UNIQ: UNIFIED CONDITIONAL FLOWS FOR OFFLINE SAFE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline safe reinforcement learning seeks to maximize reward under cost constraints using a fixed dataset. Lagrangian methods repeatedly adjust a multiplier, changing the policy objective and potentially destabilizing learning when cost estimates are inaccurate. We propose UNIQ, a conditional flow framework that replaces continued multiplier adaptation with a fixed safety rule estimated from offline data. The rule prioritizes reward in safe regions, balances reward and cost near the safety boundary, and prioritizes cost reduction in unsafe regions. Conditioned on state and accumulated cost, a single flow critic learns the combined return and guides a conditional flow policy, with both models conditioned on the physical state and accumulated cost. To train these models across a wider range of budget conditions, we resample cumulative costs for each logged transition. Our analysis establishes population Bellman consistency and fixed-policy convergence under vanishing update error, and identifies the utility-change term eliminated by the fixed safety rule. A capacity-restricted construction further demonstrates a strict value-error advantage over separately fitted component critics. On nine MetaDrive tasks, UNIQ satisfies the cost threshold on every task and achieves approximately 48% higher average normalized reward than the strongest baseline.

Anonymous Project Page & Videos

1 INTRODUCTION

Offline safe reinforcement learning (OSRL) seeks to maximize reward subject to a cost constraint using only logged interactions (Le et al., 2019; Xu et al., 2022; Liu et al., 2023a). A common approach learns reward and cost values separately and uses a Lagrange multiplier to balance them during policy improvement (Le et al., 2019; Wu et al., 2021; Hong et al., 2024). This couples two learning problems: the policy responds to the current penalty, while the penalty responds to the policy's estimated cost. These updates can oscillate and depend sensitively on multiplier initialization and learning rates (Stooke et al., 2020; Xu et al., 2021). Offline data make this interaction harder to control. A learned policy may choose actions that are poorly represented in the dataset (Levine et al., 2020; Xu et al., 2022; Polosky et al., 2022), and errors in their estimated costs can drive the multiplier toward insufficient penalties or excessive conservatism (Lee et al., 2022; Liu et al., 2023b).

We propose UNIQ to remove continued multiplier adaptation from the final training stage. Motivated by region-based offline safe RL (Zheng et al., 2024; Zhu et al., 2026; Koirala et al., 2024), UNIQ determines a safety rule using an offline dataset. Recovery states prioritize cost reduction; safe and boundary states undergo an additional accumulated-cost check to determine their reward–cost priorities. We then freeze the complete rule. Priorities can differ across states and cost histories while the mapping to the learning objective remains fixed.

UNIQ learns this objective directly with one flow critic and one conditional flow policy. Conditional flow matching provides a regression-based construction for both expressive action distributions (Lipman et al., 2023; Park et al., 2025) and temporal-difference value learning (Agrawalla et al., 2026). The critic fits the combined reward–cost Bellman target and uses its predicted mean as the value estimate. The policy learns from logged actions weighted by their unified advantage (Zhang et al., 2025; Feng et al., 2025). Both models condition on accumulated cost to represent different cost histories at the same physical state. To broaden their coverage beyond logged histories, we augment

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.

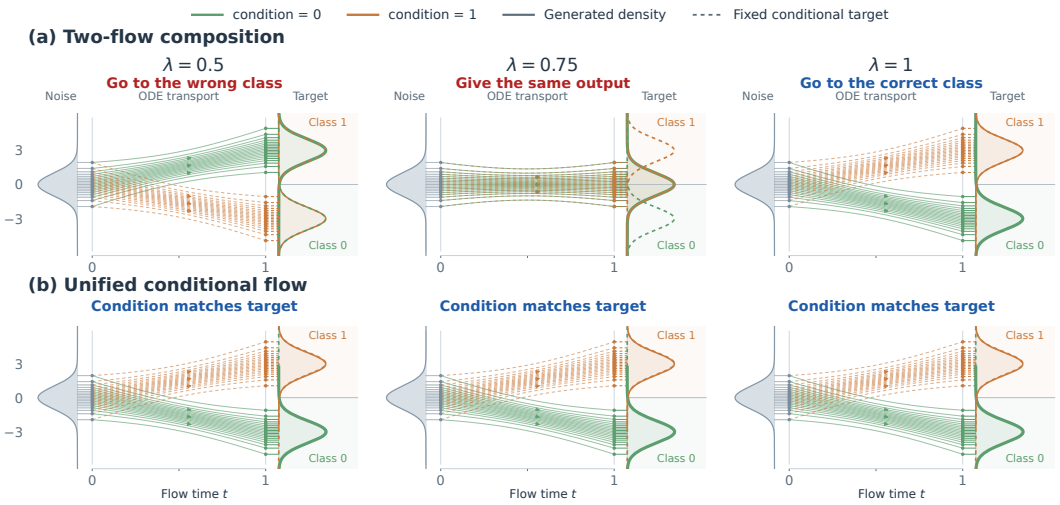


Figure 1: **Two-flow composition and direct conditional learning.** The two conditional targets remain fixed. In (a), λ is the composition weight: $\lambda = 0.5$ exchanges the condition assignments, $\lambda = 0.75$ makes the generated distributions coincide, and $\lambda = 1$ matches the targets. In (b), a unified conditional flow matches both targets directly; the same result is repeated across columns for comparison. Colors identify conditions, and solid and dashed density curves show generated and target distributions, respectively.

each transition with sampled cumulative costs and consistent successors; execution uses the actual cumulative cost.

Our theoretical results show that population flow matching recovers the intended Bellman update, and freezing the rule removes the utility-change term from critic tracking. Direct critic fitting also has a distinct approximation benefit: in a common capacity-restricted flow class, a single critic recovers the combined value exactly while separately optimal reward and cost critics fail to recover their difference (Section 4).

Figure 1 illustrates the calibration required by two-flow composition. Although the conditional targets remain fixed, changing the composition weight can reverse their assignments or make distinct conditions generate the same distribution. A calibrated weight recovers the targets, but exact component flows alone do not ensure this correspondence. UNIQ instead trains a single conditional flow directly on the desired targets. Appendix G establishes the related limitation for our regularized actor objective: additive composition of exact reward and safety flows need not generate the policy defined by the combined value. The error arises from composition itself, motivating direct policy learning from the unified value.

Our main contributions are:

- **Unified value and policy learning.** We introduce UNIQ, which trains a flow critic and a conditional flow policy under a fixed safety rule. A cumulative-cost resampling procedure broadens the conditions represented during training while preserving consistent successor costs, without further environment interaction.
- **Theoretical support for direct learning under a fixed rule.** We establish Bellman consistency, fixed-policy convergence under vanishing update error, and a critic tracking bound that isolates the term eliminated by freezing the rule. We also prove a strict approximation advantage for direct critic fitting in a common capacity-restricted flow class (Section 4).
- **Higher reward under the evaluated cost constraint.** Across nine MetaDrive tasks from DSRL (Liu et al., 2023a; Li et al., 2022), UNIQ satisfies the reported cost threshold on every task and achieves approximately 48% higher average normalized reward than FISOR, the strongest baseline feasible on all nine tasks (Table 1).

Related Work. Offline safe RL includes primal–dual methods (Le et al., 2019; Hong et al., 2024), feasibility-based learning (Zheng et al., 2024; Koirala et al., 2024; Zhu et al., 2026), and budget-conditioned policies (Liu et al., 2023b; Guo et al., 2025). Flow matching has also been used for

offline policies, critics, and safe policy learning (Park et al., 2025; Agrawalla et al., 2026; Tayal et al., 2026; Tayal & Tayal, 2026). UNIQ combines a fixed safety rule with direct estimation of its combined Bellman target and conditional policy learning from that value. Our analysis studies the resulting critic tracking and approximation properties. Appendix A compares the objectives, conditioning variables, and learning procedures in detail.

2 PROBLEM SETUP AND MOTIVATION

Offline Safe RL. We consider a discounted constrained Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, c, \rho, \gamma, \ell)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $r(s, a)$ is the reward function, $c(s, a)$ is the safety cost function, ρ is the initial state distribution, $\gamma \in (0, 1)$ is the discount factor, and $\ell > 0$ is the cost limit. We assume that r is bounded and $0 \le c(s, a) \le C_{\max}$. For a policy π, its expected discounted reward and cost are

$$J_h(\pi) = \mathbb{E}_{\rho,P,\pi}\left[\sum_{j=0}^{\infty} \gamma^j h(s_j, a_j)\right], \qquad h \in \{r, c\}. \tag{1}$$

The objective is

$$\max_{\pi} J_r(\pi) \qquad \text{subject to} \qquad J_c(\pi) \le \ell. \tag{2}$$

In offline RL, the learner has access only to a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, c_i, s_i')\}_{i=1}^{N}$ collected before training. It cannot query the transition kernel or collect additional environment interactions. The task is therefore to learn a rewarding policy whose expected cost satisfies the constraint, using the same logged data to estimate both quantities.

Lagrangian learning with separate reward and cost values. A common approach to solving the constrained optimization problem (2) introduces a nonnegative multiplier and optimizes the Lagrangian (Le et al., 2019; Altman, 1999)

$$\mathcal{L}(\pi, \lambda) = J_r(\pi) - \lambda\big(J_c(\pi) - \ell\big). \tag{3}$$

For a fixed Markov policy, define the component action values

$$Q_h^{\pi}(s, a) = \mathbb{E}_{\pi}\left[\sum_{j=0}^{\infty} \gamma^j h(s_j, a_j) \,\middle|\, s_0 = s, a_0 = a\right], \qquad h \in \{r, c\}. \tag{4}$$

The policy can then be improved using the combined Q functions $Q_r^{\pi}(s, a) - \lambda Q_c^{\pi}(s, a)$. An implementation typically fits reward and cost critics (Q functions) separately and adjusts the multiplier using an estimated policy cost. For example, at learning iteration k, $\lambda_{k+1} = [\lambda_k + \eta_\lambda\big(\widehat{J}_c(\pi_k) - \ell\big)]_+, [u]_+ = \max\{u, 0\}$, where $\eta_\lambda > 0$ is a step size. A cost estimate above the limit increases the penalty; an estimate below the limit decreases it. Policy learning and multiplier adaptation thus proceed together: the critics estimate the components, and the multiplier determines how the policy combines them.

We focus on two challenges that motivate UNIQ. First, multiplier updates change the policy's action scores even when the critic estimates remain fixed. Inaccurate cost estimates can drive these updates toward excessive or insufficient penalties. Appendix E.3 quantifies the resulting change in the regularized policy target. A critic trained on the combined return must also track a changing Bellman fixed point, with the resulting error bounded in Section 4.2. Under a fixed evaluated policy, separate reward and cost Bellman operators remain independent of the multiplier. Second, training reward and cost models separately need not recover the value or policy required by the combined objective. With limited capacity, separate critics may fit shared structure that cancels upon subtraction while missing the smaller value difference used by the policy (Section 4.3). For policy learning, additive composition of separately trained reward and safety flows can fail to generate the policy defined by the combined value, even when both component flows are exact (Appendix G).

UNIQ fixes reward and cost coefficients as functions of the safety context, allowing priorities to vary across contexts while keeping the utility fixed across training iterations. A single flow critic learns the combined value, and a conditional flow policy learns directly from it. The next section formalizes this objective and the flow matching losses; Section 5 presents cumulative-cost conditioning and the practical training procedure.

3 UNIQ: UNIFIED CONDITIONAL FLOWS FOR OFFLINE SAFE RL

UNIQ learns a single combined value under a pretrained fixed safety rule and uses this value to train a conditional flow policy. The rule specifies how reward and cost enter the learning objective; conditioning allows the critic and policy to respond to the current safety context. We define the rule, introduce the flow matching construction, and then specify the critic and policy objectives.

3.1 A PRETRAINED FIXED RULE CONDITIONED ON SAFETY

Following region-based approaches to offline safe RL (Zhu et al., 2026; Zheng et al., 2024; Koirala et al., 2024), we specify different reward and cost priorities across estimated safety regions. Let z be a Markov context containing the physical state s and the safety information needed to determine the current utility and the distribution of the next context. Write $\widetilde{P}(\cdot \mid z, a)$ for its transition kernel. Section 5 instantiates this context, and Appendix D.1 verifies its Markov property. A fixed classifier $\kappa(z)$ assigns a safety label, and a coefficient rule $\omega(z) = (\alpha_r(z), \alpha_c(z))$ defines the utility, bounded by a constant $G < \infty$:

$$g_\omega(z, a) = \alpha_r(z) r(s, a) - \alpha_c(z) c(s, a), \qquad |g_\omega(z, a)| \le G. \tag{5}$$

UNIQ uses three regimes:

$$(\alpha_r(z), \alpha_c(z)) = \begin{cases} (1, 0), & \kappa(z) = \text{safe}, \\ (1, \lambda_\omega(z)), & \kappa(z) = \text{boundary}, \\ (0, 1), & \kappa(z) = \text{recovery}. \end{cases} \tag{6}$$

These regimes prioritize reward, balance reward and cost, and prioritize cost reduction, respectively. The boundary coefficient $\lambda_\omega(z) \ge 0$ is fixed before final training. The maps κ, ω, and the context update remain unchanged during the training stage. A trajectory can move between regimes without changing these maps.

