid s)$, its conditional expectation is

$$\mathbb{E}_{(m,a_c)\sim\pi_\theta(\cdot|s)}[f(s,m,a_c)] = \sum_{m\in\mathcal{M}} \pi_{\theta_m}(m \mid s) \int_{\mathcal{A}_c^m} f(s,m,a_c) \, \pi_{\theta_c}(a_c \mid s, m)\, da_c. \tag{6}$$

Although factorized, the continuous component remains conditioned on the selected mode, so the policy defines a joint distribution over hybrid actions.

When $\mathcal{A}_c^m = \mathcal{A}_c$ for every m, the hybrid action space reduces to $\mathcal{M} \times \mathcal{A}_c$. If $\pi_{\theta_c}(a_c \mid s, m) = \pi_{\theta_c}(a_c \mid s)$ as well, the policy further reduces to an independent discrete–continuous factorization. Thus, the formulation covers the general variable-domain case, the fixed-domain special case, and the further special case of independent discrete–continuous policies.

3.3 Variational Policy Improvement

Following MPO and CVPO (Abdolmaleki et al., 2018; Liu et al., 2022), variational policy optimization introduces an intermediate nonparametric action distribution between consecutive parametric policies. At iteration i, the E-step optimizes a target $q_i^\star(\cdot \mid s)$ while holding the current policy π_i fixed, and the M-step projects this target onto the policy class to obtain π_{i+1}. The resulting update is

$$\pi_i \longrightarrow q_i^\star \longrightarrow \pi_{i+1}.$$

This separation decouples target optimization from parametric policy fitting. For hybrid actions, however, $q_i^\star(\cdot \mid s)$ is a probability measure on $\mathcal{H}$ rather than a pair of independently optimized action distributions, so its construction and projection must preserve the mode-conditioned policy structure and the associated hybrid KL geometry.

4 Hybrid Constrained Variational Policy Optimization

We now formulate the preceding variational update for the hybrid CMDP. The constrained E-step constructs a joint nonparametric target $q_i^\star(m, a_c \mid s)$ on $\mathcal{H}$ relative to the current policy π_i, and the factorized M-step projects the same target onto a parameterized hybrid policy.

4.1 Hybrid-Action Variational E-Step

At iteration i, fix a reference probability measure d_i on $\mathcal{S}$, the current hybrid policy π_i, and measurable reward and cost action-value functions $Q_r^i, Q_c^i : \mathcal{S}\times\mathcal{H} \to \mathbb{R}$. We write $\pi_i(m, a_c \mid s)$ for the μ-density of $\pi_i(\cdot \mid s)$, and all four objects remain fixed throughout one E/M cycle. Let q be a measurable stochastic kernel from $\mathcal{S}$ to $\mathcal{H}$ such that $q(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-almost every (a.e.) state s, where $\ll$ denotes absolute continuity. The conditional KL divergence is $D_{\mathrm{KL}}(q\|\pi_i) = \int_{\mathcal{H}} \log(dq/d\pi_i)\, dq$, with the state conditioning suppressed; since $\pi_i(\cdot \mid s) \ll \mu$, every admissible $q(\cdot \mid s)$ also admits a μ-density. For an E-step cost threshold $\epsilon_{E,i} > 0$ and KL radius $\epsilon_{\mathrm{KL}} > 0$, the E-step solves

$$\begin{aligned} \max_q \quad & \mathbb{E}_{s\sim d_i,\, h\sim q(\cdot|s)} \left[Q_r^i(s,h)\right] \\ \text{s.t.} \quad & \mathbb{E}_{s\sim d_i,\, h\sim q(\cdot|s)} \left[Q_c^i(s,h)\right] \le \epsilon_{E,i}, \\ & \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) \le \epsilon_{\mathrm{KL}}. \end{aligned} \tag{7}$$

Here, $\epsilon_{E,i}$ bounds the fixed-reference surrogate cost and is distinct from the global CMDP limit d.

Assumption 1 (E-step regularity)**.** *There exist finite constants $M_r, M_c > 0$ such that, for $d_i(ds)\pi_i(dh \mid s)$-a.e. (s,h), $|Q_r^i(s,h)| \le M_r$ and $0 \le Q_c^i(s,h) \le M_c$. Moreover, there exists a measurable stochastic kernel $\bar{q}_i$ from $\mathcal{S}$ to $\mathcal{H}$ such that $\bar{q}_i(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-a.e. s, whose expected cost is strictly below $\epsilon_{E,i}$ and whose averaged KL divergence from π_i is strictly below ϵ_{KL}.*

The associated Lagrangian, with cost multiplier $\lambda \ge 0$ and KL multiplier $\eta \ge 0$, is

$$\begin{aligned} \mathcal{L}_i(q,\eta,\lambda) = {} & \mathbb{E}_{s\sim d_i,\, h\sim q(\cdot|s)} \left[Q_r^i(s,h) - \lambda Q_c^i(s,h)\right] \\ & - \eta\, \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) + \lambda\epsilon_{E,i} + \eta\epsilon_{\mathrm{KL}}. \end{aligned} \tag{8}$$

For $\eta > 0$ and $\lambda \geq 0$, define, for d_i-a.e. s, the hybrid partition function

$$Z_i(s;\eta,\lambda) = \sum_{m\in\mathcal{M}} \int_{\mathcal{A}_c^m} \pi_i(m, a_c \mid s) \exp\left(\frac{Q_r^i(s,m,a_c) - \lambda Q_c^i(s,m,a_c)}{\eta}\right) da_c. \tag{9}$$

The corresponding dual objective $g_i : (0,\infty)\times[0,\infty) \to \mathbb{R}$ is

$$g_i(\eta,\lambda) = \eta\epsilon_{\mathrm{KL}} + \lambda\epsilon_{E,i} + \eta\,\mathbb{E}_{s\sim d_i}\left[\log Z_i(s;\eta,\lambda)\right]. \tag{10}$$

Theorem 1 (Strong duality and closed-form hybrid E-step)**.** *Under Assumption 1, the E-step has zero duality gap, and g_i is jointly convex on $(0,\infty)\times[0,\infty)$. If g_i attains its infimum over this domain at $(\eta_i^\star, \lambda_i^\star)$ with $\eta_i^\star > 0$, then an optimal variational target, defined for d_i-a.e. s, has the μ-density*

$$q_i^\star(m, a_c \mid s) = \frac{\pi_i(m, a_c \mid s) \exp\left(\frac{Q_r^i(s,m,a_c) - \lambda_i^\star Q_c^i(s,m,a_c)}{\eta_i^\star}\right)}{Z_i(s;\eta_i^\star,\lambda_i^\star)}. \tag{11}$$

This target is feasible for Equation (7) and satisfies complementary slackness with $(\eta_i^\star, \lambda_i^\star)$.

The reward term favors hybrid actions with larger Q_r^i, while the cost term discounts those with larger Q_c^i according to $\lambda_i^\star$. The E-step therefore reallocates probability jointly over modes and their conditional controls before projection.

4.2 Hybrid-Action KL Geometry and Variational Projection

For a fixed state s, write $q(m, a_c \mid s) = q_m(m \mid s)q_c(a_c \mid s, m)$ and $\pi(m, a_c \mid s) = \pi_m(m \mid s)\pi_c(a_c \mid s, m)$, where q_m, π_m are mode marginals and q_c, π_c are mode-conditioned continuous densities. If $q(\cdot \mid s) \ll \pi(\cdot \mid s)$, the KL chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi(\cdot \mid s)) = {} & D_{\mathrm{KL}}(q_m(\cdot \mid s) \,\|\, \pi_m(\cdot \mid s)) \\ & + \mathbb{E}_{m\sim q_m(\cdot|s)} D_{\mathrm{KL}}(q_c(\cdot \mid s, m) \,\|\, \pi_c(\cdot \mid s, m)). \end{aligned} \tag{12}$$

Under the corresponding absolute-continuity conditions, the decomposition applies to the target-to-policy pair $(q_i^\star, \pi_\theta)$ and the old-to-new pair (π_i, π_θ). Let $\pi_{i,m}, \pi_{i,c}$ be the factors of π_i, and let $\epsilon_m, \epsilon_c > 0$ be the mode and conditional-continuous trust-region radii. Assuming finite log-density expectations, the M-step solves

$$\begin{aligned} \max_\theta \quad & \mathbb{E}_{s\sim d_i,\, h\sim q_i^\star(\cdot|s)} \left[\log \pi_\theta(h \mid s)\right] \\ \text{s.t.} \quad & \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(\pi_{i,m}(\cdot \mid s) \,\|\, \pi_{\theta_m}(\cdot \mid s)) \leq \epsilon_m, \\ & \mathbb{E}_{s\sim d_i,\, m\sim\pi_{i,m}(\cdot|s)} D_{\mathrm{KL}}(\pi_{i,c}(\cdot \mid s, m) \,\|\, \pi_{\theta_c}(\cdot \mid s, m)) \leq \epsilon_c. \end{aligned} \tag{13}$$

Let $q_{i,m}^\star$ be the mode marginal of $q_i^\star$. Using the policy factorization in Equation (5), the projection objective decomposes as

$$\begin{aligned} & \mathbb{E}_{s\sim d_i,\,(m,a_c)\sim q_i^\star(\cdot|s)} \left[\log \pi_\theta(m, a_c \mid s)\right] \\ = {} & \mathbb{E}_{s\sim d_i,\, m\sim q_{i,m}^\star(\cdot|s)} \left[\log \pi_{\theta_m}(m \mid s)\right] \\ & + \mathbb{E}_{s\sim d_i,\,(m,a_c)\sim q_i^\star(\cdot|s)} \left[\log \pi_{\theta_c}(a_c \mid s, m)\right]. \end{aligned} \tag{14}$$

Thus, both policy factors are fitted to the same joint variational target.

4.3 Bounded Approximate Dual and Target Admissibility

For approximate value functions and expectations, restrict the dual variables to

$$(\eta,\lambda) \in \mathcal{D}_i = [\eta_{\min}, \eta_{\max}] \times [0, \lambda_{\max}], \tag{15}$$

where $0 < \eta_{\min} \leq \eta_{\max} < \infty$ and $0 < \lambda_{\max} < \infty$. Let $\widehat{g}_i : \mathcal{D}_i \to \mathbb{R}$ be a continuous approximation of g_i. By compactness, it attains a minimizer $(\widehat{\eta}_i, \widehat{\lambda}_i) \in \mathcal{D}_i$; Appendix A.4 proves this existence claim.

For $(\eta, \lambda) \in \mathcal{D}_i$, let the normalized approximate E-step kernel $\widehat{q}_{i,\eta,\lambda}$ satisfy $\widehat{q}_{i,\eta,\lambda}(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-a.e. s, and set $\widehat{q}_i = \widehat{q}_{i,\widehat{\eta}_i,\widehat{\lambda}_i}$. For the fixed approximate cost action-value function $\widehat{Q}_c^i : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$, define the violation residuals, using $[x]_+ = \max\{x, 0\}$, by

$$\begin{aligned} r_{c,i}(q) &= \left[\mathbb{E}_{s\sim d_i,\, h\sim q(\cdot|s)} \left[\widehat{Q}_c^i(s,h)\right] - \epsilon_{E,i}\right]_+, \\ r_{\mathrm{KL},i}(q) &= \left[\mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) - \epsilon_{\mathrm{KL}}\right]_+. \end{aligned} \tag{16}$$

For prescribed cost- and KL-residual tolerances $\tau_c, \tau_{\mathrm{KL}} \geq 0$, let $\mathcal{Q}_{i,\mathrm{valid}}$ comprise kernels q with finite cost and KL expectations, $q(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-a.e. s, and $r_{c,i}(q) \leq \tau_c$ and $r_{\mathrm{KL},i}(q) \leq \tau_{\mathrm{KL}}$. Run the M-step only if $\widehat{q}_i \in \mathcal{Q}_{i,\mathrm{valid}}$; otherwise $\pi_{i+1} = \pi_i$.

4.4 Theoretical Properties

For a positive-temperature optimal E-step dual pair $(\eta_i^\star, \lambda_i^\star)$ satisfying Theorem 1, define, for a stochastic kernel q and hybrid policy π, the fixed-reference surrogate

$$\begin{aligned} \mathcal{F}_i(q, \pi) = \mathbb{E}_{s\sim d_i,\, h\sim q(\cdot|s)} &\left[Q_r^i(s,h) - \lambda_i^\star Q_c^i(s,h)\right] \\ &- \eta_i^\star \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi(\cdot \mid s)). \end{aligned} \tag{17}$$

Proposition 1 (Fixed-reference E/M surrogate improvement)**.** *Assume exact E/M solutions and finiteness of the displayed quantities. Then, for every stochastic kernel q with $q(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-a.e. s and finite $\mathcal{F}_i(q, \pi_i)$,*

$$\mathcal{F}_i(q_i^\star, \pi_{i+1}) \geq \mathcal{F}_i(q_i^\star, \pi_i) \geq \mathcal{F}_i(q, \pi_i). \tag{18}$$

The inequalities are the exact M- and E-step improvements, respectively, with $d_i, Q_r^i, Q_c^i, \eta_i^\star$, and $\lambda_i^\star$ held fixed.

The outer CMDP analysis instead uses the true normalized discounted occupancy measure of π_i,

$$d_{\pi_i} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_{\pi_i}(s_t \in \cdot). \tag{19}$$

For $x \in \{r, c\}$, let $Q_x^{\pi_i}$ and $V_x^{\pi_i}$ be the true discounted action- and state-value functions, and define $A_x^{\pi_i}(s,h) = Q_x^{\pi_i}(s,h) - V_x^{\pi_i}(s)$. Define

$$\Delta_i = \sup_{s\in\mathcal{S}} D_{\mathrm{TV}}(\pi_{i+1}(\cdot \mid s), \pi_i(\cdot \mid s)), \tag{20}$$

where, for probability measures P, Q on a common measurable space, $D_{\mathrm{TV}}(P, Q) = \sup_C |P(C) - Q(C)|$, with C ranging over measurable sets. Define also

$$B_{x,i+1} = \sup_{s\in\mathcal{S}} \left|\mathbb{E}_{h\sim\pi_{i+1}(\cdot|s)} \left[A_x^{\pi_i}(s,h)\right]\right|, \qquad x \in \{r, c\}. \tag{21}$$

The following proposition adapts the performance bounds of Achiam et al. (2017).