3.2 CONDITIONAL FLOW MATCHING

Flow matching learns a velocity field that transports a base distribution to a target distribution (Lipman et al., 2023). UNIQ applies it to scalar critic targets (Agrawalla et al., 2026) and actions (Park et al., 2025). Let ξ denote the conditioning variables: $\xi = (z, a)$ for the critic and $\xi = z$ for the policy. Given ξ, draw independent samples $E \sim p_0$ and $Y \sim p_1(\cdot \mid \xi)$, where p_0 is the base law and p_1 the desired target law. For independent $t \sim \mathrm{Unif}[0, 1]$, define

$$X_t = (1 - t)E + tY, \qquad D = Y - E, \tag{7}$$

and train the velocity field with $\mathcal{L}(v) = \mathbb{E}\left[\|v(t, X_t, \xi) - D\|_2^2\right]$. Generation holds ξ fixed and solves the ordinary differential equation (ODE)

$$\frac{dU_t}{dt} = v_\theta(t, U_t, \xi), \qquad U_0 = E, \qquad F_\theta(E; \xi) = U_1. \tag{8}$$

Flow time $t \in [0, 1]$ is distinct from environment time j and training iteration k. Training requires only sampled interpolants; generation requires ODE integration. Under the moment and transport assumptions of Proposition 2, the population optimal field generates the conditional target law. Appendix C gives the precise statement and proof.

3.3 A UNIFIED CRITIC AND ITS INDUCED POLICY

For a policy π, the utility in (5) defines the action value and Bellman operator

$$Q_\omega^\pi(z, a) = \mathbb{E}_\pi\left[\sum\nolimits_{j=0}^{\infty} \gamma^j g_\omega(z_j, a_j) \mid z_0 = z, a_0 = a\right], \tag{9}$$

$$(T_\omega^\pi q)(z, a) = g_\omega(z, a) + \gamma \mathbb{E}_{z' \sim \widetilde{P}(\cdot \mid z, a),\, a' \sim \pi(\cdot \mid z')} q(z', a'). \tag{10}$$

Let $Q_{r,\omega}^\pi$ and $Q_{c,\omega}^\pi$ be the discounted values of $\alpha_r(z_j) r_j$ and $\alpha_c(z_j) c_j$ under the same policy. Then $Q_\omega^\pi = Q_{r,\omega}^\pi - Q_{c,\omega}^\pi$. Each future reward and cost is weighted at the context where it occurs; weighting ordinary component values only by the current coefficients generally gives a different value (Appendix D.2).

Critic objective. At iteration k, hold the current policy π_k and target critic $\bar{q}_k$ fixed during the critic update. Given a transition (z, a, r, c, z'), draw $a' \sim \pi_k(\cdot \mid z')$ and form

$$Y_Q = g_\omega(z, a) + \gamma \bar{q}_k(z', a'). \tag{11}$$

Its conditional mean is $(T_\omega^{\pi_k} \bar{q}_k)(z, a)$. With independent scalar source noise E_Q and uniform flow time t, fit

$$\mathcal{L}_Q(\theta) = \mathbb{E}\left[\left| v_\theta^Q(t, (1-t)E_Q + tY_Q, z, a) - (Y_Q - E_Q)\right|^2\right]. \tag{12}$$

The endpoint mean supplies the value estimate used for policy improvement: $q_\theta(z, a) = \mathbb{E}_{E_Q} F_\theta^Q(E_Q; z, a)$. Thus the flow models a one-step Bellman target distribution, whose mean provides the value update. Combining reward and cost in Y_Q aligns critic fitting with the value used by the policy.

Policy objective. Let $\beta(\cdot \mid z)$ be a fixed reference action distribution from the offline data. Given the critic q fixed, consider an action distribution p and temperature $\tau > 0$:

$$\mathcal{J}_z(p; q) = \mathbb{E}_{a\sim p} q(z, a) - \tau \,\mathrm{KL}(p \| \beta(\cdot \mid z)), \tag{13}$$

where KL is Kullback–Leibler divergence. For finite objectives and $0 < Z_q(z) < \infty$, its unique maximizer is

$$\pi_q(da \mid z) = \frac{e^{q(z,a)/\tau}}{Z_q(z)} \beta(da \mid z), \qquad Z_q(z) = \mathbb{E}_{a\sim\beta(\cdot|z)} e^{q(z,a)/\tau}. \tag{14}$$

Appendix C.4 proves this characterization. To fit π_q, draw $z \sim \mu$, where μ is the context distribution, $A_1 \sim \beta(\cdot \mid z)$, independent source $A_0 \sim p_0^\pi$, and independent uniform flow time t. Minimize the weighted loss

$$\mathcal{L}_\pi(\phi; q) = \mathbb{E}\left[\frac{e^{q(z,A_1)/\tau}}{Z_q(z)} \| v_\phi^\pi(t, (1-t)A_0 + tA_1, z) - (A_1 - A_0)\|_2^2\right]. \tag{15}$$

The weight changes the action target from β to π_q, so a single conditional flow learns the policy defined by the combined value (Feng et al., 2025). Appendix C.3 derives this reweighting, while Appendix G explains why additive composition of separately trained reward and safety flows need not recover π_q, even with exact component fields.

4 Main Results

In this section, we analyze three questions: whether a flow critic preserves the Bellman update (Proposition 1), which source of target variation is removed by freezing the safety rule (Theorem 1), and when fitting the combined value is more accurate than fitting its components (Theorem 2). The results concern value learning for the specified utility; they do not establish joint actor–critic convergence or feasibility of the original constrained problem. Section 6 evaluates the cost criterion empirically. Proofs appear in the appendices. All proofs are deferred to the appendices.

4.1 Bellman Consistency of the Flow Critic

UniQ uses a conditional flow to learn a scalar value at each context–action pair. We first verify that this representation preserves the Bellman update: fitting the target distribution and taking its mean must recover the value required for policy evaluation. Fix a policy π, a rule ω, and a bounded value estimate q. For the transition kernel $\widetilde{P}$ of the Markov context z, define

$$Y_q = g_\omega(z, a) + \gamma q(Z', A'), \qquad Z' \sim \widetilde{P}(\cdot \mid z, a), \quad A' \sim \pi(\cdot \mid Z'). \tag{16}$$

The randomness in the next context and action makes Y_q a conditional regression target whose mean is $(T_\omega^\pi q)(z, a)$.

Proposition 1 (Bellman consistency)**.** *Let g_ω and q be bounded. Assume the source E_Q has finite second moment and is independent of Y_q conditional on (z, a). Suppose the flow realizes the population minimizing velocity field and satisfies the transport assumption of Proposition 2. Its endpoint F_q then satisfies, almost everywhere under the context–action sampling distribution,*

$$\mathcal{L}(F_q(E_Q; z, a) \mid z, a) = \mathcal{L}(Y_q \mid z, a), \qquad \mathbb{E}_{E_Q} F_q(E_Q; z, a) = (T_\omega^\pi q)(z, a), \tag{17}$$

where $\mathcal{L}(\cdot \mid z, a)$ denotes a conditional probability law.

Proposition 1 identifies the value update implemented by exact population fitting (Appendix C.2). We next allow imperfect updates and changes in the evaluated policy and utility.

4.2 Value Tracking under a Fixed Safety Rule

Freezing the safety rule removes one source of variation in the value being learned. If the reward–cost coefficients change, the combined value can move even under a fixed policy. The following bound separates this utility variation from policy changes and critic update error.

Consider a common Markov context space with fixed transition kernel $\widetilde{P}$. At iteration k, let g_k, π_k, and q_k denote the utility, evaluated policy, and critic estimate. The exact value being estimated is the Bellman fixed point $Q^{\pi_k}_{g_k}$. We allow an approximate Bellman update,

$$\|q_{k+1} - T^{\pi_k}_{g_k} q_k\|_\infty \le \zeta_k, \tag{18}$$

where ζ_k bounds the total value-update error. Write $e_k = \|q_k - Q^{\pi_k}_{g_k}\|_\infty$ for the critic error, and measure the changes in utility and policy by $d_k = \|g_{k+1} - g_k\|_\infty$, $p_k = \sup_z \mathrm{TV}(\pi_{k+1}(\cdot \mid z), \pi_k(\cdot \mid z))$. Here $\mathrm{TV}(p, p') = \sup_B |p(B) - p'(B)|$, and all supremum norms use the common context–action domain.

Theorem 1. *Assume $|g_k(z,a)| \le G$ for every k, z, a, bounded q_0, and finite errors in* (18).

$$e_{k+1} \le \gamma e_k + \zeta_k + \underbrace{\frac{d_k}{1-\gamma}}_{\text{utility change}} + \underbrace{\frac{2\gamma G}{(1-\gamma)^2} p_k}_{\text{policy change}}. \tag{19}$$

For $g_k = r - \lambda_k c$ with bounded multipliers, $d_k \le C_{\max} \|\lambda_{k+1} - \lambda_k\|_\infty$; this includes scalar multipliers and context-dependent coefficient functions. For a fixed rule $g_k = g_\omega$, $d_k = 0$.

For a fixed rule, $d_k = 0$, leaving only the previous critic error, update error, and policy change. This also holds when a trajectory crosses safety regions: the context changes, but the utility remains the same function. Appendix E.6 gives exact examples of the effect of utility changes. Fixing the evaluated policy gives the following convergence result.

Corollary 1. *Under Theorem 1, fix $g_k = g_\omega$ and $\pi_k = \pi$. For $e_k = \|q_k - Q^\pi_\omega\|_\infty$,*

$$e_K \le \gamma^K e_0 + \sum_{k=0}^{K-1} \gamma^{K-1-k} \zeta_k. \tag{20}$$

If $\zeta_k \to 0$, then $q_K \to Q^\pi_\omega$ in supremum norm. If $\zeta_k \le \bar{\zeta}$ for every k, then

$$e_K \le \gamma^K e_0 + \frac{1-\gamma^K}{1-\gamma}\bar{\zeta}, \qquad \limsup_{K\to\infty} e_K \le \frac{\bar{\zeta}}{1-\gamma}. \tag{21}$$

Exact value updates converge geometrically, while uniformly bounded update error gives asymptotic error at most $\bar{\zeta}/(1-\gamma)$. During joint training, policy changes still contribute through Theorem 1. Appendix E.2 relates flow approximation, numerical error, data error, and target-critic lag to ζ_k. The tracking bound applies to approximate Bellman updates; Proposition 1 connects those updates to the flow critic. Appendix E.7 further bounds reward and cost deviations from a reference policy under additional regularity assumptions. Its feasibility conclusion requires a reference safety margin large enough to absorb those errors.

4.3 Direct Approximation of the Combined Value

We now show why UniQ learns the combined value directly. The policy uses the difference between reward and cost values, whereas separate critics minimize their component errors. Under limited capacity, fitting each component well can leave this difference poorly estimated. We construct a conditional flow class in which fitting the combined target directly gives strictly smaller value error.

Fix a critic fitting problem. Let $x = (z, a)$ denote the critic input, sampled from a distribution μ_{cap}, and let Y_r and Y_c be jointly distributed scalar regression targets for the reward and cost components. Direct fitting uses $Y_U = Y_r - Y_c$. The value to be estimated in either case is its conditional mean,

$$m(x) = \mathbb{E}[Y_U \mid x] = \mathbb{E}[Y_r \mid x] - \mathbb{E}[Y_c \mid x].$$

Algorithm 1 UNIQ

Require: Offline data $\mathcal{D}$, cost limit ℓ, sampler p_U, N_u, and target-update rate $\eta_Q \in (0, 1]$.

1: **Safety rule.** Fit $\widehat{Q}_c, \widehat{V}_c$; construct κ by (23); freeze the cost networks and κ, ω.
2: Initialize critic v_θ^Q, baseline V_ν, policy v_ϕ^π, and target parameters $\bar{\theta} \leftarrow \theta$.
3: **for** each training step **do**
4: **Budget sampling.** Sample transitions $(s, a, r, c, s') \sim \mathcal{D}$ and N_u independent draws $u \sim p_U$ per transition.
5: For each draw u, set $u' = u + c$, $z = (s, u)$, $z' = (s', u')$, and evaluate $\kappa(s, u)$ and $\kappa(s', u')$.
6: **Critic.** Draw $a' \sim \pi_\phi(\cdot \mid s', u')$; form $Y_Q = g_\omega(z, a) + \gamma \bar{q}(z', a')$.
7: Update θ with (12), holding Y_Q fixed. Estimate $q_\theta(z, a)$ and update ν with (25).
8: **Actor.** Compute A, w by (26); update ϕ with (27), holding w fixed.
9: **Target critic.** $\bar{\theta} \leftarrow (1 - \eta_Q)\bar{\theta} + \eta_Q \theta$.
10: **end for**

Each fitting problem uses the same conditional flow class. For $i \in \{r, c, U\}$, let q_i^* be the endpoint mean of a flow minimizing the population flow matching loss for Y_i. We compare q_U^* with $q_r^* - q_c^*$ using $\|f\|_{L_2(\mu_{\text{cap}})} = (\mathbb{E}_{x \sim \mu_{\text{cap}}} |f(x)|^2)^{1/2}$.