Proposition 2 (Conditional reward and cost bounds)**.** *Assume $B_{r,i+1}$ and $B_{c,i+1}$ are finite. Then*

$$\begin{aligned} J_r(\pi_{i+1}) &\geq J_r(\pi_i) + \frac{1}{1-\gamma} \mathbb{E}_{\substack{s\sim d_{\pi_i} \\ h\sim\pi_{i+1}(\cdot|s)}} [A_r^{\pi_i}(s,h)] - \frac{2\gamma B_{r,i+1}}{(1-\gamma)^2} \Delta_i, \\ J_c(\pi_{i+1}) &\leq J_c(\pi_i) + \frac{1}{1-\gamma} \mathbb{E}_{\substack{s\sim d_{\pi_i} \\ h\sim\pi_{i+1}(\cdot|s)}} [A_c^{\pi_i}(s,h)] + \frac{2\gamma B_{c,i+1}}{(1-\gamma)^2} \Delta_i. \end{aligned} \tag{22}$$

Define the cost-increase upper bound

$$U_{c,i} = \frac{1}{1-\gamma} \mathbb{E}_{\substack{s\sim d_{\pi_i} \\ h\sim\pi_{i+1}(\cdot|s)}} [A_c^{\pi_i}(s,h)] + \frac{2\gamma B_{c,i+1}}{(1-\gamma)^2} \Delta_i. \tag{23}$$

If $J_c(\pi_i) \leq d - \zeta$ for some constraint margin $\zeta > 0$ and $U_{c,i} \leq \zeta$, then $J_c(\pi_{i+1}) \leq d$, giving a sufficient feasibility margin condition.

Proposition 1 establishes exact fixed-reference surrogate improvement. Proposition 2 and the margin condition characterize the outer CMDP update using true occupancies and advantages. These bounds are conditional on the statewise shift Δ_i, which is not controlled by the averaged M-step KL constraints, and do not imply monotone empirical learning curves.

5 EXPERIMENTS

We compare H-CVPO with constrained hybrid-action baselines at three cost limits under matched interaction budgets and conduct component-wise ablations. The main text presents the experimental setup, baseline comparisons, and main ablation results. Appendix B provides training and task details, learning curves at additional cost limits, late-stage trends, an extended-training comparison with CHPO, and complete ablation results.

5.1 EXPERIMENTAL SETUP

Tasks and baselines. Following CHPO (Zhou et al., 2025), we evaluate constrained versions of *Moving*, *Sliding*, *HardMove*, and *Parking*, which respectively test mode–parameter coordination, inertial dynamics, combinatorial mode spaces, and collision-aware maneuvering. We compare H-CVPO with CHPO, HPPO-Lag, and PDQN-Rco: CHPO switches between reward and cost objectives, while HPPO-Lag and PDQN-Rco use Lagrangian and reward-constrained updates. H-CVPO instead separates joint target construction from factorized policy projection. Appendix B.2 provides task details and clarifies the scope of the theoretical and empirical comparisons.

Protocol. Unless stated otherwise, all experiments use training seeds 11, 22, and 33, with matched environments, interaction budgets, cost limits, and evaluation protocols across methods. For the standard 320k experiments, all tabulated final-policy results use the policy obtained at the fixed training endpoint for every method and seed. We do not select checkpoints after training, keeping both the training budget and checkpoint-selection rule matched across methods.

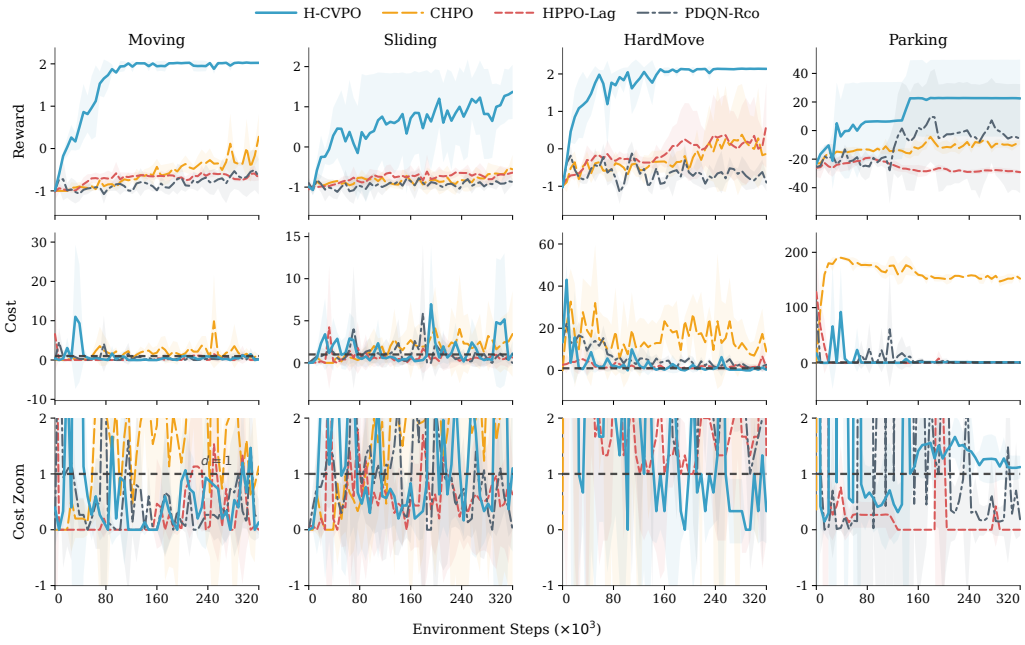

Figure 1: Learning curves for H-CVPO and the baselines under the primary cost limit $d = 1.0$. Lines and shaded regions show the mean and one standard deviation over the three training seeds. Rows report reward, cost on the full scale, and cost near the constraint boundary; horizontal dashed lines mark the cost limit.

5.2 COMPARISON WITH BASELINES

Performance across hybrid-control tasks. Figure 1 compares H-CVPO with the baselines under the primary cost limit $d = 1.0$. By roughly 120k environment steps, mean rewards on Moving and HardMove approach 2 and subsequently stabilize, while their pronounced initial cost spikes largely subside. Parking undergoes a second stage of improvement around 140–160k steps, when

Table 1: Final performance after 320k environment steps under different cost limits. For each training seed, we average 40 deterministic evaluation episodes using fixed evaluation seeds 22000–22039 and report the mean and sample standard deviation across the three training seeds.

		H-CVPO		CHPO		HPPO-Lag		PDQN-Rco	
Environment	d	Reward	Cost	Reward	Cost	Reward	Cost	Reward	Cost
Moving	1.0	2.07±0.03	0.85±0.97	-0.04±0.19	0.98±0.73	-0.69±0.09	1.30±1.51	-0.72±0.43	0.12±0.20
	1.5	1.41±1.18	1.07±0.60	-0.12±0.35	1.83±0.52	-0.62±0.08	1.33±1.50	-0.60±0.49	0.28±0.30
	2.0	2.08±0.02	1.00±0.43	-0.12±0.35	1.83±0.52	-0.46±0.10	0.37±0.16	0.02±0.43	5.83±7.87
Sliding	1.0	1.08±0.61	0.67±0.47	-0.63±0.21	3.79±1.71	-0.67±0.16	0.61±0.38	-0.95±0.05	0.49±0.48
	1.5	1.55±0.85	0.68±0.30	-0.62±0.26	2.06±0.55	-0.64±0.04	1.07±1.24	-0.43±0.38	2.19±0.83
	2.0	1.99±0.10	0.84±0.06	-0.79±0.15	2.08±0.46	-0.64±0.09	0.46±0.14	-0.63±0.10	3.52±5.55
HardMove	1.0	2.05±0.00	0.42±0.52	-0.33±0.43	10.58±4.93	0.34±1.25	1.17±0.38	-0.73±0.12	2.25±2.17
	1.5	2.06±0.01	2.17±3.32	-0.29±0.39	9.00±2.22	-0.31±0.22	3.25±2.61	-0.31±0.39	2.00±2.38
	2.0	2.05±0.02	0.50±0.43	-0.31±0.42	13.00±6.01	-0.33±0.18	1.67±0.38	-0.85±0.15	2.33±0.88
Parking	1.0	23.26±26.94	1.06±0.16	7.45±3.31	91.98±3.59	-23.78±2.12	0.15±0.05	0.37±33.49	0.19±0.29
	1.5	38.82±0.01	0.85±0.28	7.45±3.31	91.98±3.59	-23.36±0.35	0.12±0.00	-14.60±9.24	63.83±104.30
	2.0	35.92±5.03	1.23±0.13	7.45±3.31	91.98±3.59	-22.54±0.59	0.12±0.02	-22.86±9.95	59.64±49.87

reward rises from its intermediate plateau to above 20; mean cost is below the limit at several earlier checkpoints but rises above it during this reward increase. Sliding improves more gradually, with reward fluctuations and recurrent cost spikes throughout the 320k-step budget.

CHPO shows modest late-stage reward gains on Moving and HardMove, but continues to incur substantial costs on HardMove and especially Parking through 320k steps. HPPO-Lag and PDQN-Rco generally maintain lower costs, but achieve substantially lower rewards than H-CVPO on Moving, Sliding, and HardMove. On Parking, both baselines also achieve substantially lower late-stage rewards than H-CVPO.

Feasibility requires mean episodic evaluation cost not exceeding d. Evaluations during training identify multiple feasible H-CVPO checkpoints per task (Figure 1). Following Section 5.1, we compare only 320k-endpoint policies in Table 1, with matched training budgets and uniform checkpoint-selection and evaluation protocols, without post-hoc selection of earlier feasible checkpoints.

Performance under different cost limits. Table 1 reports final-policy performance across 12 task–limit settings with $d \in \{1.0, 1.5, 2.0\}$. Under the evaluation-based feasibility criterion above, H-CVPO satisfies the limit in 10 settings and achieves the highest mean reward in all 12. Its two endpoint cost violations occur on Parking at $d = 1.0$, where the mean cost is 1.06, and on HardMove at $d = 1.5$, where the mean cost is 2.17. No evaluated method is feasible in the latter setting.

CHPO is feasible only on Moving at $d = 1.0$ and 2.0, with costs of 9.00–13.00 on HardMove and 91.98 on Parking. Its Sliding cost of 2.08 at $d = 2.0$, like H-CVPO's Parking cost of 1.06 at $d = 1.0$, counts as infeasible despite the small exceedance. HPPO-Lag is feasible in 9 settings but has lower mean rewards than H-CVPO in all 12, often with costs well below the limits. PDQN-Rco is feasible in 4 settings, with large cost exceedances and cross-seed variability in several cases. Thus, H-CVPO has both the highest feasibility count and the highest mean rewards.

5.3 Ablation Studies

We examine three single-component variants under $d = 1.0$. *w/o Cost-Aware E-Step* removes only the cost term from the joint target in Equation (11); the E-step itself, the cost critic, and all other update components are retained. *Joint Hybrid KL* replaces the two old-to-new M-step KL constraints in Equation (13) with a single joint KL constraint at the same total KL allowance. *w/o Target Admissibility* removes only the pre-projection filter based on the cost and KL residuals in Equation (16), disabling the *Skip* branch in Algorithm 1 while retaining post-update acceptance checks. All other training and evaluation settings remain fixed. We focus on the cost-aware E-step here; Appendix B.4 reports all three variants across $d \in \{1.0, 1.5, 2.0\}$.

Figure 2 compares full H-CVPO with the reward-only E-step variant. The ablation's mean cost exceeds $d = 1.0$ for most of training on Sliding, HardMove, and Parking. Full H-CVPO has substantially lower costs on these tasks, although Sliding and HardMove retain cost spikes and

Parking remains above the limit late in training. Removing cost awareness also delays reward improvement on Moving and yields lower rewards on HardMove and Parking; Sliding has broadly comparable rewards later in training.

The fixed 320k-endpoint evaluations in Table 5 provide a complementary comparison. At $d = 1.0$, removing cost awareness changes Moving and HardMove from feasible to infeasible, raising mean costs from 0.85 to 1.73 and from 0.42 to 3.25, respectively. On Parking, full H-CVPO already slightly exceeds the limit; the ablation lowers mean reward from 23.26 to -18.30 and raises mean cost from 1.06 to 51.39. Sliding is an exception: both endpoint policies are feasible, and the ablation raises mean reward from 1.08 to 1.31 while lowering mean cost from 0.67 to 0.52, but increases the across-seed reward standard deviation from 0.61 to 1.32. Across all three cost limits, removing cost-aware reweighting changes eight of the twelve task–limit settings from feasible to infeasible. These results support cost-aware variational reweighting as an important contributor to constraint control, with task- and limit-dependent exceptions.

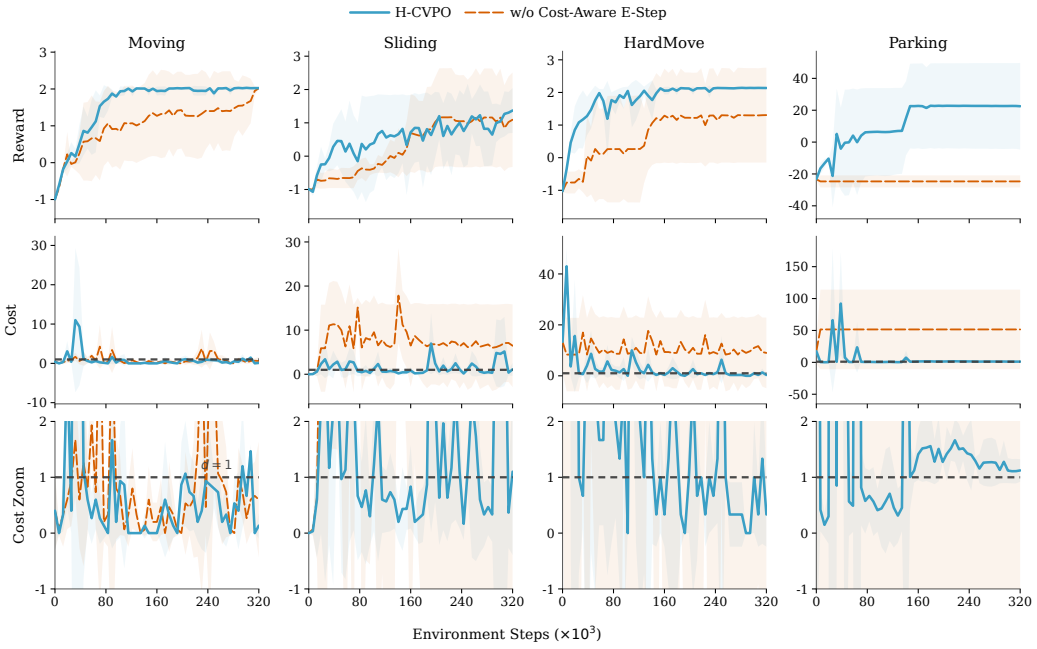

Figure 2: Ablation of the cost-aware variational E-step under the primary cost limit $d = 1.0$. Curves compare full H-CVPO with the variant removing the cost contribution from the variational target. Curves and shading denote the mean and one standard deviation over the three training seeds; dashed lines mark the cost limit.

6 Conclusion

We address constrained reinforcement learning in hybrid action spaces with H-CVPO, a variational framework that separates joint target construction from factorized policy projection. Its E-step forms a reward–cost-aware joint nonparametric target, while its M-step fits both policy components under separate discrete and conditional-continuous KL controls. Under the stated assumptions, exact E/M updates yield fixed-reference surrogate improvement and conditional reward–cost bounds; approximate updates use cost and KL residuals to determine target projection. Across four tasks and three cost limits, H-CVPO shows favorable reward–cost trade-offs against hybrid-action baselines. Ablations identify cost-aware reweighting as important for constraint control, while the effects of separate KL controls and target-admissibility filtering vary by task and cost limit. As discrete–continuous decision structures arise broadly in engineered systems, we expect H-CVPO to offer a general approach to constrained hybrid control across industrial systems and beyond.

AI USE STATEMENT

Generative AI tools were used to improve the language, clarity, and presentation of the manuscript. All scientific contributions were developed independently, and all reported results were independently verified. All AI-assisted text was reviewed, and the authors take full responsibility for the content of the manuscript.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018. URL https://arxiv.org/abs/1806.06920.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999. ISBN 978-0-8493-0382-1.

Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019. URL https://arxiv.org/abs/1901.10031.

Olivier Delalleau, Maxim Peter, Eloi Alonso, and Adrien Logut. Discrete and continuous action representation for practical RL in video games. *arXiv preprint arXiv:1912.11077*, 2019. URL https://arxiv.org/abs/1912.11077.

Jingzheng Fan, Bingfeng Zu, Jianwei Zhou, Zhen Wang, and Haopeng Wang. Adaptive mode selection strategy for series-parallel hybrid electric vehicles based on variable power reserve. *Energies*, 14(11):3171, 2021. doi: 10.3390/en14113171.

Zhou Fan, Rui Su, Weinan Zhang, and Yong Yu. Hybrid actor–critic reinforcement learning in parameterized action space. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2279–2285, 2019. doi: 10.24963/ijcai.2019/316.

Matthew Hausknecht and Peter Stone. Deep reinforcement learning in parameterized action space. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016. URL https://arxiv.org/abs/1511.04143.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018. URL https://arxiv.org/abs/1805.00909.

Lijuan Li, Xue Yang, Shipin Yang, and Xiaowei Xu. Optimization of oxygen system scheduling in hybrid action space based on deep reinforcement learning. *Computers & Chemical Engineering*, 171:108168, 2023. doi: 10.1016/j.compchemeng.2023.108168.

Yuan Lin, Xiao Liu, and Zishun Zheng. Discretionary lane-change decision and control via parameterized soft actor–critic for hybrid action space. *Machines*, 12(4):213, 2024. doi: 10.3390/machines12040213.

Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13644–13668. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22b.html.

Warwick Masson, Pravesh Ranchod, and George Konidaris. Reinforcement learning with parameterized actions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1934–1940. AAAI Press, 2016. doi: 10.1609/aaai.v30i1.10226.

Luis Montero, Antonio Bello, and Javier Reneses. A review on the unit commitment problem: Approaches, techniques, and resolution methods. *Energies*, 15(4):1296, 2022. doi: 10.3390/en15041296.

Michael Neunert, Abbas Abdolmaleki, Markus Wulfmeier, Thomas Lampe, Tobias Springenberg, Roland Hafner, Francesco Romano, Jonas Buchli, Nicolas Heess, and Martin Riedmiller. Continuous-discrete reinforcement learning for hybrid control in robotics. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 735–751. PMLR, 30 Oct–01 Nov 2020. URL https://proceedings.mlr.press/v100/neunert20a.html.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

K. Shanti Swarup and P. V. Simi. Neural computation using discrete and continuous Hopfield networks for power system economic dispatch and unit commitment. *Neurocomputing*, 70(1–3): 119–129, 2006. doi: 10.1016/j.neucom.2006.05.002.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfrvsA9FX.

Jiechao Xiong, Qing Wang, Zhuoran Yang, Peng Sun, Lei Han, Yang Zheng, Haobo Fu, Tong Zhang, Ji Liu, and Han Liu. Parametrized deep Q-networks learning: Reinforcement learning with discrete–continuous hybrid action space. *arXiv preprint arXiv:1810.06394*, 2018. URL https://arxiv.org/abs/1810.06394.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rke3TJrtPS.

Ao Zhou, Jiayi Guan, Li Shen, Fan Lu, Sanqing Qu, Junqiao Zhao, Ziqiao Wang, Ya Wu, and Guang Chen. CHPO: Constrained hybrid-action policy optimization for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 73157–73185, 2025. doi: 10.52202/085713-2201.

Weichao Zhuang, Xiaowu Zhang, Daofei Li, Liangmo Wang, and Guodong Yin. Mode shift map design and integrated energy management control of a multi-mode hybrid electric vehicle. *Applied Energy*, 204:476–488, 2017. doi: 10.1016/j.apenergy.2017.07.059.

A Proofs and Derivations

We provide full proofs and derivations of the results in Section 4.

A.1 Proof of Strong Duality and Closed-Form Hybrid E-Step

We retain the notation of Section 4.1, including the reference measure d_i, current policy π_i, action-value functions Q_r^i, Q_c^i, hybrid reference measure μ, and E-step bounds $\epsilon_{E,i}, \epsilon_{\mathrm{KL}}$.

Density-ratio formulation. Define the joint probability measure ν_i on $\mathcal{S} \times \mathcal{H}$ by

$$\nu_i(ds, dh) = d_i(ds)\pi_i(dh \mid s). \tag{24}$$

For an admissible kernel q, write $q(h \mid s)$ and $\pi_i(h \mid s)$ for their μ-densities. Absolute continuity gives the density ratio

$$w(s, h) = \frac{q(h \mid s)}{\pi_i(h \mid s)}, \tag{25}$$

defined ν_i-a.e. Let $L^1(\nu_i)$ denote the space of ν_i-integrable functions identified up to ν_i-a.e. equality, with norm $\|w\|_{L^1(\nu_i)} = \int |w|\, d\nu_i$; define $L^1(d_i)$ analogously. Let $\mathcal{W}_i$ denote the admissible density ratios:

$$\mathcal{W}_i = \left\{ w \in L^1(\nu_i) : w \geq 0, \quad \int_{\mathcal{H}} w(s,h)\pi_i(dh \mid s) = 1 \text{ for } d_i\text{-a.e. } s \right\}. \tag{26}$$

For each $w \in \mathcal{W}_i$, choose a jointly measurable nonnegative representative and set $q(h \mid s) = w(s,h)\pi_i(h \mid s)$ on the states where the normalization holds. On the exceptional measurable d_i-null set, set $q(\cdot \mid s) = \pi_i(\cdot \mid s)$. This defines an admissible kernel without changing any ν_i-integral. We use the same null-set convention for subsequent kernel constructions. The conditional-integration map is bounded and linear from $L^1(\nu_i)$ to $L^1(d_i)$. Together with the weak closedness of the nonnegative cone, this shows that $\mathcal{W}_i$ is weakly closed and convex.

For $w \in \mathcal{W}_i$, define

$$\begin{aligned} R_i(w) &= \int_{\mathcal{S}\times\mathcal{H}} w(s,h) Q_r^i(s,h)\, \nu_i(ds,dh), \\ C_i(w) &= \int_{\mathcal{S}\times\mathcal{H}} w(s,h) Q_c^i(s,h)\, \nu_i(ds,dh), \\ H_i(w) &= \int_{\mathcal{S}\times\mathcal{H}} w(s,h) \log w(s,h)\, \nu_i(ds,dh), \end{aligned} \tag{27}$$

with $0 \log 0 = 0$. Here, $R_i(w)$ and $C_i(w)$ are the fixed-reference expectations of the reward and cost action-values, while $H_i(w)$ is the averaged KL divergence in Equation (7).

Primal attainment. Assumption 1 makes R_i and C_i weakly continuous linear functionals on $L^1(\nu_i)$. The entropy functional H_i is proper, convex, and weakly lower semicontinuous. Every $w \in \mathcal{W}_i$ satisfies $\int w\, d\nu_i = 1$. For a finite entropy bound $K \geq 0$ and a tail threshold $t > 1$,

$$\sup_{\substack{w\in\mathcal{W}_i \\ H_i(w)\leq K}} \int_{\{w>t\}} w\, d\nu_i \leq \frac{K + 1/e}{\log t}, \tag{28}$$

since $w \log w \geq -1/e$. Thus, these sublevel sets are uniformly integrable and hence relatively weakly compact by the Dunford–Pettis theorem. They are also weakly closed, by the weak lower semicontinuity of H_i and weak closedness of $\mathcal{W}_i$, and are therefore weakly compact. Let $p_i^\star$ denote the optimal E-step value:

$$\begin{aligned} p_i^\star = \sup_{w\in\mathcal{W}_i} \quad & R_i(w) \\ \text{s.t.} \quad & C_i(w) \leq \epsilon_{E,i}, \qquad H_i(w) \leq \epsilon_{\mathrm{KL}} \end{aligned} \tag{29}$$

The feasible set is nonempty by strict feasibility and weakly compact by the preceding argument and weak continuity of C_i. Hence the weakly continuous objective R_i attains its supremum.

Strong duality. To establish strong duality, let u_c and u_{KL} denote perturbed upper bounds on cost and KL divergence, respectively, and let u_r denote a perturbed reward level. Define

$$\mathcal{V}_i = \left\{ (u_c, u_{\mathrm{KL}}, u_r) \in \mathbb{R}^3 : \begin{array}{l} \text{there exists } w \in \mathcal{W}_i \text{ such that} \\ C_i(w) \leq u_c,\ H_i(w) \leq u_{\mathrm{KL}},\ R_i(w) \geq u_r \end{array} \right\}. \tag{30}$$

Convexity of H_i makes $\mathcal{V}_i$ convex. To see that it is closed, consider a convergent sequence in $\mathcal{V}_i$. Its entropy coordinates are bounded, so the corresponding witnesses lie in a common weakly compact entropy sublevel set. Passing to a weakly convergent subsequence gives a limit in $\mathcal{W}_i$; weak continuity of R_i, C_i and weak lower semicontinuity of H_i preserve the three defining inequalities.

The point $(\epsilon_{E,i}, \epsilon_{\mathrm{KL}}, p_i^\star)$ lies on the boundary of $\mathcal{V}_i$. Its monotonicity in the three coordinates implies that a supporting hyperplane can be written with the cost, KL, and objective coefficients $\lambda, \eta, \alpha \geq 0$, not all zero, as

$$\lambda(u_c - \epsilon_{E,i}) + \eta(u_{\mathrm{KL}} - \epsilon_{\mathrm{KL}}) - \alpha(u_r - p_i^\star) \geq 0, \qquad (u_c, u_{\mathrm{KL}}, u_r) \in \mathcal{V}_i. \tag{31}$$

If $\alpha = 0$, evaluating this inequality at the density ratio associated with the strictly feasible kernel in Assumption 1 forces $\lambda = \eta = 0$, a contradiction. Thus, $\alpha > 0$, and rescaling gives $\alpha = 1$. Consequently, for every $w \in \mathcal{W}_i$ with finite $H_i(w)$,

$$R_i(w) - \lambda\big(C_i(w) - \epsilon_{E,i}\big) - \eta\big(H_i(w) - \epsilon_{\mathrm{KL}}\big) \leq p_i^\star. \tag{32}$$

This inequality extends to the full Lagrangian domain. If $\eta > 0$, ratios with $H_i(w) = \infty$ have Lagrangian value $-\infty$. If $\eta = 0$, the entropy term is omitted, and finite-entropy ratios suffice by the following approximation. For any $w \in \mathcal{W}_i$ and integer $n \geq 1$, define the retained conditional mass and a bounded density ratio by

$$b_n(s) = \int_{\mathcal{H}} \min\{w(s,h), n\}\pi_i(dh \mid s), \qquad w_n(s,h) = \min\{w(s,h), n\} + 1 - b_n(s). \tag{33}$$

Then $w_n \in \mathcal{W}_i$ and $0 \leq w_n \leq n+1$ almost everywhere. Moreover,

$$\|w_n - w\|_{L^1(\nu_i)} \leq 2 \int \big(w - \min\{w, n\}\big)\, d\nu_i \longrightarrow 0. \tag{34}$$

Thus, $H_i(w_n) < \infty$, while boundedness of Q_r^i and Q_c^i gives $R_i(w_n) \to R_i(w)$ and $C_i(w_n) \to C_i(w)$. This extends the inequality to every $w \in \mathcal{W}_i$ when $\eta = 0$, without requiring entropy convergence. The dual objective at the supporting multipliers is therefore at most $p_i^\star$; weak duality gives the reverse inequality. Hence the E-step has zero duality gap on the full multiplier domain.