Theorem 2. *There exist a finite critic-input distribution μ_{cap}, square-integrable joint targets (Y_r, Y_c), and a common capacity-restricted conditional flow class whose unique population minimizing endpoint means satisfy*

$$\|q_U^* - m\|_{L_2(\mu_{\text{cap}})} = 0, \qquad \|q_r^* - q_c^* - m\|_{L_2(\mu_{\text{cap}})} = \Delta > 0. \tag{22}$$

All three fits use the same source distribution, independent source–target pairing, linear interpolation, uniform flow time, and per-model capacity. For learned means satisfying $\|\widehat{q}_i - q_i^*\|_{L_2(\mu_{\text{cap}})} \leq \delta_i$, $i \in \{U, r, c\}$, direct fitting still has strictly smaller value error whenever $\delta_U + \delta_r + \delta_c < \Delta$.

In the construction, reward and cost targets share a large component that consumes the capacity of both separate critics. Their fitted values cancel, leaving the smaller difference unrepresented. Direct fitting allocates its capacity to that difference and recovers it exactly. The gap holds at the population optima, even though separate fitting uses two models with the same per-model capacity. This is an existence result for illustrative Gaussian targets in a restricted class, rather than a universal ordering of neural critics. Appendix F gives the construction and equality cases; Corollary 2 shows that the gap persists under small target perturbations.

5 UNIQ: PRACTICAL ALGORITHM

Algorithm 1 has two stages: estimate and freeze the safety rule, then train the unified critic and conditional policy across cumulative-cost conditions.

Estimating the Safety Rule. We fit a cost critic $\widehat{Q}_c(s, a)$ and value $\widehat{V}_c(s)$ using a cost-minimizing variant of IQL (Kostrikov et al., 2022; Zheng et al., 2024), which favors low-cost actions supported by the offline data. The critic estimates discounted cost after taking action a in state s, and the value summarizes these estimates through lower-cost expectile regression. We use $\widehat{V}_c$ to define the preliminary state classification:

$$\kappa_0(s) = \begin{cases} \text{safe}, & \widehat{V}_c(s) \leq \ell - \varepsilon_{\text{reg}}, \\ \text{boundary}, & |\widehat{V}_c(s) - \ell| < \varepsilon_{\text{reg}}, \\ \text{recovery}, & \widehat{V}_c(s) \geq \ell + \varepsilon_{\text{reg}}, \end{cases} \tag{23}$$

where $\varepsilon_{\text{reg}} > 0$. For safe and boundary states, a fixed budget-dependent check refines the classification and reward–cost weighting. We denote the resulting context label by $\kappa(s, u)$, distinguishing it from $\kappa_0(s)$; recovery states directly prioritize cost reduction. The cost networks and complete maps κ, ω are frozen throughout final training. Appendix B provides the cost-value losses.

Training across Budget Conditions. Let $z = (s, u)$, where u is accumulated, undiscounted cost. During execution, the policy receives the cost actually incurred along its trajectory, initialized at

$u_0 = 0$ and updated by $u_{j+1} = u_j + c(s_j, a_j)$. This input records cost history; the reward and cost objective remains the discounted formulation in (2) (Appendix D.3).

Each logged transition is observed under a particular cost history, so offline trajectories may provide limited coverage of cumulative-cost conditions at a given state. To broaden this coverage, we pair each logged transition with N_u cumulative-cost values sampled according to

$$p_U = \rho_U \,\mathrm{Unif}([0, \ell]) + (1 - \rho_U) p_{\mathrm{over}}, \tag{24}$$

where p_{over} is the empirical distribution of cumulative costs above ℓ, obtained from logged trajectories or stored annotations. In implementation, we draw $\lceil \rho_U N_u \rceil$ samples from $\mathrm{Unif}([0, \ell])$ and the remaining $N_u - \lceil \rho_U N_u \rceil$ samples from p_{over}. We fix $\rho_U = 0.6$ heuristically without tuning this parameter. For each draw, set $u' = u + c$, $z = (s, u)$, and $z' = (s', u')$. This augmentation reuses observed physical transitions to train the shared critic and policy across different cumulative-cost conditions. Lemma 3 establishes the conditional Bellman identity preserved by this augmentation, including its empirical-kernel interpretation for finite replay data. Additional draws expand the cumulative-cost inputs paired with logged transitions; they do not add physical state–action observations.

The critic receives $(s, a, u, \kappa(s, u))$ and the actor receives (s, u). For a state that passes the cost-value test, resampling u can change its safety label and reward–cost coefficients. Each draw therefore requires reevaluating the same fixed rule at the new context. This changes the training input, not the rule itself. Appendix D.5 quantifies the value and decision errors that can arise when this information is omitted. At deployment, u is updated from observed costs without resampling.

Critic updates. Use (12) with target (11), computed from the current policy and $\bar{q} = q_{\bar{\theta}}$ and fixed during each update. Estimate q_θ by averaging endpoints from M independent Gaussian sources after K Euler steps. Fit $V_\nu(z) = V_\nu(s, u, \kappa(s, u))$, used only for actor weighting, by minimizing

$$\mathcal{L}_V(\nu) = \mathbb{E}_{\mathcal{D}, p_U} L^2_{\tau_V}(\mathrm{sg}[q_\theta(z, a)] - V_\nu(z)) \,. \tag{25}$$

Here $L^2_{\tau_V}(x) = |\tau_V - \mathbf{1}\{x < 0\}| x^2$ with $\tau_V \in (0, 1)$, and sg stops gradients.

Actor Update. For temperature $\tau > 0$ and cap $w_{\max} > 0$, define

$$A(z, a) = q_\theta(z, a) - V_\nu(z), \qquad w(z, a) = \min\{e^{A(z,a)/\tau}, w_{\max}\}. \tag{26}$$

Given a logged action a, independent source $a_0 \sim p_0^\pi$, and $t \sim \mathrm{Unif}[0, 1]$, set $a_t = (1 - t)a_0 + ta$ and minimize

$$\mathcal{L}_\pi^{\mathrm{impl}}(\phi) = \mathbb{E}_{\mathcal{D}, p_U, a_0, t} \left[\mathrm{sg}[w(z, a)] \| v_\phi^\pi(t, a_t, s, u) - (a - a_0) \|_2^2 \right] . \tag{27}$$

Without clipping, the action-independent factor $e^{-V_\nu(z)/\tau}$ cancels when the weights are normalized at each context, recovering (14). Independent cumulative-cost sampling leaves the logged reference unchanged: $\beta(da \mid s, u) = \beta_0(da \mid s)$, where β_0 is the logged action distribution. Appendix B.2 derives the modified target induced by the clipped weights.

6 Experiments

Benchmarks and metrics. We use nine MetaDrive tasks from DSRL (Liu et al., 2023a), combining three road-difficulty levels (Easy, Medium, and Hard) with three traffic-density levels (Sparse, Mean, and Dense). All methods are trained on fixed offline datasets. We report normalized reward and cost and classify a result as feasible when its reported normalized cost is below 1. Rewards are compared among feasible methods on each task. Gray marks infeasible results, bold marks feasible results, and blue highlights the highest feasible reward.

Baselines. We compare UniQ with eight baselines: BC-safe, CPQ (Xu et al., 2022), COptiDICE (Lee et al., 2022), FISOR (Zheng et al., 2024), CCAC (Guo et al., 2025), CAPS(IQL) (Chemingui et al., 2025), SafeFQL (Tayal et al., 2026), and PD-Flow, covering imitation, constrained value learning, feasibility-based policy learning, and budget-conditioned control. PD-Flow is our primal–dual flow baseline with flow-matching regularization and an adaptive Lagrange multiplier. See Appendix A for more details.

Main results. Table 1 shows that UniQ satisfies the reported cost threshold on all nine tasks and achieves the highest reward among feasible methods on every task. Its average normalized reward

Table 1: Normalized MetaDrive benchmark results. Each entry reports normalized **reward/cost**, where higher reward (↑) and lower cost (↓) are better. Gray: Unsafe agents. **Bold**: Safe agents whose normalized cost is smaller than 1. **Blue**: Safe agents with the highest reward. PD-Flow denotes the Primal-Dual Flow Actor baseline.

Task	BC-safe	CPQ	COptiDICE	FISOR	CCAC	CAPS (IQL)	SafeFQL	PD-Flow†	**UniQ** (ours)
EasySparse	0.11 / 1.42	**-0.06 / 0.16**	0.93 / 9.27	**0.50 / 0.87**	**-0.06 / 0.03**	**0.04 / 0.47**	**-0.06 / 0.07**	**-0.05 / 0.10**	**0.54 / 0.91**
EasyMean	**0.10 / 0.10**	**-0.06 / 0.05**	0.74 / 7.23	**0.41 / 0.48**	**-0.07 / 0.10**	**0.07 / 0.72**	**-0.07 / 0.07**	**-0.04 / 0.11**	**0.52 / 0.90**
EasyDense	**0.33 / 0.09**	**-0.06 / 0.12**	0.54 / 5.27	**0.44 / 0.37**	**-0.06 / 0.10**	**0.12 / 0.43**	**-0.06 / 0.07**	**-0.01 / 0.41**	**0.60 / 0.91**
MediumSparse	**0.13 / 0.39**	**-0.09 / 0.10**	0.79 / 4.48	**0.43 / 0.09**	**-0.06 / 0.23**	**0.27 / 0.36**	**-0.08 / 0.09**	**0.06 / 0.62**	**0.78 / 0.93**
MediumMean	**0.12 / 0.10**	**-0.08 / 0.12**	0.34 / 2.25	**0.44 / 0.31**	**-0.03 / 0.35**	**0.32 / 0.68**	**-0.07 / 0.09**	**0.19 / 0.43**	**0.75 / 0.93**
MediumDense	**0.07 / 0.14**	**-0.08 / 0.10**	0.76 / 4.08	**0.51 / 0.12**	**-0.04 / 0.18**	**0.33 / 0.48**	**-0.06 / 0.05**	**0.06 / 0.63**	**0.83 / 0.97**
HardSparse	0.08 / 2.04	**-0.05 / 0.10**	0.45 / 4.58	**0.27 / 0.01**	**-0.05 / 0.07**	**0.12 / 0.62**	**-0.04 / 0.05**	**-0.03 / 0.24**	**0.47 / 0.94**
HardMean	**0.02 / 0.07**	**-0.05 / 0.14**	0.41 / 4.21	**0.28 / 0.21**	**-0.05 / 0.09**	0.25 / 1.48	**-0.05 / 0.05**	**-0.02 / 0.17**	**0.43 / 0.96**
HardDense	**0.05 / 0.32**	**-0.04 / 0.10**	0.22 / 2.47	**0.32 / 0.37**	**-0.04 / 0.15**	**0.26 / 0.37**	**-0.04 / 0.05**	**-0.03 / 0.15**	**0.39 / 0.91**
Average	**0.11 / 0.52**	**-0.06 / 0.11**	0.58 / 4.87	**0.40 / 0.31**	**-0.05 / 0.14**	**0.20 / 0.62**	**-0.06 / 0.06**	**0.02 / 0.32**	**0.59 / 0.93**

reaches 0.59, approximately 48% **higher** than FISOR's 0.40, the strongest baseline feasible on all nine tasks. This advantage holds across all road-difficulty and traffic-density levels, with absolute reward gains over FISOR ranging from 0.04 to 0.35. UNIQ achieves these gains while operating closer to the cost threshold: normalized costs range from 0.90 to 0.97, averaging 0.93, compared with FISOR's 0.31. These results show that UNIQ uses more of the available cost allowance to obtain consistently higher reward while remaining feasible across the full benchmark. Here, feasibility refers to aggregate task-level cost and does not guarantee that every trajectory satisfies the budget.

6.1 Ablation Studies

Number of cumulative-cost samples. We vary $N_u \in \{8, 16, 24\}$ while keeping the remaining settings fixed (Table 3). All nine task–sample-count combinations have normalized cost below one, ranging from 0.90 to 0.99. Reward varies by at most 0.06 within each task: $N_u = 24$ performs best on EasySparse, whereas $N_u = 8$ performs best on EasyMean and EasyDense. Increasing the sample count gives no consistent reward gain over this range.

Cumulative-cost sampling ratio. We vary the nominal ratio $\rho_U \in \{0.25, 0.6, 0.75\}$ while fixing the total sample count and the sampling distribution within each region. Mean normalized cost remains below one on all three Easy tasks, and reward varies by at most 0.07 within each task (Figure 2). No ratio performs best on every task. This comparison measures sensitivity to the sampling mixture; it does not isolate the benefit of resampling over using logged cumulative costs alone.

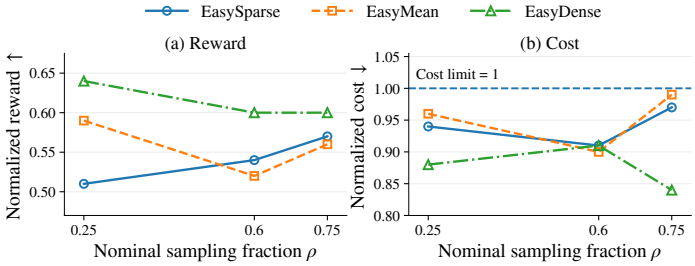


Figure 2: Sensitivity to the budget sampling ratio ρ on three tasks.

Illustrative rollouts. Appendix H illustrates driving behavior under different cost allowances, including speed response, overtaking, and reactions to a vehicle ahead (Figure 3- 5). In the overtaking example, the vehicle with a cost allowance of 10 completes a pass at $t = 5.6$ s, while the vehicle with an allowance of 1 does not complete a pass during the displayed interval. These qualitative examples complement the task-level results. Videos are available on our anonymous project page.

7 Conclusion

UNIQ learns a combined reward–cost value and conditional flow policy under a fixed safety rule, avoiding dual updates during final training. Cumulative-cost resampling broadens training coverage while preserving consistent successor costs. Our analysis establishes Bellman consistency, isolates the utility-change term eliminated by freezing the rule, and demonstrates a strict advantage for direct critic fitting in a capacity-restricted construction. On nine MetaDrive tasks, UNIQ satisfies the reported cost threshold and achieves the highest reward among feasible methods on every task, improving average normalized reward by approximately 48% over the strongest baseline feasible on all nine tasks.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing the manuscript, improving its organization, checking mathematical arguments, and editing code. We also used AI-generated manuscript reviews to identify potential issues and guide revisions. These tools were not used to originate the research problem or design the experiments. All co-authors reviewed the AI-assisted material for grammatical errors, unsupported claims, and consistency with our intended meaning and technical contributions. We take full responsibility for the final content, including all text, claims, code, and other artifacts produced with AI assistance.

ETHICS STATEMENT

This work aims to improve the safety of reinforcement learning agents through theoretical analysis and simulation-based evaluation. Our experiments do not involve human participants, and we do not identify direct ethical concerns arising from the experimental procedures.

REPRODUCIBILITY STATEMENT

We will release the code and instructions needed to reproduce the reported results.

REFERENCES

Bhavya Agrawalla, Michal Nauman, Khush Agrawal, and Aviral Kumar. floq: Training critics via flow-matching for scaling compute in value-based RL. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59d15ab03bd84330d44313eb203c3bb6-Abstract-Conference.html.

Marvin Alles, Nutan Chen, Patrick van der Smagt, and Botond Cseke. FlowQ: Energy-guided flow policies for offline reinforcement learning. *arXiv preprint arXiv:2505.14139*, 2025. URL https://arxiv.org/abs/2505.14139.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Yassine Chemingui, Aryan Deshwal, Honghao Wei, Alan Fern, and Jana Doppa. Constraint-adaptive policy switching for offline safe reinforcement learning. In *AAAI Conf. Artificial Intelligence*, volume 39, pp. 15722–15730, 2025.

Rufeng Chen, Zhaofan Zhang, Zhejiang Yang, Hechang Chen, and Sihong Xie. Decoupled guidance diffusion for adaptive offline safe reinforcement learning. *arXiv preprint arXiv:2605.02777*, 2026. URL https://arxiv.org/abs/2605.02777.

Perry Dong, Chongyi Zheng, Chelsea Finn, Dorsa Sadigh, and Benjamin Eysenbach. Value flows. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.07650.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16993–17029. PMLR, 2025. URL https://proceedings.mlr.press/v267/feng25s.html.

Zijian Guo, Weichao Zhou, Shengao Wang, and Wenchao Li. Constraint-conditioned actor-critic for offline safe reinforcement learning. In *Int. Conf. on Learning Representations (ICLR)*, 2025.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. A primal-dual-critic algorithm for offline constrained reinforcement learning. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 280–288. PMLR, 2024.

Prajwal Koirala, Zhanhong Jiang, Soumik Sarkar, and Cody Fleming. FAWAC: Feasibility informed advantage weighted regression for persistent safety in offline reinforcement learning. *arXiv preprint arXiv:2412.08880*, 2024. URL https://arxiv.org/abs/2412.08880.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *Int. Conf. on Learning Representations (ICLR)*. OpenReview.net, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Int. Conf. Machine Learning (ICML)*, pp. 3703–3712. PMLR, 2019.

Jongmin Lee, Cosmin Paduraru, Daniel J. Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation. In *Int. Conf. on Learning Representations (ICLR)*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. URL https://arxiv.org/abs/2005.01643.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022.

Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Safe offline reinforcement learning with real-time budget constraints. In *Int. Conf. Machine Learning (ICML)*, pp. 21127–21152. PMLR, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Int. Conf. on Learning Representations (ICLR)*, 2023.

Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023a.

Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *Int. Conf. Machine Learning (ICML)*, pp. 21611–21630. PMLR, 2023b.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48104–48127. PMLR, 2025. URL https://proceedings.mlr.press/v267/park25f.html.

Nicholas Polosky, Bruno C. da Silva, Madalina Fiterau, and Jithin Jagannath. Constrained offline policy optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17801–17810. PMLR, 2022. URL https://proceedings.mlr.press/v162/polosky22a.html.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *Int. Conf. Machine Learning (ICML)*, pp. 9133–9143. PMLR, 2020.

Huikang Su, Dengyun Peng, Zifeng Zhuang, YuHan Liu, Qiguang Chen, Donglin Wang, and Qinghe Liu. Boundary-to-region supervision for offline safe reinforcement learning. *arXiv preprint arXiv:2509.25727*, 2025. URL https://arxiv.org/abs/2509.25727.

Manan Tayal and Mumuksh Tayal. Epigraph-guided flow matching for safe and performant offline reinforcement learning. *arXiv preprint arXiv:2602.08054*, 2026. URL https://arxiv.org/abs/2602.08054.

Mumuksh Tayal, Manan Tayal, and Ravi Prakash. Safe flow Q-learning: Offline safe reinforcement learning with reachability-based flow policies. *arXiv preprint arXiv:2603.15136*, 2026. URL https://arxiv.org/abs/2603.15136.

Dhiran Venula, Shrinidi Kuppurajan, Om Vivek Mahajan, Returaj Burnwal, Nirav Pravinbhai Bhatt, and Balaraman Ravindran. Can agents learn safe behavior from non-preferred demonstrations? In *ICRA Workshop on Reinforcement Learning in the Era of Imitation Learning*, 2026. URL https://rl4il-icra.github.io/assets/papers/P8_canagents.pdf.

Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 25439–25451, 2021.

Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized q-learning for safe offline reinforcement learning. In *AAAI Conf. Artificial Intelligence*, volume 36, pp. 8753–8760, 2022.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *Int. Conf. Machine Learning (ICML)*, volume 139, pp. 11480–11491. PMLR, 2021.

Hsin-Jung Yang, Zhanhong Jiang, Prajwal Koirala, Qisai Liu, Cody Fleming, and Soumik Sarkar. LexiSafe: Offline safe reinforcement learning with lexicographic safety-reward hierarchy. *arXiv preprint arXiv:2602.17312*, 2026. URL https://arxiv.org/abs/2602.17312.

Qin Zhang, Linrui Zhang, Haoran Xu, Li Shen, Bowen Wang, Yongzhe Chang, Xueqian Wang, Bo Yuan, and Dacheng Tao. SaFormer: A conditional sequence modeling approach to offline safe reinforcement learning. *arXiv preprint arXiv:2301.12203*, 2023. URL https://arxiv.org/abs/2301.12203.

Shiyuan Zhang, Weitong Zhang, and Quanquan Gu. Energy-weighted flow matching for offline reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2d9c6cdb4cfe93869c090fea7375044b-Abstract-Conference.html.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023. URL https://arxiv.org/abs/2311.13443.

Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. *arXiv preprint arXiv:2401.10700*, 2024.

Jiahui Zhu, Lei Ying, and Honghao Wei. Divide and conquer: Selective value learning and policy optimization for offline safe reinforcement learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=4KYrv6qYMl.

APPENDIX CONTENTS

A Extended Related Work **15**

A.1 Safety objectives and fixed rules . . . 15

A.2 Budget conditioning . . . 15

A.3 Flow policies and critics . . . 15

A.4 Comparison with safe flow methods . . . 15

B More Algorithm Deatils **16**

B.1 Stage 1: cost-value estimation . . . 16

B.2 Stage 2: the implemented actor target . . . 16

C Conditional Flow Matching **16**

C.1 Population target and transport . . . 16

C.2 Proof of Proposition 1 . . . 17

C.3 Weighted conditional flow matching . . . 17

C.4 The optimal regularized policy . . . 18

D The Markov Context and Learning Utility **18**

D.1 Which state is required for the Bellman equation? . . . 18

D.2 Weighting immediate signals versus combining their values . . . 19

D.3 Budget reconstruction and replay . . . 20

D.4 Bellman consistency of random budget sampling . . . 20

D.5 Value and decision error when the budget is omitted . . . 21

E Error Propagation and Stability **22**

E.1 From velocity error to endpoint error . . . 22

E.2 From the implemented regression target to the Bellman update . . . 23

E.3 Movement of the regularized policy target . . . 23

E.4 Proof of Theorem 1 . . . 24

E.5 Proof of Corollary 1 . . . 24

E.6 Exact tracking examples . . . 25

E.7 Complete value error and policy consequences . . . 25

F Capacity and Scalar Critic Approximation **27**

F.1 The equality case for linear classes . . . 27

F.2 Proof of Theorem 2 . . . 27

F.3 The strict gap persists under target perturbations . . . 29

G Composition of Conditional Policy Flows **30**

G.1 Policy targets and additive composition . . . 30

G.2 Proof of Theorem 3 . . . 31

G.3 Error transfer between path distributions . . . 31

G.4 An exact endpoint comparison . . . 32

H Illustrative Driving Rollouts **32**

H.1 Speed Response 33

H.2 Overtaking 33

H.3 Response to a Vehicle Ahead 34

I Detail Experiment Results **34**

I.1 Main Results 34

I.2 Number of cumulative-cost samples. 34

I.3 Cumulative-cost sampling ratio. 35

A Extended Related Work

A.1 Safety objectives and fixed rules

Primal–dual methods optimize reward and cost through coupled policy and multiplier updates (Le et al., 2019; Hong et al., 2024). Other approaches address the reliability of offline constraint estimates: CPQ penalizes out-of-distribution actions through the cost critic (Xu et al., 2022), COptiDICE estimates stationary distribution corrections with conservative cost evaluation (Lee et al., 2022), and COPO uses an offline cost-projection step (Polosky et al., 2022). Safety priorities can also be specified more explicitly. FISOR maximizes reward within an estimated feasible region and reduces risk outside it (Zheng et al., 2024); FAWAC incorporates cost advantages into advantage-weighted policy learning (Koirala et al., 2024); and LexiSafe gives safety precedence over reward through lexicographic optimization (Yang et al., 2026). UniQ uses cost-value assessment followed by a cumulative-cost check to define one combined reward–cost return. It fixes the rule before final training and directly learns the value of that return, avoiding continued multiplier adaptation during this stage.

A.2 Budget conditioning

Budget conditioning is an established approach to offline safe RL. Constrained Decision Transformer conditions sequence prediction on reward and cost targets (Liu et al., 2023b), while SaFormer conditions action proposals on remaining cost and verifies their safety (Zhang et al., 2023). B2R modifies cost-to-go supervision to represent a safety boundary rather than an exact target (Su et al., 2025). Among actor–critic approaches, CCAC conditions learning on constraint thresholds (Guo et al., 2025), and CAPS switches among policies with different reward–cost trade-offs at deployment (Chemingui et al., 2025). TREBI uses diffusion-based trajectory planning for real-time budgets (Lin et al., 2023); SDGD separates cost-conditioned generation from reward-gradient guidance (Chen et al., 2026). UniQ incorporates accumulated cost into a shared critic and policy and trains them across sampled budget conditions under the same fixed utility. For states that pass the initial cost-value assessment, accumulated cost also enters the safety rule. Resampling this input exposes the shared models to the resulting conditional objectives while keeping the rule itself fixed.

A.3 Flow policies and critics

Flow matching learns generative velocity fields through regression (Lipman et al., 2023). Flow Q-Learning uses an expressive behavior flow to regularize a one-step actor (Park et al., 2025). Energy-Weighted Flow Matching and FlowQ incorporate value information into flow policy training through weighted regression and energy-guided probability paths, respectively (Zhang et al., 2025; Alles et al., 2025), while guided-flow analyses characterize how conditioning or reweighting changes the desired velocity field (Zheng et al., 2023; Feng et al., 2025). For value learning, floq trains flow critics with temporal-difference targets (Agrawalla et al., 2026), and Value Flows models return distributions through distributional Bellman updates (Dong et al., 2026). UniQ builds on these constructions by fitting the combined Bellman target and reading out its mean. Our analysis studies when directly fitting this value is more accurate than subtracting separately fitted reward and cost values under a common capacity restriction.