The positive-temperature Gibbs target. We next identify the Lagrangian maximizer. For fixed $\eta > 0$ and $\lambda \geq 0$, define the probability kernel $q_{\eta,\lambda}$, with the fixed iteration index i suppressed, using the partition function Z_i from Equation (9). Its μ-density is

$$q_{\eta,\lambda}(h \mid s) = \frac{\pi_i(h \mid s) \exp\Big(\frac{Q_r^i(s,h) - \lambda Q_c^i(s,h)}{\eta}\Big)}{Z_i(s;\eta,\lambda)}. \tag{35}$$

The boundedness in Assumption 1 makes the partition function finite and strictly positive for d_i-a.e. s. With the null-set convention above, this defines a measurable probability kernel with $q_{\eta,\lambda}(\cdot \mid s) \ll \pi_i(\cdot \mid s)$. Its log-density ratio relative to π_i is bounded ν_i-a.e. for fixed (η, λ), so its averaged KL divergence is finite. For any admissible kernel q with finite averaged KL divergence, direct expansion gives, for d_i-a.e. s,

$$\begin{aligned} &\mathbb{E}_{h\sim q(\cdot|s)}\left[Q_r^i(s,h) - \lambda Q_c^i(s,h)\right] - \eta D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) \\ &\quad = \eta \log Z_i(s;\eta,\lambda) - \eta D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, q_{\eta,\lambda}(\cdot \mid s)). \end{aligned} \tag{36}$$

The last term is nonpositive and vanishes precisely at $q = q_{\eta,\lambda}$ up to null sets. Hence $q_{\eta,\lambda}$ maximizes the Lagrangian; kernels of infinite averaged KL divergence have Lagrangian value $-\infty$. Substitution yields the dual objective $g_i(\eta,\lambda)$ in Equation (10).

The zero-temperature boundary. The multiplier domain of the Lagrange dual includes $\eta = 0$, whereas Equation (10) is written for $\eta > 0$. We verify that its boundary limit equals the actual Lagrangian supremum with the entropy term omitted. For fixed $\lambda \geq 0$, define

$$S_{i,\lambda}(s) = \operatorname*{ess\,sup}_{h\sim\pi_i(\cdot|s)} \left\{Q_r^i(s,h) - \lambda Q_c^i(s,h)\right\}. \tag{37}$$

The Laplace principle gives, for d_i-a.e. s,

$$\eta \log Z_i(s;\eta,\lambda) \longrightarrow S_{i,\lambda}(s) \qquad \text{as } \eta \downarrow 0. \tag{38}$$

This limit provides a measurable version of $S_{i,\lambda}$. Both sides are bounded in absolute value by $M_r + \lambda M_c$, so dominated convergence permits integration over d_i. Every admissible kernel has expected reward–cost action-value at most $\mathbb{E}_{s\sim d_i}[S_{i,\lambda}(s)]$. Conversely, evaluating the Gibbs identity at $q_{\eta,\lambda}$ and using nonnegativity of KL gives

$$\begin{aligned} \eta\mathbb{E}_{s\sim d_i}[\log Z_i(s;\eta,\lambda)] &\leq \mathbb{E}_{s\sim d_i,\, h\sim q_{\eta,\lambda}(\cdot|s)}[Q_r^i(s,h) - \lambda Q_c^i(s,h)] \\ &\leq \mathbb{E}_{s\sim d_i}[S_{i,\lambda}(s)]. \end{aligned} \tag{39}$$

Letting $\eta \downarrow 0$ proves that these kernels approach the upper bound. Therefore the boundary dual objective is

$$\begin{aligned} g_i(0,\lambda) &= \lambda\epsilon_{E,i} + \sup_{w\in\mathcal{W}_i} \{R_i(w) - \lambda C_i(w)\} \\ &= \lambda\epsilon_{E,i} + \mathbb{E}_{s\sim d_i}[S_{i,\lambda}(s)] = \lim_{\eta\downarrow 0} g_i(\eta,\lambda). \end{aligned} \tag{40}$$

Thus, allowing $\eta = 0$ does not lower the infimum, and the full dual value is $\inf_{\eta>0,\,\lambda\geq 0} g_i(\eta,\lambda)$. This argument does not assert that a minimizing pair has positive temperature.

Convexity and optimality conditions. It remains to establish convexity and the optimality conditions. For d_i-a.e. state s, let $\widetilde{q}(\cdot \mid s)$ range over probability measures on $\mathcal{H}$ that are absolutely continuous with respect to $\pi_i(\cdot \mid s)$. The preceding KL identity is equivalently the Gibbs variational formula

$$\eta \log Z_i(s;\eta,\lambda) = \sup_{\widetilde{q}(\cdot|s)\ll\pi_i(\cdot|s)} \left\{ \mathbb{E}_{h\sim\widetilde{q}(\cdot|s)} \left[Q_r^i(s,h) - \lambda Q_c^i(s,h) \right] - \eta D_{\mathrm{KL}}(\widetilde{q}(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) \right\}. \tag{41}$$

Measures with infinite KL divergence contribute $-\infty$ and may be excluded from the supremum. For each remaining $\widetilde{q}$, the expression in braces is affine in (η, λ); its supremum is therefore jointly convex. Averaging over d_i and adding the affine terms $\eta\epsilon_{\mathrm{KL}} + \lambda\epsilon_{E,i}$ preserves convexity. Consequently, g_i is jointly convex on $(0,\infty) \times [0,\infty)$. Moreover, the boundedness of Q_r^i and Q_c^i supplies local dominating functions, so differentiation may pass through the state and action integrals for $\eta > 0$. The Gibbs kernel gives

$$\begin{aligned} \frac{\partial g_i}{\partial \eta} &= \epsilon_{\mathrm{KL}} - \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q_{\eta,\lambda}(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)), \\ \frac{\partial g_i}{\partial \lambda} &= \epsilon_{E,i} - \mathbb{E}_{s\sim d_i,\, h\sim q_{\eta,\lambda}(\cdot|s)} \left[Q_c^i(s,h) \right]. \end{aligned} \tag{42}$$

Suppose that $(\eta_i^\star, \lambda_i^\star)$ minimizes g_i with $\eta_i^\star > 0$, and set $q_i^\star = q_{\eta_i^\star,\lambda_i^\star}$. The temperature variable is interior, whereas λ is constrained to be nonnegative. First-order optimality therefore yields

$$\begin{aligned} \mathbb{E}_{s\sim d_i} D_{\mathrm{KL}}(q_i^\star(\cdot \mid s) \,\|\, \pi_i(\cdot \mid s)) &= \epsilon_{\mathrm{KL}}, \\ \mathbb{E}_{s\sim d_i,\, h\sim q_i^\star(\cdot|s)} \left[Q_c^i(s,h) \right] &\le \epsilon_{E,i}, \\ \lambda_i^\star \left(\mathbb{E}_{s\sim d_i,\, h\sim q_i^\star(\cdot|s)} \left[Q_c^i(s,h) \right] - \epsilon_{E,i} \right) &= 0. \end{aligned} \tag{43}$$

These are primal feasibility and complementary slackness. Since $q_i^\star$ maximizes the Lagrangian at the dual optimum, zero duality gap establishes its primal optimality and completes the proof of Theorem 1.

A.2 Derivation of Hybrid-Action KL Decomposition

Consider the joint densities q, π and their factors q_m, π_m, q_c, π_c introduced in Section 4.2. The mode set $\mathcal{M}$, conditional action spaces $\mathcal{A}_c^m$, dimensions p_m, and reference-measure convention are as defined in Equations (3)– (4). Fix a state s such that $q(\cdot \mid s) \ll \pi(\cdot \mid s)$. This absolute continuity implies $q_m(\cdot \mid s) \ll \pi_m(\cdot \mid s)$ and, for every mode satisfying $q_m(m \mid s) > 0$, $q_c(\cdot \mid s, m) \ll \pi_c(\cdot \mid s, m)$. Hence, for $q(\cdot \mid s)$-a.e. (m, a_c),

$$\log \frac{q(m, a_c \mid s)}{\pi(m, a_c \mid s)} = \log \frac{q_m(m \mid s)}{\pi_m(m \mid s)} + \log \frac{q_c(a_c \mid s, m)}{\pi_c(a_c \mid s, m)}. \tag{44}$$

Integrating this identity under $q(\cdot \mid s)$ and using $\int_{\mathcal{A}_c^m} q_c(a_c \mid s, m)\, da_c = 1$ gives

$$\begin{aligned} & D_{\mathrm{KL}}(q(\cdot \mid s) \,\|\, \pi(\cdot \mid s)) \\ &= \sum_{m\in\mathcal{M}} q_m(m \mid s) \int_{\mathcal{A}_c^m} q_c(a_c \mid s, m) \left[\log \frac{q_m(m \mid s)}{\pi_m(m \mid s)} + \log \frac{q_c(a_c \mid s, m)}{\pi_c(a_c \mid s, m)} \right] da_c \\ &= \sum_{m\in\mathcal{M}} q_m(m \mid s) \log \frac{q_m(m \mid s)}{\pi_m(m \mid s)} \\ &\quad + \sum_{m\in\mathcal{M}} q_m(m \mid s) D_{\mathrm{KL}}(q_c(\cdot \mid s, m) \,\|\, \pi_c(\cdot \mid s, m)). \end{aligned} \tag{45}$$

The first sum is $D_{\mathrm{KL}}(q_m(\cdot \mid s) \,\|\, \pi_m(\cdot \mid s))$, and the second is the $q_m(\cdot \mid s)$-expectation of the mode-conditioned KL divergence. Terms with $q_m(m \mid s) = 0$ are assigned value zero. The identity also holds when a KL divergence is infinite: the negative parts of the log-density ratios are integrable, since $-t \log t \le 1/e$ for $0 < t \le 1$, so the decomposition involves no subtraction of infinities. This proves Equation (12); the unit-mass convention also covers modes with $p_m = 0$.

A.3 Derivation of Factorized Variational Projection

As in Section 4.2, assume finite log-density expectations under $d_i(ds)q_i^\star(dh \mid s)$. By Equation (5), the hybrid policy satisfies the following identity almost everywhere under this measure:

$$\log \pi_\theta(m, a_c \mid s) = \log \pi_{\theta_m}(m \mid s) + \log \pi_{\theta_c}(a_c \mid s, m). \tag{46}$$

Recall that $q_{i,m}^\star$ is the mode marginal of the joint target $q_i^\star$ in Equation (11); for d_i-a.e. s,

$$q_{i,m}^\star(m \mid s) = \int_{\mathcal{A}_c^m} q_i^\star(m, a_c \mid s)\, da_c. \tag{47}$$

Since $\log \pi_{\theta_m}(m \mid s)$ is independent of a_c, linearity of expectation and the preceding marginalization give

$$\begin{aligned} &\mathbb{E}_{s\sim d_i,\,(m,a_c)\sim q_i^\star(\cdot|s)} \left[\log \pi_\theta(m, a_c \mid s)\right] \\ &= \mathbb{E}_{s\sim d_i,\, m\sim q_{i,m}^\star(\cdot|s)} \left[\log \pi_{\theta_m}(m \mid s)\right] \\ &\quad + \mathbb{E}_{s\sim d_i,\,(m,a_c)\sim q_i^\star(\cdot|s)} \left[\log \pi_{\theta_c}(a_c \mid s, m)\right]. \end{aligned} \tag{48}$$

The discrete-policy term therefore depends only on $q_{i,m}^\star$, whereas the conditional-continuous term retains the full joint target $q_i^\star$. This is Equation (14).

A.4 Existence of an Approximate Dual Minimizer

By Equation (15), $\mathcal{D}_i$ is the Cartesian product of the nonempty closed intervals $[\eta_{\min}, \eta_{\max}]$ and $[0, \lambda_{\max}]$. It is therefore closed and bounded in $\mathbb{R}^2$, and hence compact by the Heine–Borel theorem. The approximate dual objective $\widehat{g}_i : \mathcal{D}_i \to \mathbb{R}$ is assumed continuous in Section 4.3. The extreme-value theorem then yields a pair $(\widehat{\eta}_i, \widehat{\lambda}_i) \in \mathcal{D}_i$ satisfying

$$\widehat{g}_i(\widehat{\eta}_i, \widehat{\lambda}_i) = \min_{(\eta,\lambda)\in\mathcal{D}_i} \widehat{g}_i(\eta, \lambda). \tag{49}$$

Thus, the bounded approximate dual admits a minimizer, as claimed in Section 4.3.

A.5 Proof of Fixed-Reference E/M Surrogate Improvement

We use the exact-update and finiteness assumptions of Proposition 1, including the finite M-step log-density expectations specified in Section 4.2. All quantities defining the fixed-reference surrogate—d_i, Q_r^i, Q_c^i, the KL multiplier $\eta_i^\star > 0$, and the cost multiplier $\lambda_i^\star \geq 0$—are held fixed throughout the E/M cycle. Let q be any stochastic kernel from $\mathcal{S}$ to $\mathcal{H}$ satisfying $q(\cdot \mid s) \ll \pi_i(\cdot \mid s)$ for d_i-a.e. s and having finite $\mathcal{F}_i(q, \pi_i)$. Equations (8) and (17) give

$$\mathcal{L}_i(q, \eta_i^\star, \lambda_i^\star) = \mathcal{F}_i(q, \pi_i) + \lambda_i^\star \epsilon_{E,i} + \eta_i^\star \epsilon_{\mathrm{KL}}. \tag{50}$$

The last two terms are independent of q, so the Lagrangian and $\mathcal{F}_i(\cdot, \pi_i)$ have the same maximizers on this domain. By Theorem 1, the exact E-step target $q_i^\star$ is such a maximizer. Therefore,

$$\mathcal{F}_i(q_i^\star, \pi_i) \geq \mathcal{F}_i(q, \pi_i) \tag{51}$$

for every such q.

Next fix $q_i^\star$. The current policy π_i is M-step feasible because both old-to-new component KL divergences in Equation (13) vanish when evaluated at π_i. Hence the exact M-step solution π_{i+1} satisfies

$$\mathbb{E}_{s\sim d_i,\, h\sim q_i^\star(\cdot|s)} \left[\log \pi_{i+1}(h \mid s) - \log \pi_i(h \mid s)\right] \geq 0. \tag{52}$$

For fixed $q_i^\star$, only the KL term in $\mathcal{F}_i(q_i^\star, \pi)$ depends on π. Expanding its difference gives

$$\begin{aligned} &\mathcal{F}_i(q_i^\star, \pi_{i+1}) - \mathcal{F}_i(q_i^\star, \pi_i) \\ &= \eta_i^\star \mathbb{E}_{s\sim d_i,\, h\sim q_i^\star(\cdot|s)} \left[\log \pi_{i+1}(h \mid s) - \log \pi_i(h \mid s)\right] \geq 0. \end{aligned} \tag{53}$$

Combining the two inequalities gives the chain in Proposition 1.

A.6 Proof of Conditional Reward and Cost Bounds

We use the occupancies $d_{\pi_i}, d_{\pi_{i+1}}$, returns J_x, advantages $A_x^{\pi_i}$, and statewise policy shift Δ_i from Section 4.4, with $x \in \{r, c\}$ indexing reward and cost. These are true normalized discounted occupancies, distinct from the fixed reference d_i. The transition kernel P, initial-state distribution ρ_0, discount factor γ, and cost limit d are as defined in Section 3. Define the new-policy expected advantage relative to π_i by

$$\bar{A}_{x,i+1}(s) = \mathbb{E}_{h\sim\pi_{i+1}(\cdot|s)}\left[A_x^{\pi_i}(s,h)\right]. \tag{54}$$

By the definition of $B_{x,i+1}$ in Section 4.4, $|\bar{A}_{x,i+1}(s)| \le B_{x,i+1}$ for every $s \in \mathcal{S}$. The performance-difference identity, followed by adding and subtracting the old-policy occupancy term, gives

$$\begin{aligned} J_x(\pi_{i+1}) - J_x(\pi_i) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\pi_{i+1}}}\left[\bar{A}_{x,i+1}(s)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\pi_i}}\left[\bar{A}_{x,i+1}(s)\right] + \mathcal{R}_{x,i}, \end{aligned} \tag{55}$$

where the occupancy correction is

$$\mathcal{R}_{x,i} = \frac{1}{1-\gamma}\int_{\mathcal{S}} \bar{A}_{x,i+1}(s)\left(d_{\pi_{i+1}} - d_{\pi_i}\right)(ds). \tag{56}$$

For a finite signed measure σ on $\mathcal{S}$, let $|\sigma|$ denote its variation measure and define its total-variation norm by $\|\sigma\|_1 = |\sigma|(\mathcal{S})$. The preceding pointwise bound gives

$$|\mathcal{R}_{x,i}| \le \frac{B_{x,i+1}}{1-\gamma}\left\|d_{\pi_{i+1}} - d_{\pi_i}\right\|_1. \tag{57}$$

It remains to control this occupancy shift. For a policy π and a measurable state set $E \subseteq \mathcal{S}$, define the induced state transition kernel by

$$P_\pi(E \mid s) = \int_{\mathcal{H}} P(E \mid s,h)\pi(dh \mid s). \tag{58}$$

For such a signed measure σ, define its propagation under P_π by

$$\left(P_\pi^\top \sigma\right)(E) = \int_{\mathcal{S}} P_\pi(E \mid s)\,\sigma(ds). \tag{59}$$

Thus, $P_\pi^\top$ is the measure-propagation operator associated with the policy-induced transition kernel. As a Markov operator, it contracts total variation:

$$\left\|P_\pi^\top \sigma\right\|_1 \le \|\sigma\|_1. \tag{60}$$

Moreover, the data-processing inequality for total variation gives, for every $s \in \mathcal{S}$,

$$\begin{aligned} \left\|P_{\pi_{i+1}}(\cdot \mid s) - P_{\pi_i}(\cdot \mid s)\right\|_1 &= 2D_{\mathrm{TV}}\left(P_{\pi_{i+1}}(\cdot \mid s), P_{\pi_i}(\cdot \mid s)\right) \\ &\le 2D_{\mathrm{TV}}(\pi_{i+1}(\cdot \mid s), \pi_i(\cdot \mid s)) \le 2\Delta_i. \end{aligned} \tag{61}$$

Because d_{π_i} is a probability measure, this statewise bound implies

$$\begin{aligned} &\left\|\left(P_{\pi_{i+1}}^\top - P_{\pi_i}^\top\right) d_{\pi_i}\right\|_1 \\ &\quad\le \int_{\mathcal{S}} \left\|P_{\pi_{i+1}}(\cdot \mid s) - P_{\pi_i}(\cdot \mid s)\right\|_1 d_{\pi_i}(ds) \le 2\Delta_i. \end{aligned} \tag{62}$$

The normalized discounted occupancy obeys the fixed-point equation

$$d_\pi = (1-\gamma)\rho_0 + \gamma P_\pi^\top d_\pi. \tag{63}$$

Subtracting the equations for π_{i+1} and π_i gives

$$\begin{aligned} d_{\pi_{i+1}} - d_{\pi_i} &= \gamma P_{\pi_{i+1}}^\top\left(d_{\pi_{i+1}} - d_{\pi_i}\right) \\ &\quad + \gamma\left(P_{\pi_{i+1}}^\top - P_{\pi_i}^\top\right) d_{\pi_i}. \end{aligned} \tag{64}$$

Applying the two preceding bounds yields

$$\begin{aligned}\left\|d_{\pi_{i+1}} - d_{\pi_i}\right\|_1 &\leq \gamma \left\|d_{\pi_{i+1}} - d_{\pi_i}\right\|_1 \\ &\quad + \gamma \left\|\left(P_{\pi_{i+1}}^\top - P_{\pi_i}^\top\right) d_{\pi_i}\right\|_1 \\ &\leq \gamma \left\|d_{\pi_{i+1}} - d_{\pi_i}\right\|_1 + 2\gamma\Delta_i.\end{aligned} \tag{65}$$

Rearranging gives

$$\left\|d_{\pi_{i+1}} - d_{\pi_i}\right\|_1 \leq \frac{2\gamma}{1-\gamma}\Delta_i. \tag{66}$$

Substitution into the occupancy-correction bound yields

$$|\mathcal{R}_{x,i}| \leq \frac{2\gamma B_{x,i+1}}{(1-\gamma)^2}\Delta_i. \tag{67}$$

Finally, the definition of $\bar{A}_{x,i+1}$ gives

$$\mathbb{E}_{s\sim d_{\pi_i}}\left[\bar{A}_{x,i+1}(s)\right] = \mathbb{E}_{\substack{s\sim d_{\pi_i} \\ h\sim\pi_{i+1}(\cdot|s)}}\left[A_x^{\pi_i}(s,h)\right]. \tag{68}$$

Using $\mathcal{R}_{r,i} \geq -2\gamma B_{r,i+1}\Delta_i/(1-\gamma)^2$ for reward and $\mathcal{R}_{c,i} \leq 2\gamma B_{c,i+1}\Delta_i/(1-\gamma)^2$ for cost therefore proves both stated inequalities.

With $U_{c,i}$ denoting the cost-increase upper bound defined in Section 4.4, Proposition 2 gives $J_c(\pi_{i+1}) \leq J_c(\pi_i) + U_{c,i}$. If $J_c(\pi_i) \leq d - \zeta$ for a margin $\zeta > 0$ and $U_{c,i} \leq \zeta$, then

$$J_c(\pi_{i+1}) \leq J_c(\pi_i) + U_{c,i} \leq (d-\zeta) + \zeta = d. \tag{69}$$

This proves the sufficient feasibility-margin statement following Proposition 2.

B Training and Experimental Details

This section details the implementation of H-CVPO developed in Section 4 and summarized in Algorithm 1. It also specifies the tasks and shared training and evaluation settings used in Section 5.

B.1 H-CVPO Training Procedure

H-CVPO combines off-policy critic learning with the fixed-reference E/M update in Section 4. Algorithm 1 summarizes transition collection, critic and target-critic updates, and scheduled variational updates after the prescribed critic warm-up.

Fixed-reference cycle. Each E/M cycle uses a uniform reference batch $\{s_b\}_{b=1}^B$ drawn from replay, separately from the samples used for critic regression. This batch provides the empirical counterpart of d_i: the same batch averages are used in the E-step, M-step, and empirical KL checks, regardless of the critic sampling scheme. At *Freeze*, the current actor is copied to π_i, and the reference batch, π_i, and critic estimates Q_r^i, Q_c^i are held fixed throughout the cycle. The sampled particles remain fixed throughout the probe and any subsequent dual solve; the resulting target $\widehat{q}_i$ remains fixed during actor fitting. Thus, actor parameters may change in the M-step, but the reference policy does not. Throughout this subsection, Q_c^i denotes the frozen cost critic and plays the role of $\widehat{Q}_c^i$ in Equation (16).

E-step: joint target construction. For each reference state s_b and mode $m \in \mathcal{M}$, draw a positive integer number $K_{b,m}$ of continuous particles from the old conditional policy:

$$a_{b,m,j} \sim \pi_{i,c}(\cdot \mid s_b, m), \qquad j = 1, \ldots, K_{b,m}. \tag{70}$$

The corresponding hybrid candidate $h_{b,m,j} = (m, a_{b,m,j})$ is assigned the old-policy quadrature mass

$$\omega_{b,m,j}^i = \frac{\pi_{i,m}(m \mid s_b)}{K_{b,m}}. \tag{71}$$

The masses for a given mode sum to its old-policy probability, preserving the discrete marginal rather than replacing it with a uniform distribution. The *Draw* and *Form* steps therefore provide a quadrature rule for the hybrid sum–integral in Equation (9). The *Evaluate* step applies the frozen critics to every candidate.

The reward-oriented probe sets $\lambda = 0$; let $\widehat{C}_i^{(0)}$ denote the resulting batch-weighted cost used to test the E-step constraint. The algorithm uses the probe solution with $\widehat{\lambda}_i = 0$ when $\widehat{C}_i^{(0)} \le \rho\epsilon_{E,i}$, where $\rho \in (0,1)$ is a dead-zone fraction that places the probe acceptance threshold below $\epsilon_{E,i}$. Otherwise, the algorithm solves the bounded approximate dual over $\mathcal{D}_i$, yielding $(\widehat{\eta}_i, \widehat{\lambda}_i) \in \mathcal{D}_i$ for the reference batch. This bounded solve approximates Equation (10) on the domain in Equation (15). In either branch, $\widehat{\eta}_i$ controls the sharpness of reweighting; a positive $\widehat{\lambda}_i$ discounts costly candidates. The surrogate threshold $\epsilon_{E,i}$ is distinct from the episodic CMDP limit d. When enabled, the prior-support safeguard is applied only in the inactive-cost branch ($\widehat{\lambda}_i = 0$), before target construction. For each reference state s_b, the particle weights are normalized as

$$\widehat{w}^i_{b,m,j} = \frac{\omega^i_{b,m,j} \exp\left(\frac{Q^i_r(s_b,h_{b,m,j}) - \widehat{\lambda}_i Q^i_c(s_b,h_{b,m,j})}{\widehat{\eta}_i}\right)}{\sum_{m' \in \mathcal{M}} \sum_{j'=1}^{K_{b,m'}} \omega^i_{b,m',j'} \exp\left(\frac{Q^i_r(s_b,h_{b,m',j'}) - \widehat{\lambda}_i Q^i_c(s_b,h_{b,m',j'})}{\widehat{\eta}_i}\right)}. \tag{72}$$

These weights sum to one over (m, j) for each b. The *Construct* $\widehat{q}_i$ step uses them to approximate expectations under the joint target in Equation (11). The particles serve as quadrature points; the absolute-continuity requirement in Section 4.3 concerns the underlying target kernel, not the atomic empirical measure formed by the sampled points.

M-step: target validation and projection. Before projection, the algorithm evaluates empirical counterparts of the cost and target-to-old-policy KL residuals in Equation (16) on the fixed reference batch. The target must satisfy the finiteness and support conditions in Section 4.3, with residuals no larger than τ_c and τ_{KL}, respectively. If this admissibility test fails, the *Skip* step leaves the actor unchanged for the cycle. Selecting the reward-oriented probe does not bypass this test.

For an admissible target, the algorithm saves the current actor parameters and performs M_{M} updates using the empirical weighted log-likelihood in Equation (14). Both actor factors are fitted to the same fixed joint target. The *Measure* step then evaluates the old-to-new KL divergences in Equation (13) on the reference batch. The discrete term compares mode probabilities; the conditional-continuous term averages the mode-conditioned KL divergences under the old mode probabilities $\pi_{i,m}$, as required by Equation (12). The candidate is accepted only when all measured quantities are finite and both trust-region limits hold; otherwise, the saved actor parameters are restored. These post-update checks are distinct from the pre-projection target-admissibility test. The continuous mean and standard-deviation controls used in the experiments are reported in Table 3. The algorithm returns the policy at the predefined training endpoint, consistent with the checkpoint protocol in Section 5.1.

B.2 Experimental Details

Comparison protocol. We compare all methods empirically and separately discuss their theoretical guarantees. Under the stated assumptions, Theorem 1 establishes E-step strong duality for H-CVPO, while Section 4.4 provides fixed-reference surrogate improvement and conditional reward–cost bounds for its exact E/M updates. CHPO switches between reward and cost optimization; its analysis derives finite-iteration bounds on the policy-objective gap and the excess of expected cost over its threshold under its own assumptions (Zhou et al., 2025). We make no additional convergence or feasibility claims for HPPO-Lag or PDQN-Rco.

For the standard 320k-step experiments, every final-policy table evaluates the policy obtained at that fixed endpoint for each method and training seed, following Section 5.1. We do not substitute an earlier checkpoint on the basis of its reward or estimated feasibility. This keeps the interaction budget and checkpoint-selection rule matched across methods.

Following the task setup of CHPO (Zhou et al., 2025), we use constrained versions of *Moving*, *Sliding*, and *HardMove*, three hybrid-action benchmarks from DI-engine, together with the parameterized

Algorithm 1 Practical H-CVPO training procedure.