A.4 Comparison with safe flow methods

SafeFQL combines separate reward and reachability critics with a behavior flow distilled into a one-step actor (Tayal et al., 2026). EpiFlow learns an auxiliary value that jointly represents safety and performance through an epigraph reformulation of hard state constraints, then uses that value to train a weighted flow policy (Tayal & Tayal, 2026). With different data assumptions, NOPE learns implicit preferences from non-preferred and unlabeled demonstrations to train a value-weighted flow policy (Venula et al., 2026). UniQ studies the discounted return of an additive reward–cost utility whose coefficients remain fixed during final training. A flow critic estimates this combined value directly, and a conditional flow policy learns from its advantage. Our analysis isolates the benefit of freezing the utility and constructs a strict advantage for direct critic fitting under a common capacity restriction (Section 4). We also analyze when additive composition of exact component flows fails to recover the combined policy target (Appendix G).

B More Algorithm Deatils

B.1 Stage 1: cost-value estimation

Following Zheng et al. (2024), we fit $\widehat{V}_c$ and $\widehat{Q}_c$ using a cost-minimizing variant of IQL. To distinguish the cost expectile from the actor temperature τ, denote its parameter by $\tau_c \in (1/2, 1)$ and define $L_c^{\tau_c}(e) = |\tau_c - \mathbf{1}\{e > 0\}|e^2$. The losses are

$$\mathcal{L}_{\widehat{V}_c} = \mathbb{E}_{(s,a)\sim\mathcal{D}} L_c^{\tau_c}(\widehat{Q}_c(s,a) - \widehat{V}_c(s)), \tag{28}$$

$$\mathcal{L}_{\widehat{Q}_c} = \mathbb{E}_{(s,a,c,s')\sim\mathcal{D}}\big(c + \gamma\widehat{V}_c(s') - \widehat{Q}_c(s,a)\big)^2. \tag{29}$$

With this sign convention, $\tau_c > 1/2$ gives greater weight to negative residuals and favors lower cost values. The networks provide an estimated safety assessment for (23); the cost networks, classifier, and coefficient rule are frozen after this stage.

B.2 Stage 2: the implemented actor target

The flow critic is trained on (11). Its endpoint mean q_θ is the value estimate, and the expectile baseline V_ν in (25) is used only to form the actor advantage $A = q_\theta - V_\nu$. The critic bootstraps from the target action value, rather than from this baseline.

To identify the actor's target, fix the critic and baseline during an actor update. For the clipped positive weight w in (26), define

$$Z_w(z) = \mathbb{E}_{a\sim\beta(\cdot|z)} w(z,a), \qquad \pi_w(da \mid z) = \frac{w(z,a)}{Z_w(z)}\beta(da \mid z). \tag{30}$$

For $0 < Z_w(z) < \infty$, Lemma 1 shows that normalized weighted flow matching fits this conditional endpoint law under the stated moment and transport assumptions. The implemented loss omits $Z_w(z)$. At each fixed context, this changes the loss by a positive scale and leaves an unrestricted conditional minimizer unchanged. Across contexts, however, it weights the regression problems in proportion to $Z_w(z)$, which can affect approximation in a shared restricted model class.

Without clipping, $w = e^{(q_\theta - V_\nu)/\tau}$, and the action-independent factor $e^{-V_\nu(z)/\tau}$ cancels in (30). The conditional target is then π_{q_θ} from (14), the maximizer of the regularized actor objective. When clipping is active, (30) describes the modified target. Thus the exact regularized-policy characterization applies to the unclipped weights; the practical capped objective has its own explicitly defined endpoint distribution.

Conventions for the proofs. Throughout the proofs, vector norms are Euclidean; $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ denote matrix Frobenius and spectral norms. We use

$$\mathrm{TV}(p,q) = \sup_B |p(B) - q(B)|, \qquad W_r(p,q) = \inf_{\Gamma\in\Pi(p,q)} \left(\mathbb{E}_\Gamma\|X - Y\|^r\right)^{1/r}, \quad r \in \{1,2\},$$

where $\Pi(p,q)$ is the set of couplings with marginals p, q. All policy comparisons use the same retained state, transition kernel, and initial distribution. A uniform conditional bound does not follow from an integrated regression bound without an additional assumption.

C Conditional Flow Matching

C.1 Population target and transport

For the interpolation in (7), let ν denote the joint law of (t, X_t, ξ). Define

$$v^*(t,x,\xi) = \mathbb{E}[Y - E \mid t, X_t = x, \xi]. \tag{31}$$

The corresponding excess risk is

$$\mathcal{L}(v) - \mathcal{L}(v^*) = \mathbb{E}\|v(t,X_t,\xi) - v^*(t,X_t,\xi)\|_2^2 = \|v - v^*\|_{L_2(\nu)}^2. \tag{32}$$

Proposition 2 (Population target of conditional flow matching)**.** *Suppose* $\mathbb{E}[\|E\|_2^2 + \|Y\|_2^2] < \infty$*, and let* ν *be the joint law of* (t, X_t, ξ)*. The population minimizer over square integrable fields is, uniquely up to sets of zero* ν *measure, the conditional expectation in* (31)*. For every square integrable field* v*, the excess-risk identity* (32) *holds. The conditional distributions of* X_t *satisfy the continuity equation with velocity* v^**. If this equation has a unique solution generated by its ODE, that ODE transports the base law to* $p_1(\cdot \mid \xi)$ *at* $t = 1$*, for almost every* ξ*.*

Proof. Let $\mathcal{H} = \sigma(t, X_t, \xi)$ and set $m = \mathbb{E}[D \mid \mathcal{H}]$. Since $D \in L_2$, its conditional expectation exists in L_2. A predictor evaluated at (t, X_t, ξ) is $\mathcal{H}$ measurable. The identity

$$\|v - D\|^2 = \|v - m\|^2 + \|m - D\|^2 + 2\langle v - m, m - D\rangle$$

has a zero expected cross term, because

$$\mathbb{E}[\langle v - m, m - D\rangle \mid \mathcal{H}] = \langle v - m, m - \mathbb{E}[D \mid \mathcal{H}]\rangle = 0.$$

Consequently $\mathcal{L}(v) = \mathcal{L}(m) + \mathbb{E}\|v - m\|^2$. This proves (31) and (32), including uniqueness of the minimizer up to sets of zero path measure.

To establish the transport statement, fix a context outside a conditional null set. For $\varphi \in C_c^1(\mathbb{R}^d)$, the path X_t is absolutely continuous with derivative D. Its conditional finite first moment and the bounded derivative of φ give

$$\mathbb{E}[\varphi(X_t) \mid \xi] - \mathbb{E}[\varphi(X_0) \mid \xi] = \int_0^t \mathbb{E}[\nabla\varphi(X_u) \cdot D \mid \xi] du.$$

Taking the conditional expectation of D given X_u inside this integral gives

$$\frac{d}{dt} \int \varphi(x) p_t(dx \mid \xi) = \int \nabla\varphi(x) \cdot v^*(t, x, \xi) p_t(dx \mid \xi) \tag{33}$$

for almost every t. This is the weak continuity equation, with initial law $\mathcal{L}(E \mid \xi)$ and terminal law $\mathcal{L}(Y \mid \xi)$. Under the stated uniqueness assumption for the ODE generated solution, its law equals p_t at every time. □

C.2 Proof of Proposition 1

Proof. Let ρ_Q denote the context–action sampling distribution and write $x = (z, a)$. Boundedness of g_ω and q gives $|Y_q| \le \|g_\omega\|_\infty + \gamma\|q\|_\infty$, so the target and the assumed source have finite second moments. For almost every x under ρ_Q, take $Z' \sim \widetilde{P}(\cdot \mid z, a)$ and $A' \sim \pi(\cdot \mid Z')$. Conditional expectation gives

$$\begin{aligned}\mathbb{E}[Y_q \mid z, a] &= g_\omega(z, a) + \gamma \int\int q(z', a')\pi(da' \mid z')\widetilde{P}(dz' \mid z, a) \\ &= (T_\omega^\pi q)(z, a).\end{aligned}$$

Define $X_t = (1 - t)E_Q + tY_q$. Proposition 2 gives the population minimizing field

$$v_q^*(t, y, z, a) = \mathbb{E}[Y_q - E_Q \mid t, X_t = y, z, a].$$

By realizability and the stated conditional transport assumption, its ODE endpoint has law $\mathcal{L}(Y_q \mid z, a)$. Equality of these finite-moment laws implies equality of their means, which proves (17).

□

C.3 Weighted conditional flow matching

Fix a context distribution and a reference target law $p_{\mathrm{ref}}(\cdot \mid \xi)$. For a positive weight w, define

$$Z_w(\xi) = \mathbb{E}_{p_{\mathrm{ref}}} w(Y, \xi), \qquad p_w(dy \mid \xi) = \frac{w(y, \xi)}{Z_w(\xi)} p_{\mathrm{ref}}(dy \mid \xi). \tag{34}$$

For an independent source E and target $Y \sim p_{\mathrm{ref}}(\cdot \mid \xi)$, let $X_t = (1 - t)E + tY$ and $D = Y - E$. The normalized weighted loss is

$$\mathcal{L}_w(v) = \mathbb{E}\left[\frac{w(Y, \xi)}{Z_w(\xi)} \|v(t, X_t, \xi) - D\|^2\right]. \tag{35}$$

With $\xi = z$, $p_{\mathrm{ref}} = \beta$, and $w = e^{q/\tau}$, these are exactly the target and training loss of the actor in Section 3.3.

Lemma 1. *Let $E \sim p_0(\cdot \mid \xi)$ and $Y \sim p_{\mathrm{ref}}(\cdot \mid \xi)$ be conditionally independent, and let t be independent and uniform. Let $w(Y,\xi) > 0$ have finite positive conditional normalizer $Z_w(\xi)$. Assume*

$$\mathbb{E}\left[\frac{w(Y,\xi)}{Z_w(\xi)}(\|E\|^2 + \|Y\|^2)\right] < \infty.$$

Let ν_w be the law of (t, X_t, ξ) after weighting the source–target samples by $w(Y,\xi)/Z_w(\xi)$. Then (35) *is CFM for the endpoint law* (34)*. Its minimizing field is*

$$v_w^*(t,x,\xi) = \frac{\mathbb{E}_{\mathrm{ref}}[wD \mid t, X_t = x, \xi]}{\mathbb{E}_{\mathrm{ref}}[w \mid t, X_t = x, \xi]},$$

which specializes to (81)*, and*

$$\mathcal{L}_w(v) - \mathcal{L}_w(v_w^*) = \|v - v_w^*\|^2_{L_2(\nu_w)}. \tag{36}$$

The ODE sampling interpretation holds under the transport condition of Proposition 2.

Proof. At each context, change the reference product measure on (E, Y, t) by the density $w(Y,\xi)/Z_w(\xi)$. Since the source and target are conditionally independent, and the weight does not depend on the source or time, the new measure has product law

$$p_0(dE \mid \xi)\, p_w(dY \mid \xi)\, dt.$$

It has total mass one at every context, so the law of ξ is unchanged. The moment assumption ensures that Proposition 2 applies under this measure.

Let $\mathcal{H} = \sigma(t, X_t, \xi)$. Conditional expectation under the changed measure obeys the conditional Bayes identity

$$\mathbb{E}_w[D \mid \mathcal{H}] = \frac{\mathbb{E}_{\mathrm{ref}}[wD \mid \mathcal{H}]}{\mathbb{E}_{\mathrm{ref}}[w \mid \mathcal{H}]}.$$

To verify it, multiply by an arbitrary bounded $\mathcal{H}$ measurable test function and integrate under the changed measure. Conditioning the likelihood ratio on $\mathcal{H}$ cancels the denominator, producing exactly the defining integral for $\mathbb{E}_w[D \mid \mathcal{H}]$. The factor $Z_w(\xi)$ cancels because it is $\mathcal{H}$ measurable. Positivity and integrability of w make the denominator finite and positive almost everywhere under the reweighted law. The regression identity and transport statement follow from Proposition 2.

□

C.4 The optimal regularized policy

Fix z and suppress it in this proof. For $\pi_q = \beta e^{q/\tau}/Z_q$ and any $p \ll \beta$ with finite objective,

$$\log \frac{p(a)}{\pi_q(a)} = \log \frac{p(a)}{\beta(a)} - \frac{q(a)}{\tau} + \log Z_q.$$

Integration against p gives

$$\tau\, \mathrm{KL}(p\|\pi_q) = \tau\, \mathrm{KL}(p\|\beta) - \mathbb{E}_p q + \tau \log Z_q.$$

Substituting $p = \pi_q$ gives $\mathcal{J}(\pi_q; q) = \tau \log Z_q$. Restoring the context index yields

$$\mathcal{J}_z(\pi_q; q) - \mathcal{J}_z(p; q) = \tau\, \mathrm{KL}\big(p\|\pi_q(\cdot \mid z)\big). \tag{37}$$

Relative entropy is nonnegative and vanishes only when the two laws agree almost everywhere, proving optimality and uniqueness.

D The Markov Context and Learning Utility

D.1 Which state is required for the Bellman equation?