Require: Hybrid actor π_θ with factors $\pi_{\theta_m}, \pi_{\theta_c}$; reward and cost critics and their target copies; replay buffer $\mathcal{B}_{\text{rep}}$; reference batch size B; E-step limits $\epsilon_{E,i}, \epsilon_{\text{KL}}$; admissibility tolerances τ_c, τ_{KL}; M-step limits ϵ_m, ϵ_c; dual domain $\mathcal{D}_i$; dead-zone fraction $\rho \in (0, 1)$; particle counts $\{K_{b,m}\}$; M-step update count M_{M}
Ensure: Hybrid policy π_θ at the predefined final training step
Initialize the actor, critics, target critics, replay buffer, and optimization states
while the training interaction budget is not exhausted **do**
  Collect complete transitions $(s_t, h_t, r_t, c_t, s_{t+1})$ and termination flags using π_θ
  Append the transitions to $\mathcal{B}_{\text{rep}}$
  Update the reward and cost critics from replay
  Update the target critics by Polyak averaging
  if a variational update is scheduled and the prescribed critic warm-up is complete **then**
    Sample a uniform reference batch $\{s_b\}_{b=1}^{B}$ from $\mathcal{B}_{\text{rep}}$
    Freeze $\pi_i \leftarrow \pi_\theta$, the reference batch, and Q_r^i, Q_c^i for this E/M cycle
  E-step: construct the joint hybrid target
    for $b = 1, \ldots, B$ and each $m \in \mathcal{M}$ **do**
      Draw the particles in Equation (70)
      Form $h_{b,m,j} = (m, a_{b,m,j})$ and assign the masses in Equation (71)
      Evaluate $Q_r^i(s_b, h_{b,m,j})$ and $Q_c^i(s_b, h_{b,m,j})$
    end for
    Solve the reward-oriented probe with $\lambda = 0$ and compute its batch-weighted cost $\widehat{C}_i^{(0)}$
    If $\widehat{C}_i^{(0)} \le \rho\epsilon_{E,i}$, use the probe solution with $\widehat{\lambda}_i = 0$; otherwise solve the bounded approximate dual over $\mathcal{D}_i$
    When enabled, apply a prior-support safeguard within the inactive-cost branch ($\widehat{\lambda}_i = 0$)
    Construct $\widehat{q}_i$ using the weights in Equation (72)
  M-step: validate and project the joint target
    if $\widehat{q}_i \notin \mathcal{Q}_{i,\text{valid}}$ **then**
      Skip the projection and leave the actor unchanged
    else
      Save the current actor parameters
      Perform M_{M} updates of both actor factors by weighted joint maximum likelihood on the fixed target and reference batch
      Measure the empirical discrete and conditional-continuous old-to-new KL divergences
      if all measured quantities are finite and both trust-region limits hold **then**
        Accept the candidate
      else
        Restore the saved actor parameters
      end if
    end if
  end if
end while
return the policy at the predefined final training step

Parking task considered in that work. Each task pairs a discrete mode with a continuous control valid for that mode, consistent with Equation (3). The tasks differ in observation and action dimensions, dynamics, and cost definitions. Table 2 summarizes their interfaces, effective episode horizons, and primary experimental settings. The primary cost limits are denoted by d, as in Equation (2).

Table 2: Task interfaces and primary experimental settings.

Task	Observation dimension	Modes	Continuous action	Episode horizon	Primary d	Training steps
Moving	18	3	2-D, $[0, 1] \times [-1, 1]$	201	1	320,000
Sliding	20	3	2-D, $[0, 1] \times [-1, 1]$	201	1	320,000
HardMove	18	16	4-D, $[-1, 1]^4$, mode-masked	26	1	320,000
Parking	17	3	1-D, $[-1, 1]$	50	1	320,000

Moving. The agent must reach and stop within a target region while avoiding circular hazards. Each action selects acceleration, turning, or braking and includes $a_c \in [0, 1] \times [-1, 1]$. Acceleration uses the first continuous component, turning uses the second, and braking uses neither; inactive components do not affect actuation. Let δ_t denote the distance to the target. On ordinary transitions,

the reward $r_t = \delta_{t-1} - \delta_t$ measures progress toward the target. The per-step cost equals twice the number of circular hazard regions occupied by the agent. Success requires the agent to be within the target radius with zero speed. Episodes end on success, departure from the field, or expiry of the time limit.

Sliding. Sliding retains Moving's objective, three modes, and two-dimensional continuous control, but adds inertia: motion combines the vector induced by the current action with the previous velocity vector. The experimental observation has 20 dimensions, augmenting the underlying 18-dimensional observation with $(v_x, v_y) = (v\cos\phi, v\sin\phi)$, where v and ϕ denote speed and heading. Rewards and success conditions are unchanged from Moving, as are the out-of-bounds and time-limit conditions for ending an episode. The per-step cost, however, equals the number of circular hazard regions occupied by the agent, without Moving's factor of two.

HardMove. HardMove expands the discrete mode space through four binary actuators. Its 16 modes enumerate all on–off masks, paired with continuous control $a_c \in [-1, 1]^4$. The selected mask determines which continuous components are active; the others are set to zero and do not affect actuation. Motion is the sum of the displacement vectors generated by the active actuators. The agent must reach the target while avoiding circular hazards. Reward is based on distance reduction, with a success bonus and a negative terminal reward for failure or timeout. The per-step cost is ten times the number of circular hazard regions occupied by the agent. Reaching the target constitutes success; episodes end on success, departure from the field, or expiry of the time limit.

Parking. Parking requires perpendicular parking within prescribed position and heading tolerances. Its three modes are right, straight, and left, with $a_c \in [-1, 1]$ specifying travel under the selected maneuver. Right and left produce radius-5 arc motion through a kinematic center update; straight produces linear travel. Let d_{tar} denote the distance to the target, $\Delta\psi$ the heading error, and r_{target} the successful-parking indicator. The per-step reward is

$$r_t = 0.5\left(-\frac{d_{\mathrm{tar}}}{\sqrt{10^2 + 17.6^2}}\right) - \frac{|\Delta\psi|}{\pi} + 4r_{\mathrm{target}}. \tag{73}$$

Collision, overlap, and timeout penalties have zero weight in the reward. The per-step cost combines collision and proximity penalties with an overall scale factor of $1/10$; the proximity penalty activates when the obstacle distance is below 0.2. Let Δx and Δy denote target-relative position errors. Successful parking requires $|\Delta x| < 0.15$, $|\Delta y| < 0.35$, and $|\Delta\psi| < 10^\circ$. Success terminates the episode, whereas collision or proximity cost alone does not. Episodes that do not end in success are truncated after 50 steps.

Training configuration. All experiments are run on CPU using 32 vCPUs of an AMD EPYC 9654 processor and 60 GB of memory. Table 3 reports the H-CVPO network architecture and selected training settings for Algorithm 1. The mode-policy M-step limit is ϵ_m. Separate limits on the continuous mean and standard deviation implement the conditional-continuous trust-region checks represented abstractly by ϵ_c in Equation (13).

Table 3: H-CVPO network architecture and selected training settings.

Setting	Value
Policy network	$[128, 128]$, ReLU
Reward critic network	$[256, 256]$, ReLU
Cost critic network	$[256, 256]$, ReLU
Optimizer	Adam
Learning rates	actor 3×10^{-4}; reward and cost critics 10^{-3}
Critic / reference batch size	$256/256$
Replay capacity / critic warm-up	100,000 transitions / 5,000 environment steps
Discount factor	$\gamma = 0.99$
Target-network update rate	0.005
E-step KL radius	$\epsilon_{\mathrm{KL}} = 0.02$
M-step KL controls	mode $\epsilon_m = 0.01$; continuous mean 0.005; continuous standard deviation 0.0005

B.3 Additional Baseline Comparisons

We complement the main empirical comparison with three additional analyses. First, we present complete learning trajectories under the cost limits $d = 1.5$ and $d = 2.0$ to supplement the final-policy results. Second, we compare the fixed-endpoint results with performance averaged over the last five periodic evaluation checkpoints. These averages characterize late-stage training behavior across successive policies; they do not establish the feasibility of any single policy. Third, we evaluate H-CVPO and CHPO over an extended training horizon while varying the fraction of CHPO policy updates devoted to cost minimization, as specified by the C/A ratio. This analysis assesses CHPO's sensitivity to update allocation and whether the observed reward–cost differences between the two methods persist under a larger interaction budget.

B.3.1 Learning Performance under Additional Cost Limits

Figures 3 and 4 show the reward and cost learning curves under the additional cost limits $d = 1.5$ and $d = 2.0$, complementing the final-policy results in Table 1. We analyze the reward and cost trends shown in these two figures as follows.

Results at $d = 1.5$. In Figure 3, H-CVPO's mean reward settles near 1.3 on Moving after about 100k steps and fluctuates around 1.3–1.6 on Sliding. HardMove exceeds 2.0 around 110k steps and stays slightly above that level. Parking reaches 35.0–38.0 by 80–100k steps and later stays near 38.0 with some fluctuations. Later mean costs are mostly below $d = 1.5$ on Moving and HardMove, although Moving has spikes near 240k, 260k, and 280k steps. HardMove's cost is near zero around 210–260k steps. Sliding's cost falls from early peaks but still sometimes exceeds the limit. Parking's later cost is often 1.0–1.5, with spikes near 30.0.

All baselines have substantially lower late-stage rewards. CHPO's costs fluctuate above and below the limit on Moving and Sliding and remain above it on HardMove and Parking, with Parking near 150.0–160.0 later in training. HPPO-Lag's later costs stay below the limit on Moving, sometimes exceed it on Sliding and Parking, and frequently exceed it on HardMove. PDQN-Rco's later Moving cost is usually below the limit, with a spike near 250k steps; Sliding's cost fluctuates above and below the limit, and HardMove's often exceeds it. Parking has repeated large cost increases, including a peak near 100.0 around 200–210k steps.

Results at $d = 2.0$. In Figure 4, H-CVPO's mean reward on Moving rises from about 1.3 around 110k steps to 2.0 around 230–250k. Sliding and HardMove reach about 2.0 around 120k steps, with HardMove staying slightly higher. Parking reaches about 22.0 around 210k and exceeds 35.0 near the end. Mean costs on Moving and Sliding are at or below $d = 2.0$ throughout the second half of training; HardMove is mostly below it, with occasional spikes. Parking has early cost spikes above 50.0 and another near 230k, but after about 250k its cost stays below the limit, mostly around 1.0–1.5.

All baselines have substantially lower late-stage rewards. CHPO's Moving cost crosses $d = 2.0$, while Sliding stays above it over roughly the last 80k steps. HardMove costs range from about 5.0 to 15.0 over that interval, and Parking stays near 150.0–160.0 in the second half. HPPO-Lag's later costs are below the limit on Moving, mostly below it on Sliding, between 1.0 and 2.0 on HardMove over the last 80k, and near zero on Parking. PDQN-Rco's later costs are mostly below the limit on Moving and Sliding but often above it on HardMove. Parking's reward increase is not sustained; costs repeatedly exceed 100.0 and stay above 2.0 over roughly the last 80k steps.

Overall, under both cost limits, H-CVPO achieves higher late-stage mean rewards than the baselines across all four tasks. Its mean costs are often below the corresponding limits later in training, although some tasks still show cost spikes and exceed the limits at certain checkpoints. The baselines have lower rewards without consistently better constraint control.

B.3.2 Late-Stage Performance Trends

To assess whether the final checkpoint reflects nearby training behavior, we average reward and cost over the five periodic evaluation checkpoints at 294.4k, 300.8k, 307.2k, 313.6k, and 320k environment steps for each training seed, then report the mean and standard deviation across three seeds. These

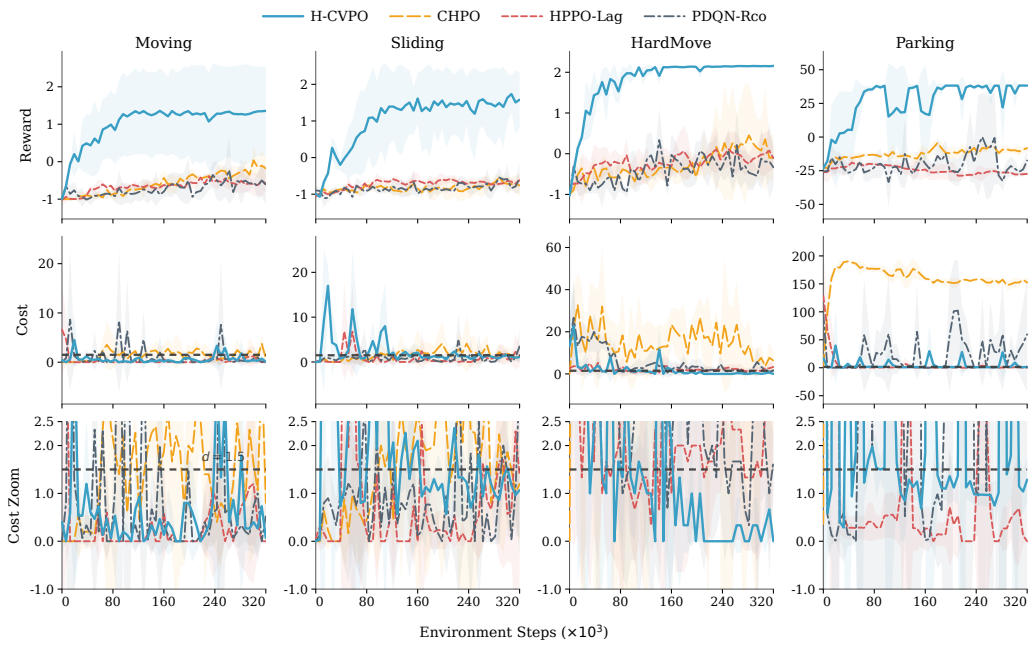

Figure 3: Learning curves for H-CVPO and the baselines under the cost limit $d = 1.5$. Lines and shaded regions show the mean and one standard deviation over the three training seeds. Rows report reward, cost on the full scale, and cost near the constraint boundary; horizontal dashed lines mark the cost limit.

Table 4: Late-stage performance over the final five periodic evaluation checkpoints (294.4k–320k environment steps). For each training seed, we average the five periodic evaluations, each computed from 10 deterministic episodes with fixed evaluation seeds 21000–21009, and report the mean and sample standard deviation across the three training seeds.

		H-CVPO		CHPO		HPPO-Lag		PDQN-Rco	
Environment	d	Reward	Cost	Reward	Cost	Reward	Cost	Reward	Cost
Moving	1.0	2.02±0.01	0.69±0.29	-0.14±0.19	1.03±0.85	-0.67±0.05	0.37±0.65	-0.62±0.44	0.48±0.49
	1.5	1.33±1.22	0.35±0.10	-0.15±0.18	1.91±1.24	-0.58±0.30	0.55±0.95	-0.58±0.20	0.48±0.25
	2.0	1.95±0.05	0.72±0.84	-0.15±0.18	1.91±1.24	-0.50±0.14	0.04±0.07	-0.30±0.45	0.35±0.18
Sliding	1.0	1.20±0.57	3.21±4.35	-0.62±0.17	2.42±3.47	-0.68±0.14	0.66±0.94	-0.88±0.13	0.55±0.30
	1.5	1.57±0.79	1.21±0.55	-0.72±0.21	1.38±1.00	-0.67±0.10	1.31±2.14	-0.67±0.14	1.59±0.46
	2.0	2.02±0.00	0.97±0.37	-0.73±0.13	3.71±1.53	-0.55±0.01	0.85±0.43	-0.66±0.11	0.95±0.36
HardMove	1.0	2.14±0.00	0.60±0.60	0.01±0.85	10.40±4.33	0.24±0.86	2.53±1.70	-0.78±0.21	1.73±1.15
	1.5	2.15±0.01	0.27±0.23	0.07±0.77	6.13±4.71	-0.03±0.40	2.33±2.14	-0.28±0.37	2.73±2.50
	2.0	2.15±0.01	0.27±0.46	0.09±0.78	9.67±6.52	-0.23±0.18	1.60±1.04	-0.66±0.30	3.00±1.91
Parking	1.0	22.63±27.00	1.15±0.19	-9.59±1.85	152.82±4.22	-28.55±1.57	0.00±0.00	-4.61±37.20	0.56±0.60
	1.5	37.05±2.06	2.88±3.24	-9.59±1.85	152.82±4.22	-27.42±1.06	0.37±0.51	-22.47±4.15	35.62±50.98
	2.0	28.54±16.80	1.16±0.51	-9.59±1.85	152.82±4.22	-27.07±1.13	0.00±0.00	-24.49±3.27	54.28±58.51

averages combine different policies and describe late-stage performance, not the feasibility of a single deployed policy.

Table 4 shows that H-CVPO has the highest mean reward in all 12 task–limit settings, with mean costs below the limits in nine. On HardMove, rewards are 2.14–2.15 and costs are 0.27–0.60 across all three limits. Three settings differ from the final-policy evaluation in whether mean cost exceeds the limit. On HardMove at $d = 1.5$, the late-stage mean cost is 0.27, below the limit, while the final-policy mean cost is 2.17, above it. On Sliding at $d = 1.0$ and Parking at $d = 1.5$, the late-stage mean costs are 3.21 and 2.88, respectively, both above the limits. Their corresponding final-policy mean costs are 0.67 and 0.85, both below the limits. Parking at $d = 1.0$ remains above the limit in

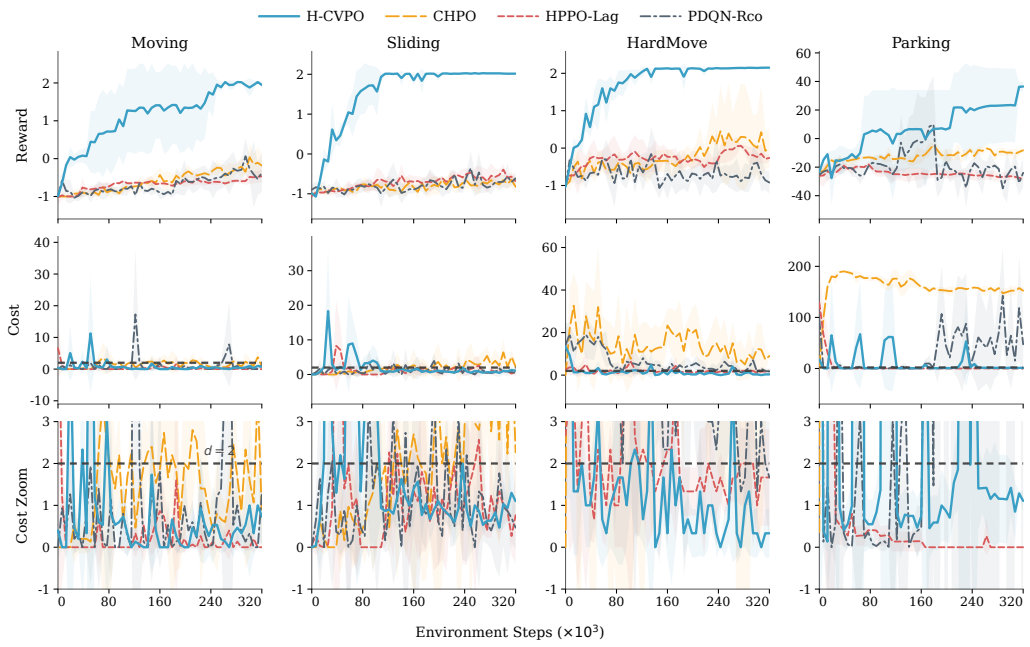

Figure 4: Learning curves for H-CVPO and the baselines under the cost limit $d = 2.0$. Lines and shaded regions show the mean and one standard deviation over the three training seeds. Rows report reward, cost on the full scale, and cost near the constraint boundary; horizontal dashed lines mark the cost limit.

both summaries. The cost standard deviations of 4.35 on Sliding at $d = 1.0$ and 3.24 on Parking at $d = 1.5$ also show substantial differences across training seeds.

The broader reward–cost pattern remains similar to the final-policy results. CHPO's mean costs are 6.13–10.40 on HardMove and 152.82 on Parking, above every tested limit. HPPO-Lag stays below the limits on Moving, Sliding, and Parking, but all its mean rewards on these tasks are negative; its HardMove costs exceed $d = 1.0$ and 1.5. PDQN-Rco stays below all three limits on Moving, but exceeds them on HardMove. Its Parking costs at $d = 1.5$ and 2.0 are 35.62 and 54.28, respectively, with negative mean rewards. Both Table 1 and Table 4 show the same pattern: within each table, CHPO's reported reward and cost means and standard deviations are identical on Moving at $d = 1.5$ and 2.0, and on Parking across all three cost limits.

Overall, the late-stage averages support the main reward–cost comparison beyond the final checkpoint alone. They include the final policy but also average over four earlier checkpoints and use different evaluation episodes. Differences between the two summaries therefore do not reflect training changes alone. The final-policy evaluation remains the basis for assessing policy feasibility.

B.3.3 Extended-Training Comparison with CHPO

We examine whether the reward–cost differences between H-CVPO and CHPO persist when both methods are trained for more steps and CHPO's update allocation is varied. In CHPO, the C/A ratio is the fraction of policy-update iterations devoted to cost minimization and therefore controls the reward–cost update balance (Zhou et al., 2025). We extend both methods to 960k environment steps under the primary cost limit $d = 1.0$ and evaluate CHPO with C/A ratios $1/10$, $1/5$, $1/4$, and $1/3$, where $1/4$ is used for CHPO in the main-text comparison in Figure 1. At 320k steps, these runs recover the corresponding trajectories from the main experiment and thus provide direct continuations of the matched-budget comparison.

Figure 5 shows that the separation between H-CVPO and CHPO persists as training is extended. On Moving, H-CVPO reaches a reward near 2.0 within roughly the first 120k steps and maintains this level thereafter; following the initial transients, its mean cost fluctuates mostly around or below

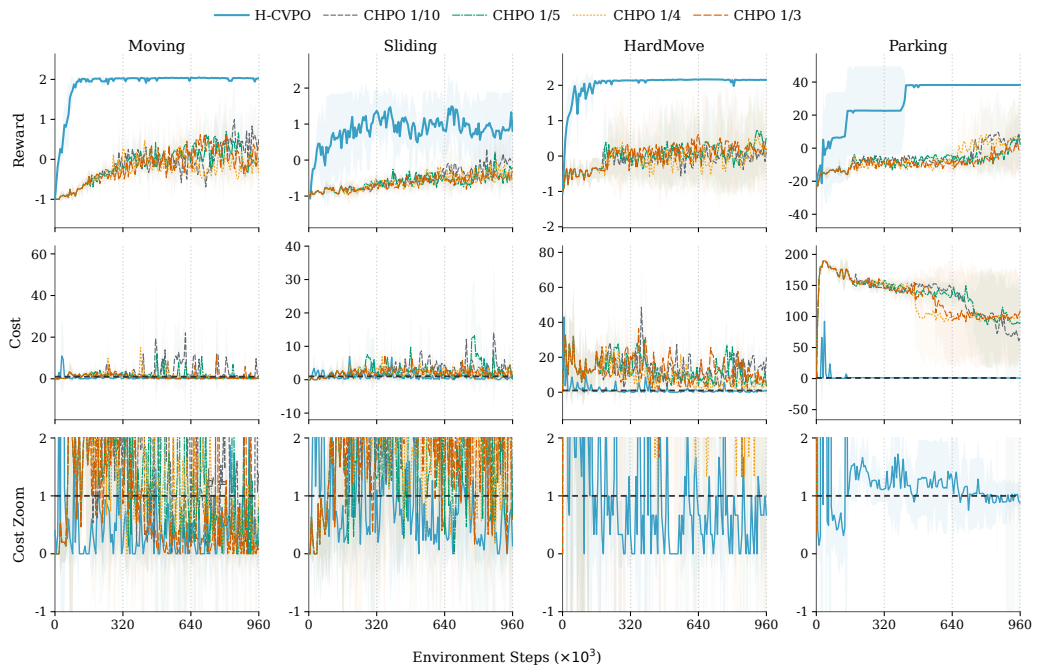

Figure 5: Extended-budget comparison between H-CVPO and CHPO under $d = 1.0$ over 960k environment steps. CHPO uses C/A ratios $1/10$, $1/5$, $1/4$, and $1/3$. Lines and shaded regions show the mean and one standard deviation over the three training seeds. Rows report reward, cost on the full scale, and cost near the constraint boundary; vertical dotted lines mark 320k, 640k, and 960k steps, and horizontal dashed lines mark the cost limit.

the constraint boundary. HardMove follows a similar pattern, approaching a reward of 2.0 by approximately 160k steps and remaining slightly above 2.0 through 960k, while its mean cost declines toward the boundary over the first 200k steps and stays comparatively low thereafter.

Parking improves in two distinct stages: its reward first rises to about 22.0 by roughly 160k steps, then increases to approximately 38.0 around 420k and remains near that level. Its mean cost settles around 1.2–1.5 over much of the middle training period before approaching 1.0 near the end. Sliding exhibits a different trade-off. Its reward rises to approximately 1.0 by 240k steps and reaches roughly 1.3 around 320k, before fluctuating mostly between 0.6 and 1.4. Despite substantial variation, its mean cost falls from about 1.1 at 320k to near 0.1 at 960k.

The CHPO trajectories vary substantially with the C/A ratio, and no ratio provides a consistent reward–cost advantage across the four tasks. On Moving and Sliding, some configurations exhibit late-stage mean costs that oscillate around the constraint boundary, but their rewards remain below those of H-CVPO. On HardMove and Parking, all four configurations maintain mean costs well above the limit through 960k steps despite gradual reward improvement, with the largest and most persistent gap occurring on Parking.

Overall, extending training from 320k to 960k steps and varying the C/A ratio do not eliminate the performance differences observed at 320k. The extended trajectories instead show distinct long-horizon behavior: H-CVPO sustains high reward as mean cost trends toward or below the constraint boundary, whereas CHPO remains sensitive to its update allocation and retains substantial costs on HardMove and Parking.

B.4 Ablation Studies

We examine three components of H-CVPO: cost-aware E-step reweighting, separate hybrid KL controls, and target-admissibility filtering. Following Section 5.3, we compare full H-CVPO with

Table 5: Final-policy performance of H-CVPO and its three single-component ablations across cost limits. The ablations are A1 (w/o Cost-Aware E-Step), A2 (Joint Hybrid KL), and A3 (w/o Target Admissibility). Entries report the mean and sample standard deviation across training seeds 11, 22, and 33, with each final policy evaluated over 40 deterministic episodes. † marks a mean episodic cost above the corresponding cost limit.

		H-CVPO		A1		A2		A3	
Environment	d	Reward	Cost	Reward	Cost	Reward	Cost	Reward	Cost
Moving	1.0	2.07 ± 0.03	0.85 ± 0.97	1.88 ± 0.29	$1.73 \pm 0.68^{\dagger}$	2.01 ± 0.04	0.43 ± 0.29	2.07 ± 0.03	0.85 ± 0.97
Moving	1.5	1.41 ± 1.18	1.07 ± 0.60	2.02 ± 0.13	$6.58 \pm 9.32^{\dagger}$	2.06 ± 0.04	1.33 ± 0.97	1.41 ± 1.18	1.07 ± 0.60
Moving	2.0	2.08 ± 0.02	1.00 ± 0.43	1.97 ± 0.15	0.87 ± 0.93	2.07 ± 0.03	1.78 ± 0.51	2.08 ± 0.02	1.00 ± 0.43
Sliding	1.0	1.08 ± 0.61	0.67 ± 0.47	1.31 ± 1.32	0.52 ± 0.45	1.71 ± 0.44	0.91 ± 0.12	1.08 ± 0.61	0.67 ± 0.47
Sliding	1.5	1.55 ± 0.85	0.68 ± 0.30	0.66 ± 1.28	$4.99 \pm 7.41^{\dagger}$	1.15 ± 0.81	0.55 ± 0.42	1.55 ± 0.85	0.68 ± 0.30
Sliding	2.0	1.99 ± 0.10	0.84 ± 0.06	0.56 ± 1.34	$2.14 \pm 1.90^{\dagger}$	1.96 ± 0.04	0.83 ± 0.13	1.99 ± 0.10	0.84 ± 0.06
HardMove	1.0	2.05 ± 0.00	0.42 ± 0.52	1.17 ± 1.50	$3.25 \pm 4.99^{\dagger}$	2.04 ± 0.01	$1.42 \pm 0.95^{\dagger}$	2.05 ± 0.00	0.42 ± 0.52
HardMove	1.5	2.06 ± 0.01	$2.17 \pm 3.32^{\dagger}$	0.31 ± 1.45	$9.17 \pm 8.43^{\dagger}$	1.97 ± 0.14	0.50 ± 0.50	2.06 ± 0.01	$2.17 \pm 3.32^{\dagger}$
HardMove	2.0	2.05 ± 0.02	0.50 ± 0.43	0.30 ± 1.51	$13.00 \pm 11.92^{\dagger}$	1.98 ± 0.11	0.75 ± 0.66	2.05 ± 0.02	0.50 ± 0.43
Parking	1.0	23.26 ± 26.94	$1.06 \pm 0.16^{\dagger}$	-18.30 ± 4.46	$51.39 \pm 64.32^{\dagger}$	38.81 ± 0.02	0.78 ± 0.19	23.27 ± 26.95	$1.07 \pm 0.15^{\dagger}$
Parking	1.5	38.82 ± 0.01	0.85 ± 0.28	-18.64 ± 5.31	$51.75 \pm 72.03^{\dagger}$	38.83 ± 0.00	0.75 ± 0.27	38.82 ± 0.01	0.78 ± 0.27
Parking	2.0	35.92 ± 5.03	1.23 ± 0.13	-16.78 ± 6.24	$108.05 \pm 32.27^{\dagger}$	38.81 ± 0.02	1.25 ± 0.38	23.31 ± 26.87	0.97 ± 0.57

three variants that respectively remove the cost term from the E-step target, replace the separate discrete and conditional-continuous M-step KL constraints with a single joint KL constraint at the same total KL allowance, or disable the pre-projection target-admissibility filter. Each variant changes only one component and is evaluated under $d \in \{1.0, 1.5, 2.0\}$, with all other training and evaluation settings unchanged. Figures 6, 7, and 8 show the full reward and cost learning curves. Table 5 summarizes final-policy performance across the three cost limits.