Lemma 2. *Let the physical process be Markov with kernel P and nonnegative observed cost $c(s,a)$. Retain either $z = s$ or $z = (s,u)$, where $u \in \mathbb{R}_+$ and $u' = u + c(s,a)$. Let $\kappa(z)$ and $\omega(z)$ be fixed measurable functions. In the first case set $\widetilde{P} = P$; in the second set*

$$\widetilde{P}(ds', du' \mid s, u, a) = P(ds' \mid s, a)\, \delta_{u + c(s,a)}(du'). \tag{38}$$

A fixed actor using the complete context induces a Markov policy $\pi(da \mid z)$ on this space. An explicit input $\kappa(z)$ does not add information because it is a fixed function of z. The law of the next context and its label is

$$K(dz', d\kappa' \mid z, a) = \widetilde{P}(dz' \mid z, a)\, \delta_{\kappa(z')}(d\kappa'). \tag{39}$$

For bounded g_ω, the discounted value Q^π_ω satisfies (10) *and is its unique bounded fixed point. Replacing (s, u) by $(s, \kappa(s, u))$ need not preserve these properties.*

Proof. For $z = s$, the physical Markov property gives the next-state law $P(\cdot \mid s, a)$ conditional on the full history and current action. For $z = (s, u)$, the next cumulative cost is a deterministic function of (s, u, a). Combining these observations proves (38).

Let $\Pi(da \mid z, \kappa)$ be the action law of an actor that receives an explicit label, including its independent source noise. Evaluating that label gives the measurable kernel

$$\pi(da \mid z) = \Pi(da \mid z, \kappa(z)).$$

The next label is the fixed function κ applied to z', which proves (39).

Split the discounted sum in (9) into the current utility and its tail. The Markov property identifies the conditional expected tail as $\gamma Q^\pi_\omega(z', a')$. Bounded utility justifies exchanging the expectation and sum, so

$$Q^\pi_\omega(z, a) = g_\omega(z, a) + \gamma \mathbb{E}[Q^\pi_\omega(z', a') \mid z, a].$$

For bounded measurable q_1, q_2,

$$\|T^\pi_\omega q_1 - T^\pi_\omega q_2\|_\infty \le \gamma \|q_1 - q_2\|_\infty.$$

The space of bounded measurable functions is complete in this norm. Contraction gives uniqueness, and the bounded discounted sum supplies the fixed point.

For the last assertion, consider one physical state with zero- and unit-cost actions. Label $u < 2$ as safe and $u \ge 2$ as recovery. From $u_0 = 0$, either action reaches a safe context, with $u = 0$ or $u = 1$. A subsequent unit-cost action gives $u' = 1$ or $u' = 2$, respectively, so identical retained inputs (s, κ) and actions have different next labels. If utility depends on that label, their Bellman targets can also differ.

□

D.2 Weighting immediate signals versus combining their values

Fix the rule ω and a policy π on the retained Markov context. All values in this subsection use this same policy and transition law. For $h \in \{r, c\}$, define the ordinary component value on this context by

$$Q^\pi_h(z, a) = \mathbb{E}_\pi \left[\sum_{j \ge 0} \gamma^j h(s_j, a_j) \,\middle|\, z_0 = z, a_0 = a \right].$$

These values use the physical reward and cost without context-dependent weights. By contrast, $Q^\pi_{r,\omega}$ and $Q^\pi_{c,\omega}$ use the weighted signals $\alpha_r(z_j) r_j$ and $\alpha_c(z_j) c_j$ at each future step. Bounded coefficients and signals justify exchanging expectation and summation, giving

$$\begin{aligned} Q^\pi_\omega(z, a) &= \mathbb{E}_\pi \left[\sum_{j \ge 0} \gamma^j \alpha_r(z_j) r_j \,\middle|\, z_0 = z, a_0 = a \right] \\ &\quad - \mathbb{E}_\pi \left[\sum_{j \ge 0} \gamma^j \alpha_c(z_j) c_j \,\middle|\, z_0 = z, a_0 = a \right] = Q^\pi_{r,\omega}(z, a) - Q^\pi_{c,\omega}(z, a). \end{aligned} \tag{40}$$

Applying the current coefficients to the ordinary values instead gives the following error.

Proposition 3. *Assume $|r| \le R_{\max}$, $0 \le c \le C_{\max}$, and bounded coefficient functions α_r, α_c. Define*

$$Q^\pi_{\mathrm{cur}}(z, a) = \alpha_r(z) Q^\pi_r(z, a) - \alpha_c(z) Q^\pi_c(z, a), \qquad B_\omega = Q^\pi_{\mathrm{cur}} - Q^\pi_\omega.$$

For a bounded real function f on contexts, write $\operatorname{osc}_z(f) = \sup_z f(z) - \inf_z f(z)$. Then

$$B_\omega(z,a) = \mathbb{E}_\pi\left[\sum_{j\geq 1}\gamma^j\left\{(\alpha_r(z)-\alpha_r(z_j))r_j + (\alpha_c(z_j)-\alpha_c(z))c_j\right\}\,\middle|\, z_0 = z, a_0 = a\right], \tag{41}$$

$$\|B_\omega\|_\infty \leq \frac{\gamma}{1-\gamma}\left[R_{\max}\operatorname{osc}_z(\alpha_r) + C_{\max}\operatorname{osc}_z(\alpha_c)\right]. \tag{42}$$

In the special case $\alpha_r \equiv 1$ and $\alpha_c = \lambda_\omega$, the error satisfies

$$Q_r^\pi(z,a) - \lambda_\omega(z)Q_c^\pi(z,a) - Q_\omega^\pi(z,a) = \mathbb{E}_\pi\left[\sum_{j\geq 0}\gamma^j(\lambda_\omega(z_j)-\lambda_\omega(z))c_j\,\middle|\, z_0 = z, a_0 = a\right], \tag{43}$$

$$\|B_\omega\|_\infty \leq \frac{\gamma C_{\max}}{1-\gamma}\operatorname{osc}_z(\lambda_\omega). \tag{44}$$

The bound in (44) *is attainable.*

Proof. At a fixed initial pair (z,a), the current coefficients are constants and can be moved inside the expectations defining Q_r^π and Q_c^π. Subtracting (40) yields a discounted sum of the differences between current and future coefficients. Its term at $j=0$ vanishes, which proves (41). Every later reward term has absolute value at most $R_{\max}\operatorname{osc}_z(\alpha_r)$, and every later cost term has absolute value at most $C_{\max}\operatorname{osc}_z(\alpha_c)$. Taking absolute values, then summing $\sum_{j\geq 1}\gamma^j = \gamma/(1-\gamma)$, proves (42) uniformly in (z,a). Setting $\alpha_r = 1$ and $\alpha_c = \lambda_\omega$ gives (43) and (44).

For attainment, take a single action and two contexts with $z_0 \to z_1 \to z_1$. Set rewards to zero, costs to 0 at z_0 and $C_{\max}$ at z_1, and coefficients to $\lambda_\omega(z_0) = 0$, $\lambda_\omega(z_1) = D > 0$. Then

$$Q_{\text{cur}}^\pi(z_0,a) = 0, \qquad Q_\omega^\pi(z_0,a) = -\frac{\gamma D C_{\max}}{1-\gamma}, \qquad \operatorname{osc}_z(\lambda_\omega) = D,$$

which attains (44). For $D = 0$, both sides are zero.

□

D.3 Budget reconstruction and replay

The cumulative-cost update $u_{j+1} = u_j + c(s_j,a_j)$ in Section 5 starts from $u_0 = 0$ and gives

$$u_j = \sum_{i=0}^{j-1} c(s_i,a_i). \tag{45}$$

Induction on j proves the identity. The input $u \in \mathbb{R}_+$ records undiscounted cumulative cost, while (2) uses discounted expected cost.

Ordered trajectories or stored annotations provide the cumulative costs used to form the empirical over-budget distribution. Replay then pairs a sampled u with the successor $u' = u + c$. Policy comparisons use the same initial context distribution, $\rho \otimes \delta_0$ for the augmented process and ρ for a state-only rule.

D.4 Bellman consistency of random budget sampling

Lemma 3. *Let $d(ds,da)$ be a physical state–action sampling distribution. Draw $(s,a) \sim d$, $s' \sim P(\cdot \mid s,a)$, and $u \sim \varrho(\cdot \mid s)$ independently of (a,s') conditional on s, where ϱ is a fixed probability kernel on $\mathbb{R}_+$. Assume reward and cost are the functions $r(s,a), c(s,a)$ from Section 2 and do not depend on u. Set $u' = u + c(s,a)$, $z = (s,u)$, and $z' = (s',u')$. For a bounded target critic $\bar{q}$ and a fixed policy π, sample $a' \sim \pi(\cdot \mid z')$. Then, almost everywhere under the augmented sampling distribution,*

$$\mathbb{E}[g_\omega(z,a) + \gamma\bar{q}(z',a') \mid z,a] = (T_\omega^\pi\bar{q})(z,a). \tag{46}$$

For replay from a finite dataset, the identity holds with its empirical conditional transition kernel in place of P.

Proof. The joint law of (s, a, u, s') factors as

$$d(ds, da)\, \varrho(du \mid s)\, P(ds' \mid s, a).$$

Conditioning on (s, u, a) leaves the next physical state law equal to $P(\cdot \mid s, a)$ on the sampled support. Since u' is a deterministic function of (s, u, a), any bounded measurable test function f obeys

$$\begin{aligned}\mathbb{E}[f(z') \mid s, u, a] &= \int f(s', u + c(s, a))P(ds' \mid s, a) \\ &= \int f(z')\widetilde{P}(dz' \mid z, a).\end{aligned}$$

Apply this identity to $f(z') = \int \bar{q}(z', a')\pi(da' \mid z')$ and add the deterministic utility to obtain (46). All expectations are finite by boundedness.

For a finite dataset, write its empirical joint transition law as $\widehat{d}_{\mathcal{D}}(ds, da)\widehat{P}_{\mathcal{D}}(ds' \mid s, a)$. Independent budget resampling gives the same factorization with $\widehat{d}_{\mathcal{D}}$ and $\widehat{P}_{\mathcal{D}}$, so the proof applies to the empirical Bellman operator $\widehat{T}^{\pi}_{\omega}$. The difference from the physical operator is

$$\begin{aligned}&((\widehat{T}^{\pi}_{\omega} - T^{\pi}_{\omega})\bar{q})(s, u, a) \\ &\quad = \gamma \int \mathbb{E}_{a' \sim \pi(\cdot \mid s', u')}\bar{q}(s', u', a')(\widehat{P}_{\mathcal{D}} - P)(ds' \mid s, a), \qquad u' = u + c(s, a). \end{aligned} \tag{47}$$

□

For the algorithm's state-independent sampler, the identity holds for each draw from either region. Independence from the logged action also gives $\beta(da \mid s, u) = \beta_0(da \mid s)$, where β_0 is the logged action distribution.

D.5 Value and decision error when the budget is omitted

Proposition 4. *Fix a policy π and bounded utility g_{ω}. Let (Z, A) have any specified sampling distribution, let $\chi(Z)$ be a measurable retained representation of the context, and set $V = Q^{\pi}_{\omega}(Z, A)$. The infimum over square integrable predictors $h(\chi(Z), A)$ is*

$$\inf_{h} \mathbb{E}\big[(V - h(\chi(Z), A))^2\big] = \mathbb{E}[\mathrm{Var}(V \mid \chi(Z), A)]\,. \tag{48}$$

For $Z = (S, U)$ and $\chi(Z) = S$, this is the least possible squared value error of a critic that omits cumulative cost. At a fixed physical state and action, any common prediction $h(s, a)$ satisfies

$$\max_{u \in \{u_L, u_H\}} |h(s, a) - Q^{\pi}_{\omega}((s, u), a)| \geq \frac{1}{2}\,|Q^{\pi}_{\omega}((s, u_L), a) - Q^{\pi}_{\omega}((s, u_H), a)|\,. \tag{49}$$

The midpoint of the two values attains the latter bound.

Proof. Since $|Q^{\pi}_{\omega}| \leq G/(1 - \gamma)$, the random variable V has finite second moment. Set $m = \mathbb{E}[V \mid \chi(Z), A]$. For any admissible h, expand $V - h = (V - m) + (m - h)$. The cross term vanishes:

$$\mathbb{E}[(V - m)(m - h)] = \mathbb{E}[(m - h)\mathbb{E}[V - m \mid \chi(Z), A]] = 0.$$

Therefore

$$\mathbb{E}[(V - h)^2] = \mathbb{E}[\mathrm{Var}(V \mid \chi(Z), A)] + \mathbb{E}[(m - h)^2].$$

The first term is independent of h, and the second is minimized to zero by $h = m$, proving (48).

For (49), write the two values as v_L, v_H. The triangle inequality gives $|v_L - v_H| \leq |v_L - h(s, a)| + |h(s, a) - v_H|$. At least one error is therefore at least $|v_L - v_H|/2$, and their midpoint has exactly that error at both contexts. □

Proposition 5. *At a physical state s, suppose exactly two actions a_0, a_1 are available. Fix a continuation policy π and two cumulative costs $u_L < u_H$. Write $Q_j(a) = Q^\pi_\omega((s, u_j), a)$ for $j \in \{L, H\}$ and suppose the action rankings are opposite:*

$$\Delta_H = Q_H(a_0) - Q_H(a_1) > 0, \qquad \Delta_L = Q_L(a_1) - Q_L(a_0) > 0.$$

A decision rule using only s assigns the same probability p to a_0 at both contexts. Define its loss relative to the larger continuation value by

$$R_H(p) = Q_H(a_0) - [pQ_H(a_0) + (1-p)Q_H(a_1)],$$
$$R_L(p) = Q_L(a_1) - [pQ_L(a_0) + (1-p)Q_L(a_1)].$$

Then

$$\inf_{p\in[0,1]} \max\{R_L(p), R_H(p)\} = \frac{\Delta_L \Delta_H}{\Delta_L + \Delta_H} > 0. \tag{50}$$

A rule observing u can attain zero loss at both contexts by choosing the action with larger Q_j.

Proof. The definitions give $R_H(p) = (1-p)\Delta_H$ and $R_L(p) = p\Delta_L$. The first function decreases and the second increases. They intersect at $p^* = \Delta_H/(\Delta_H + \Delta_L)$. Below this point the maximum is R_H, which decreases toward the intersection; above it the maximum is R_L, which increases away from it. The intersection therefore minimizes their maximum and gives (50). Choosing a_0 at u_H and a_1 at u_L gives zero loss. □

E Error Propagation and Stability

E.1 From velocity error to endpoint error

Lemma 4. *Fix a context and a target path with velocity v^* whose ODE generates that path's marginals. Suppose the learned field $\widehat{v}$ is spatially Lipschitz with bound $L(t) \geq 0$ satisfying $\int_0^1 L(t)dt < \infty$, and that both ODE solutions exist and have square integrable endpoints. Let*

$$K_L = \left(\int_0^1 \exp\left(2\int_t^1 L(u)du\right) dt\right)^{1/2}, \qquad \epsilon^2 = \int_0^1 \mathbb{E}_{X_t}\|\widehat{v}(t, X_t) - v^*(t, X_t)\|^2 dt.$$

For the endpoint laws $\widehat{\nu}_1, \nu_1$ from the same source,

$$W_2(\widehat{\nu}_1, \nu_1) \leq K_L \epsilon. \tag{51}$$

For scalar endpoints, their mean difference has absolute value at most $K_L\epsilon$. A numerical endpoint approximation with root mean squared error ϵ_{num} adds at most ϵ_{num}.

Proof. Use the same initial source to couple the two ODE solutions $\widehat{X}_t, X_t$. Add and subtract $\widehat{v}(t, X_t)$ in the difference equation. The spatial Lipschitz condition implies

$$\|\widehat{X}_t - X_t\| \leq \int_0^t L(u)\|\widehat{X}_u - X_u\|du + \int_0^t \|\widehat{v}(u, X_u) - v^*(u, X_u)\|du.$$

The inhomogeneous Gronwall inequality gives

$$\|\widehat{X}_1 - X_1\| \leq \int_0^1 e^{\int_t^1 L(u)du}\|\widehat{v}(t, X_t) - v^*(t, X_t)\|dt.$$

Taking the L_2 norm over the coupled source and applying Minkowski's inequality and then Cauchy Schwarz yields

$$\begin{aligned}(\mathbb{E}\|\widehat{X}_1 - X_1\|^2)^{1/2} &\leq \int_0^1 e^{\int_t^1 L(u)du}(\mathbb{E}\|\widehat{v}(t, X_t) - v^*(t, X_t)\|^2)^{1/2}dt \\ &\leq K_L\epsilon.\end{aligned}$$

The target ODE has the prescribed path marginals, so this is the error integrated against the same distribution as the flow loss. The endpoint coupling is admissible for W_2, proving (51). For a scalar endpoint,

$$|\mathbb{E}\widehat{X}_1 - \mathbb{E}X_1| \leq (\mathbb{E}|\widehat{X}_1 - X_1|^2)^{1/2}.$$

The triangle inequality applied to a numerical endpoint coupled with $\widehat{X}_1$ adds the numerical error. □

E.2 From the implemented regression target to the Bellman update

At iteration k, fix $\bar{q}_k$ and let $m_k(x)$ be the mean of the regression target at $x = (z, a)$. Assume its difference from the physical Bellman target satisfies

$$\|m_k - T_{g_k}^{\pi_k} \bar{q}_k\|_\infty \le \epsilon_{\mathrm{data},k}.$$

Let $\widetilde{q}_{k+1}$ be the exact endpoint mean of the learned flow. Applying Lemma 4 at each context with uniform constants gives

$$\|\widetilde{q}_{k+1} - m_k\|_\infty \le K_k \epsilon_{Q,k}.$$

Here $\epsilon_{Q,k}$ bounds the conditional velocity error uniformly over inputs, and K_k is the corresponding endpoint error constant. If the implemented readout q_{k+1} obeys $\|q_{k+1} - \widetilde{q}_{k+1}\|_\infty \le \epsilon_{\mathrm{num},k}$, define the total update error

$$\zeta_k = K_k \epsilon_{Q,k} + \epsilon_{\mathrm{num},k} + \epsilon_{\mathrm{data},k} + \gamma \|\bar{q}_k - q_k\|_\infty. \tag{52}$$

The numerical term includes integration and finite sampling in the endpoint mean. The data term measures the error in the fitted target mean; the final term records target-critic lag. The triangle inequality, applied through $\widetilde{q}_{k+1}$, m_k, and $T_{g_k}^{\pi_k} \bar{q}_k$, yields

$$\begin{aligned} \|q_{k+1} - T_{g_k}^{\pi_k} q_k\|_\infty &\le \epsilon_{\mathrm{num},k} + K_k \epsilon_{Q,k} + \epsilon_{\mathrm{data},k} \\ &\quad + \|T_{g_k}^{\pi_k} \bar{q}_k - T_{g_k}^{\pi_k} q_k\|_\infty \\ &\le \epsilon_{\mathrm{num},k} + K_k \epsilon_{Q,k} + \epsilon_{\mathrm{data},k} + \gamma \|\bar{q}_k - q_k\|_\infty. \end{aligned}$$

Bellman contraction gives the last inequality, establishing (18).

E.3 Movement of the regularized policy target

Multiplier updates affect the actor even when its component critics are fixed. At a fixed context z, suppress the context argument. For bounded reward and cost scores q_r, q_c, let $\pi_\lambda \propto \beta \exp((q_r - \lambda q_c)/\tau)$ and $\Delta = \lambda' - \lambda$. The log normalizer gives

$$\mathrm{KL}(\pi_\lambda \| \pi_{\lambda'}) = \frac{\Delta^2}{\tau^2} \int_0^1 (1-u) \operatorname{Var}_{\pi_{\lambda + u\Delta}}(q_c)\, du. \tag{53}$$

For fixed z, q_r, and q_c, write

$$\psi(\lambda) = \log \mathbb{E}_\beta \exp\left(\frac{q_r - \lambda q_c}{\tau} \right).$$

On any bounded coefficient interval, bounded scores justify differentiating under expectation. Direct differentiation gives

$$\psi'(\lambda) = -\frac{1}{\tau} \mathbb{E}_{\pi_\lambda} q_c, \qquad \psi''(\lambda) = \frac{1}{\tau^2} \operatorname{Var}_{\pi_\lambda}(q_c). \tag{54}$$

The log density ratio satisfies

$$\log \frac{\pi_\lambda}{\pi_{\lambda'}} = \frac{(\lambda' - \lambda) q_c}{\tau} + \psi(\lambda') - \psi(\lambda).$$

Integrating against π_λ and using (54) yields

$$\mathrm{KL}(\pi_\lambda \| \pi_{\lambda'}) = \psi(\lambda') - \psi(\lambda) - \psi'(\lambda)(\lambda' - \lambda).$$

Taylor's formula with integral remainder proves (53) for either sign of Δ. In particular, if the variance is bounded below by $v_* > 0$ on the interval between the coefficients, then

$$\mathrm{KL}(\pi_\lambda \| \pi_{\lambda'}) \ge \frac{v_* \Delta^2}{2\tau^2}.$$

E.4 Proof of Theorem 1

Proof. For bounded q, q', the Bellman operator obeys

$$\|T_g^\pi q - T_g^\pi q'\|_\infty \le \gamma \|q - q'\|_\infty.$$

It has a unique bounded fixed point Q_g^π, and $\|Q_g^\pi\|_\infty \le G/(1-\gamma)$ when $|g| \le G$.

For a fixed policy and two utilities, insert and subtract $T_g^\pi Q_{g'}^\pi$ in the fixed point difference. Contraction gives

$$\|Q_g^\pi - Q_{g'}^\pi\|_\infty \le \gamma \|Q_g^\pi - Q_{g'}^\pi\|_\infty + \|g - g'\|_\infty.$$

Rearrangement proves

$$\|Q_g^\pi - Q_{g'}^\pi\|_\infty \le \frac{\|g - g'\|_\infty}{1-\gamma}. \tag{55}$$

For the policy term, use the convention $\mathrm{TV}(p, p') = \sup_B |p(B) - p'(B)|$. A bounded function f satisfies $|\mathbb{E}_p f - \mathbb{E}_{p'} f| \le 2\|f\|_\infty \mathrm{TV}(p, p')$. With a common utility,

$$\|Q_g^\pi - Q_g^{\pi'}\|_\infty \le \gamma \|Q_g^\pi - Q_g^{\pi'}\|_\infty + 2\gamma \|Q_g^{\pi'}\|_\infty \sup_z \mathrm{TV}(\pi(\cdot \mid z), \pi'(\cdot \mid z)).$$

It follows that

$$\|Q_g^\pi - Q_g^{\pi'}\|_\infty \le \frac{2\gamma G}{(1-\gamma)^2} \sup_z \mathrm{TV}(\pi(\cdot \mid z), \pi'(\cdot \mid z)). \tag{56}$$

To compare q_{k+1} with its new fixed point, first compare it with the current fixed point $Q_{g_k}^{\pi_k}$. The triangle inequality, (18), and contraction give

$$\begin{aligned} e_{k+1} &\le \|q_{k+1} - Q_{g_k}^{\pi_k}\|_\infty + \|Q_{g_k}^{\pi_k} - Q_{g_{k+1}}^{\pi_{k+1}}\|_\infty \\ &\le \zeta_k + \gamma e_k + \|Q_{g_k}^{\pi_k} - Q_{g_{k+1}}^{\pi_k}\|_\infty + \|Q_{g_{k+1}}^{\pi_k} - Q_{g_{k+1}}^{\pi_{k+1}}\|_\infty. \end{aligned}$$

Applying (55) and (56) proves (19). Induction gives the complete unrolling

$$e_K \le \gamma^K e_0 + \sum_{k=0}^{K-1} \gamma^{K-1-k} \left(\zeta_k + \frac{d_k}{1-\gamma} + \frac{2\gamma G}{(1-\gamma)^2} p_k \right). \tag{57}$$

For $g_k = r - \lambda_k c$ and bounded cost, $\|g_{k+1} - g_k\|_\infty \le C_{\max} \|\lambda_{k+1} - \lambda_k\|_\infty$ pointwise. For the fixed rule $g_k = g_\omega$, the utility difference is zero, so $d_k = 0$. □

E.5 Proof of Corollary 1

Proof. With $g_k = g_\omega$ and $\pi_k = \pi$, the Bellman fixed point is Q_ω^π at every iteration. Contraction and the update error give

$$e_{k+1} \le \|q_{k+1} - T_\omega^\pi q_k\|_\infty + \|T_\omega^\pi q_k - T_\omega^\pi Q_\omega^\pi\|_\infty \le \zeta_k + \gamma e_k.$$

Induction yields (20). If $\zeta_k \le \bar{\zeta}$, summing the geometric series gives (21). Taking a limit superior proves the stated asymptotic error bound.

For vanishing errors, fix $\varepsilon > 0$ and choose an index N such that $\zeta_k \le (1-\gamma)\varepsilon/2$ for all $k \ge N$. For $K > N$, split the sum in (20) as

$$e_K \le \gamma^K e_0 + \sum_{k=0}^{N-1} \gamma^{K-1-k} \zeta_k + \sum_{k=N}^{K-1} \gamma^{K-1-k} \zeta_k.$$

The last sum is at most $\varepsilon/2$. The first two terms tend to zero because they contain a fixed finite number of finite coefficients multiplied by powers of γ. For all sufficiently large K they sum to at most $\varepsilon/2$, so $e_K \le \varepsilon$. This proves uniform convergence of the value estimates to Q_ω^π. □

E.6 EXACT TRACKING EXAMPLES

Consider one state and one action, $r = 0$, $c = 1$, and $q_{k+1} = -\lambda_k + \gamma q_k$. Its instantaneous fixed point is $-\lambda_k/(1-\gamma)$. Define the signed residual $h_k = q_k + \lambda_k/(1-\gamma)$. Direct substitution gives

$$h_{k+1} = \gamma h_k + \frac{\lambda_{k+1} - \lambda_k}{1-\gamma}. \tag{58}$$

For $\lambda_k = \bar{\lambda} + (-1)^k \delta$ with $0 < \delta \leq \bar{\lambda}$, the particular solution

$$h_k^{\mathrm{per}} = \frac{2\delta}{1-\gamma^2}(-1)^k$$

satisfies (58). Every other solution differs from it by $\gamma^k(h_0 - h_0^{\mathrm{per}})$. Consequently

$$\lim_{k\to\infty} \left| q_k + \frac{\lambda_k}{1-\gamma} \right| = \frac{2\delta}{1-\gamma^2} > 0. \tag{59}$$

For a finite nondecreasing schedule and $h_0 = 0$, the same recursion yields

$$|h_K| = \frac{1}{1-\gamma} \sum_{k=0}^{K-1} \gamma^{K-1-k} |\lambda_{k+1} - \lambda_k|.$$

Thus the accumulated utility movement term in (57) is attained exactly. A frozen coefficient instead gives $h_k = \gamma^k h_0$.