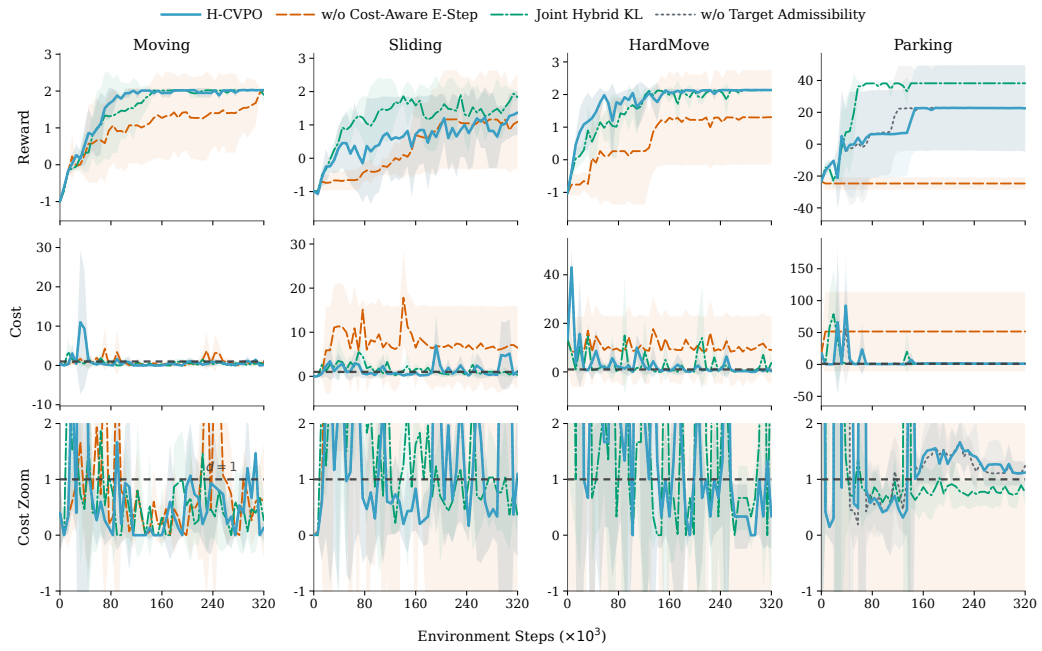

Figure 6: Complete ablation learning curves under the cost limit $d = 1.0$. The variants remove the cost contribution from the E-step target, replace the separate discrete and conditional-continuous KL controls with a single joint KL constraint, or disable target-admissibility filtering. Curves and shaded regions show the mean and one standard deviation, respectively, over three training seeds; dashed lines mark the cost limit.

Cost-aware E-step reweighting. The variant labeled *w/o Cost-Aware E-Step* in the figures removes only the cost term $-\lambda_i^{\star} Q_c^i(s, m, a_c)$ from the E-step target in Equation (11), corresponding to the

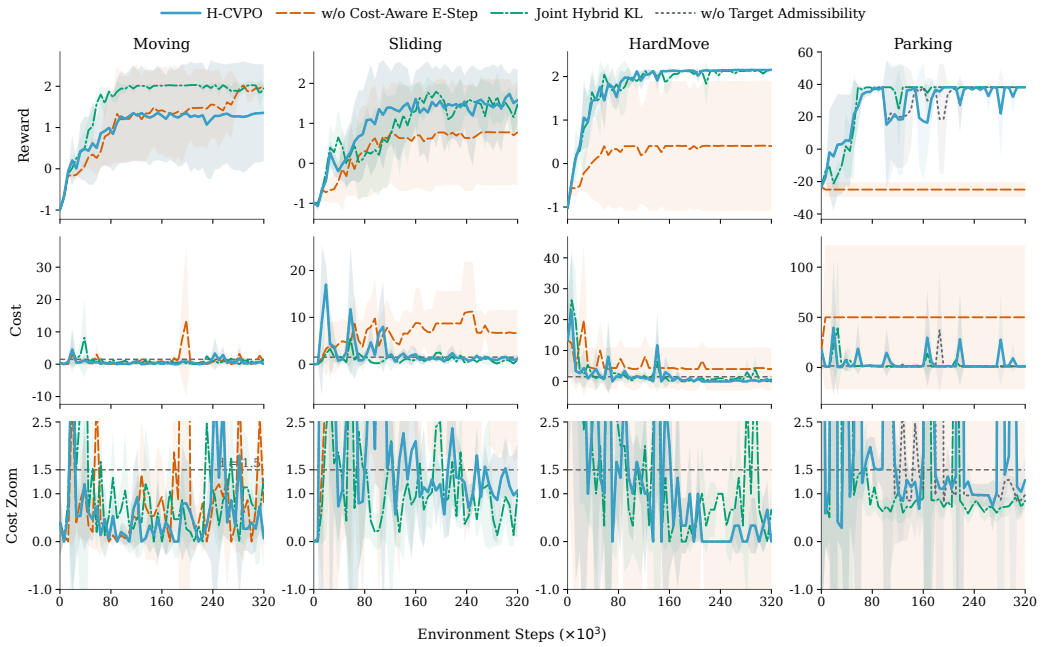

Figure 7: Complete ablation learning curves under the cost limit $d = 1.5$. The variants remove the cost contribution from the E-step target, replace the separate discrete and conditional-continuous KL controls with a single joint KL constraint, or disable target-admissibility filtering. Curves and shaded regions show the mean and one standard deviation, respectively, over three training seeds; dashed lines mark the cost limit.

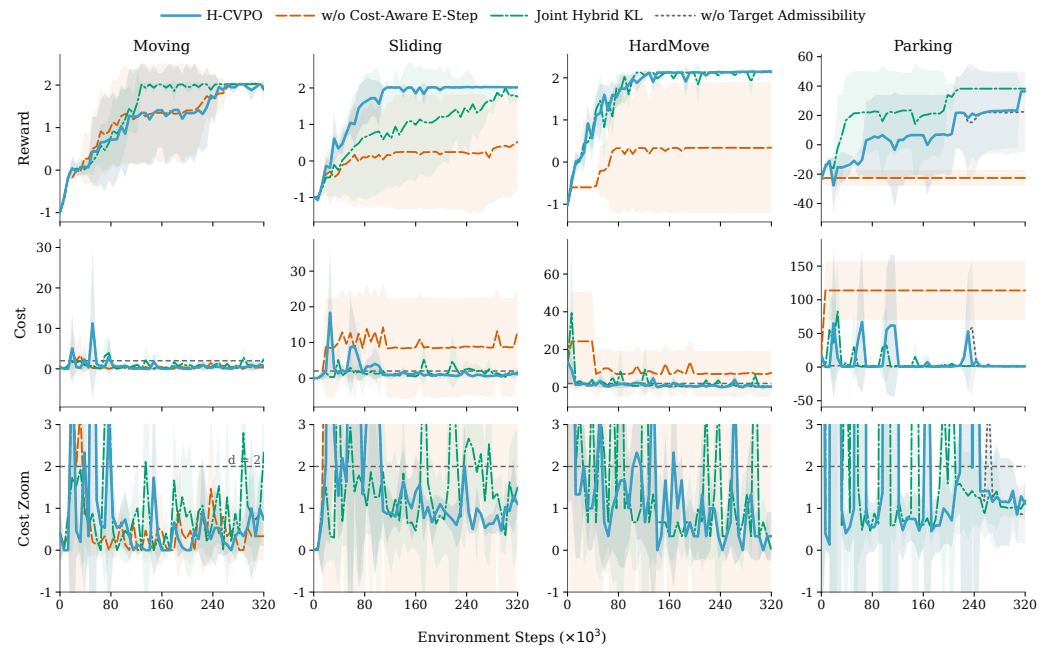

Figure 8: Complete ablation learning curves under the cost limit $d = 2.0$. The variants remove the cost contribution from the E-step target, replace the separate discrete and conditional-continuous KL controls with a single joint KL constraint, or disable target-admissibility filtering. Curves and shaded regions show the mean and one standard deviation, respectively, over three training seeds; dashed lines mark the cost limit.

target-construction step in Algorithm 1; the E-step itself is retained. Figures 6, 7, and 8 show task-dependent effects of removing cost-aware reweighting. On Moving, the ablation learns more slowly at $d = 1.0$ but reaches a mean reward of about 1.9 at the final checkpoint. At $d = 1.5$, the reward curves remain close over much of training, with the ablation reaching a higher mean reward near the end. At $d = 2.0$, both methods reach similar reward levels near the end. By contrast, on Sliding at $d = 2.0$, the ablation's mean reward rises to about 0.5 near the end, compared with about 2.0 for full H-CVPO. On HardMove at $d = 1.5$ and 2.0, the ablation's mean reward plateaus below 0.5, whereas full H-CVPO reaches about 2.1. On both tasks, the ablation's mean cost remains above the corresponding limit late in training for all three cost limits. Parking shows the clearest deterioration: the ablation's mean reward remains negative, while its late-stage mean cost stays around 50 at $d = 1.0$ and 1.5 and above 100 at $d = 2.0$. Full H-CVPO achieves positive rewards and much lower costs on Parking, although cost spikes and periods above the limits remain.

Table 5 reports the separate final-policy evaluation: the ablation exceeds the cost limit in ten of the twelve settings, compared with two for full H-CVPO. Eight settings change from below-limit to above-limit mean costs: Moving at $d = 1.0$ and 1.5, Sliding at $d = 1.5$ and 2.0, HardMove at $d = 1.0$ and 2.0, and Parking at $d = 1.5$ and 2.0. Where full H-CVPO already exceeds the limit, the ablation further raises mean cost from 2.17 to 9.17 on HardMove at $d = 1.5$ and from 1.06 to 51.39 on Parking at $d = 1.0$. Moving at $d = 2.0$ and Sliding at $d = 1.0$ remain below the limit under both methods. The ablation's final-policy mean cost on Sliding at $d = 1.0$ is 0.52, despite its above-limit late-stage curve. The evaluations use different episodes; final-policy feasibility is assessed from the table. These results support cost-aware reweighting as an important contributor to constraint control, although removing it does not lead to a final-policy cost-limit violation in every setting.

Separate hybrid KL controls. Figures 6, 7, and 8 show mixed effects of replacing the separate controls with a joint KL constraint at the same total allowance. On Moving at $d = 1.5$ and 2.0, the joint-KL variant reaches a mean reward of about 2.0 earlier than full H-CVPO. At $d = 1.5$, its mean reward remains higher through the end of training. At $d = 2.0$, full H-CVPO approaches the joint-KL variant's mean reward near the end, leaving both at similar levels. On Sliding, the variant generally has higher rewards at $d = 1.0$, whereas at $d = 2.0$ its reward improves more slowly and remains below full H-CVPO for most of training. HardMove shows broadly similar reward curves under both methods across the three limits, with cost spikes still present. On Parking at $d = 1.0$ and 2.0, the joint-KL variant reaches positive rewards earlier and maintains higher rewards over much of training. At $d = 1.5$, both methods reach similar reward levels, although full H-CVPO exhibits larger temporary declines in reward.

Table 5 reports the separate final-policy evaluation. Mean costs remain below the limits under both methods in nine settings. On HardMove at $d = 1.0$, the joint-KL variant raises mean cost from 0.42 to 1.42, exceeding the limit. In contrast, it lowers mean cost from 2.17 to 0.50 on HardMove at $d = 1.5$ and from 1.06 to 0.78 on Parking at $d = 1.0$, bringing both below their limits. On Parking at $d = 1.0$, mean reward also rises from 23.26 to 38.81. On Sliding at $d = 1.5$, however, mean reward falls from 1.55 to 1.15, while mean cost decreases from 0.68 to 0.55. Overall, the final-policy mean cost is below the limit in eleven of the twelve settings for the joint-KL variant and ten for full H-CVPO. Thus, the joint budget meets the mean-cost limit in one more setting, but this does not imply higher rewards or lower costs in every setting. Neither design achieves both higher reward and lower cost across all settings; their relative performance depends on the task and cost limit.

Target-admissibility filtering. H-CVPO assesses the approximate target $\widehat{q}_i$ using the cost and KL residuals in Equation (16) and skips its projection when $\widehat{q}_i \notin \mathcal{Q}_{i,\mathrm{valid}}$. The ablation removes only this pre-projection filter. The ablation also retains the post-update checks in Algorithm 1 for finiteness of the measured quantities and compliance with the discrete and conditional-continuous M-step KL limits. Across Figures 6, 7, and 8, the reward and cost curves for Moving, Sliding, and HardMove overlap between the full method and this ablation; visible differences appear only on Parking. At $d = 2.0$, the Parking reward curves follow similar trajectories, but separate near the end: full H-CVPO rises to a higher reward level while the ablation remains near its preceding plateau. The cost curves still show spikes, so retaining the filter does not eliminate cost-limit violations during training.

For the final policies, both methods have identical reported reward and cost statistics on Moving, Sliding, and HardMove, as shown in Table 5. On Parking at $d = 1.0$, removing the filter changes

mean reward only from 23.26 to 23.27 and mean cost from 1.06 to 1.07, with both costs above the limit. At $d = 1.5$, mean reward remains 38.82, while mean cost decreases from 0.85 to 0.78. The largest reward difference occurs at $d = 2.0$: removing the filter lowers mean reward from 35.92 to 23.31 and mean cost from 1.23 to 0.97, with both costs below the limit. These results show a reward benefit from retaining the filter on Parking at $d = 2.0$, but not a general reduction in final-policy mean cost. Its role is to check targets before projection, not to guarantee better performance.