E.7 COMPLETE VALUE ERROR AND POLICY CONSEQUENCES

To account for error in the estimated rule, let g° be the intended utility and assume $\|g_\omega - g^\circ\|_\infty \leq \epsilon_g$. Under the fixed-policy conditions of Corollary 1, with $\zeta_k \leq \bar{\zeta}$ for all k,

$$\|q_K - Q^\pi_{g^\circ}\|_\infty \leq \gamma^K e_0 + \frac{1-\gamma^K}{1-\gamma}\bar{\zeta} + \frac{\epsilon_g}{1-\gamma} =: \Delta_Q. \tag{60}$$

Let $q^\circ = Q^\pi_{g^\circ}$ and $\pi^\circ = \pi_{q^\circ}$. Define the variation of critic error across actions by

$$R_Q = \sup_z \left\{ \sup_a [q_K(z,a) - q^\circ(z,a)] - \inf_a [q_K(z,a) - q^\circ(z,a)] \right\}. \tag{61}$$

Equation (60) gives $R_Q \leq 2\Delta_Q$.

Proposition 6 (Policy sensitivity to critic and actor errors)**.** *Assume $R_Q < \infty$ and finite objectives in* (13)*. At every context,*

$$\mathcal{J}_z(\pi^\circ; q^\circ) - \mathcal{J}_z(\pi_{q_K}; q^\circ) \leq \frac{R_Q^2}{8\tau}, \tag{62}$$

$$\mathrm{TV}(\pi_{q_K}, \pi^\circ) \leq \min\{1, R_Q/(4\tau)\}. \tag{63}$$

Suppose all reference and executed actions lie in a Euclidean action space of diameter D_A and the learned actor obeys $\sup_z W_1(\widehat{\pi}(\cdot \mid z), \pi_{q_K}(\cdot \mid z)) \leq \delta_\pi$. If $Q_h^{\pi^\circ}(z,\cdot)$ is uniformly L_h Lipschitz for $h \in \{r, c\}$, then

$$|J_h(\widehat{\pi}) - J_h(\pi^\circ)| \leq \frac{L_h}{1-\gamma} \left[\delta_\pi + D_A \min\{1, R_Q/(4\tau)\} \right]. \tag{64}$$

Under the uniform transport conditions of Lemma 4, actor velocity error ϵ_π gives $\delta_\pi \leq K_\pi \epsilon_\pi + \epsilon_{\mathrm{num},\pi}$.

If the ideal target policy π° has a safety margin $m \geq 0$, meaning $J_c(\pi^\circ) \leq \ell - m$, then

$$J_c(\widehat{\pi}) \leq \ell - m + \frac{L_c}{1-\gamma} \left[\delta_\pi + D_A \min\{1, R_Q/(4\tau)\} \right]. \tag{65}$$

To prove the value bound, insert $Q^\pi_{g_\omega}$ between q_K and $Q^\pi_{g^\circ}$. Equation (21) bounds the first difference and (55) bounds the second by $\epsilon_g/(1-\gamma)$. This proves (60).

Proof of Proposition 6. Sensitivity to relative action values. Fix z and write $\delta(a) = q_K(z,a) - q^\circ(z,a)$. Let its action oscillation be $r_z \le R_Q$. Adding a constant to δ leaves the tilted policy unchanged, so we may center it in an interval of radius $r_z/2$. Define

$$\psi(t) = \log \mathbb{E}_{a\sim\pi^\circ} e^{t\delta(a)/\tau}, \qquad p_t(da) = e^{t\delta(a)/\tau - \psi(t)}\pi^\circ(da), \qquad 0 \le t \le 1.$$

Then $p_0 = \pi^\circ$ and $p_1 = \pi_{q_K}$. Boundedness of the centered δ justifies differentiation, giving

$$\psi'(t) = \frac{\mathbb{E}_{p_t}\delta}{\tau}, \qquad \psi''(t) = \frac{\mathrm{Var}_{p_t}(\delta)}{\tau^2} \le \frac{r_z^2}{4\tau^2}.$$

For the variance bound, if c is the midpoint of the range of δ, then $\mathrm{Var}_{p_t}(\delta) = \inf_b \mathbb{E}_{p_t}(\delta - b)^2 \le \mathbb{E}_{p_t}(\delta - c)^2 \le r_z^2/4$.

The log density ratio of p_1 and p_0 is $\delta/\tau - \psi(1) + \psi(0)$. Its expectation under p_1 is

$$\begin{aligned}\mathrm{KL}(p_1\|p_0) &= \psi'(1) - \psi(1) + \psi(0) \\ &= \int_0^1 [\psi'(1) - \psi'(s)]\, ds = \int_0^1 t\psi''(t)\, dt \le \frac{r_z^2}{8\tau^2}.\end{aligned}$$

Apply (37) to q° and p_1, then multiply this bound by τ. This proves (62) with $r_z \le R_Q$.

For total variation, the density derivative with respect to the common dominating measure is

$$\partial_t p_t = \frac{1}{\tau} p_t(\delta - \mathbb{E}_{p_t}\delta).$$

Integrating the derivative, applying the triangle inequality, and using the variance bound gives

$$\begin{aligned}\mathrm{TV}(p_1,p_0) &\le \frac{1}{2\tau}\int_0^1 \mathbb{E}_{p_t}|\delta - \mathbb{E}_{p_t}\delta|\, dt \\ &\le \frac{1}{2\tau}\int_0^1 \sqrt{\mathrm{Var}_{p_t}(\delta)}\, dt \le \frac{r_z}{4\tau}.\end{aligned}$$

Total variation is also at most one, proving (63). If $r_z = 0$, the tilted law is exactly p_0, so both bounds vanish as required.

From policy distance to return difference. On an action space with diameter D_A, couple the common part of p_0, p_1 identically and the unmatched parts arbitrarily. The unmatched mass is their total variation, and its transport distance is at most D_A. Hence $W_1(p_1,p_0) \le D_A\,\mathrm{TV}(p_1,p_0)$. The triangle inequality with the learned actor gives, uniformly in z,

$$W_1(\widehat{\pi}(\cdot \mid z), \pi^\circ(\cdot \mid z)) \le \delta_\pi + D_A \min\{1, R_Q/(4\tau)\} =: \delta_{\mathrm{tot}}. \tag{66}$$

For $h \in \{r, c\}$, let $V_h^{\pi^\circ}(z) = \mathbb{E}_{a\sim\pi^\circ} Q_h^{\pi^\circ}(z,a)$. The Bellman equation along a trajectory from $\widehat{\pi}$ gives

$$\mathbb{E}[h_j + \gamma V_h^{\pi^\circ}(z_{j+1}) - V_h^{\pi^\circ}(z_j)] = \mathbb{E}[Q_h^{\pi^\circ}(z_j,a_j) - V_h^{\pi^\circ}(z_j)].$$

Multiply by γ^j, sum through $j = N$, and let $N \to \infty$. Bounded rewards and costs make the terminal term vanish. For the common initial context distribution specified in Appendix D.3,

$$J_h(\widehat{\pi}) - J_h(\pi^\circ) = \frac{1}{1-\gamma}\mathbb{E}_{z\sim d^{\widehat{\pi}}}\left[\mathbb{E}_{\widehat{\pi}(\cdot|z)} Q_h^{\pi^\circ}(z,\cdot) - \mathbb{E}_{\pi^\circ(\cdot|z)} Q_h^{\pi^\circ}(z,\cdot)\right],$$

where $d^{\widehat{\pi}}$ is normalized discounted occupancy. For any coupling of the two action laws, Lipschitz continuity bounds the absolute difference of the two expectations by L_h times the expected action distance. Taking the infimum over couplings bounds it by $L_h W_1$. Equation (66) now proves (64). Inserting the assumed reference safety margin gives (65).

Finally, $W_1 \le W_2$ and Lemma 4 give $\delta_\pi \le K_\pi \epsilon_\pi + \epsilon_{\mathrm{num},\pi}$ when the constants and errors are uniform across contexts.

□

F Capacity and Scalar Critic Approximation

F.1 The equality case for linear classes

Let μ_{cap} be a probability distribution over critic inputs, and let $\mathcal{F}$ be a common closed linear subspace of $L_2(\mu_{\mathrm{cap}})$. Its orthogonal projection $P_{\mathcal{F}}$ is linear, hence

$$P_{\mathcal{F}}(Q_r - Q_c) = P_{\mathcal{F}}Q_r - P_{\mathcal{F}}Q_c.$$

The direct fit and the difference of separately optimal fits coincide in this class. The rank-constrained class used below is nonlinear.

F.2 Proof of Theorem 2

Let $x = (i, j)$ index mn critic inputs, sampled uniformly from μ_{cap}. For a matrix M, define $M(x) = M_{ij}$ and $\|M\|_{L_2(\mu_{\mathrm{cap}})} = \|M\|_F/\sqrt{mn}$. Let $S, T \in \mathbb{R}^{m\times n}$ be the shared component and the target difference. With per-model rank limit r_0, assume

$$\begin{aligned} S^\top T = 0, \quad ST^\top = 0, \quad \mathrm{rank}(S) = r_0, \\ 1 \le \mathrm{rank}(T) \le r_0, \qquad \sigma_{r_0}(S) > \|T\|_{\mathrm{op}}. \end{aligned} \tag{67}$$

Here $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_{\mathrm{op}}$ is the largest singular value, and $\sigma_{r_0}(S)$ is the smallest nonzero singular value of S. Thus S and T occupy orthogonal row and column spaces, with every nonzero singular value of S larger than those of T.

At each x, let (Y_r, Y_c) be jointly Gaussian with means $((S + T)(x), S(x))$, variances $\sigma^2 > 0$, and covariance $\sigma^2/2$. The combined target $Y_U = Y_r - Y_c$ has mean $T(x)$ and the same variance.

The common flow family uses an independent source $E \sim \mathcal{N}(0, \sigma^2)$, uniform flow time on $[0, 1]$, and independent source–target pairs. For a matrix M of rank at most r_0, define

$$\begin{aligned} v_M(t, y, x) = a_t y + b_t M(x), \qquad \mathrm{rank}(M) \le r_0, \\ a_t = \frac{2t-1}{(1-t)^2 + t^2}, \qquad b_t = \frac{1-t}{(1-t)^2 + t^2}. \end{aligned} \tag{68}$$

Here y is the scalar interpolant and x is the critic-input index; every critic has rank limit r_0.

Lemma 5. *Under* (67)*, let M_r^*, M_c^*, M_U^* denote the endpoint means minimizing the respective population flow matching losses in* (68)*. These minimizing means are unique and satisfy*

$$M_r^* = M_c^* = S, \qquad M_U^* = T, \tag{69}$$

and

$$\|M_U^* - T\|_{L_2(\mu_{\mathrm{cap}})} = 0, \qquad \|M_r^* - M_c^* - T\|_{L_2(\mu_{\mathrm{cap}})} = \frac{\|T\|_F}{\sqrt{mn}} > 0. \tag{70}$$

For learned means satisfying $\|\widehat{M}_i - M_i^*\|_{L_2(\mu_{\mathrm{cap}})} \le \delta_i$, $i \in \{U, r, c\}$, direct fitting still has strictly smaller error whenever $\delta_U + \delta_r + \delta_c < \|T\|_F/\sqrt{mn}$.

To obtain Theorem 2, identify $q_i^*(x) = M_i^*(x)$ for $i \in \{r, c, U\}$, $m(x) = T(x)$, and $\Delta = \|T\|_F/\sqrt{mn}$. The assumptions are nonempty: with $m = n = 2$ and $r_0 = 1$, the matrices $S = \mathrm{diag}(2, 0)$ and $T = \mathrm{diag}(0, 1)$ satisfy (67) and give $\Delta = 1/2$. The Gaussian targets have finite second moments. The lemma therefore proves the existence statement in the main text, while retaining the explicit dependence of the gap on the target means.

Proof. Conditional flow objective. Fix a context and one of the three targets. Let its true mean be m_* and variance be σ^2. The independent source has mean zero and the same variance. Writing $d_t = (1 - t)^2 + t^2$, Gaussian moments give

$$\mathbb{E}[X_t] = tm_*, \quad \mathrm{Var}(X_t) = \sigma^2 d_t, \quad \mathbb{E}[D] = m_*, \quad \mathrm{Cov}(D, X_t) = \sigma^2(2t - 1).$$

Therefore

$$v^*(t, y, x) = m_* + \frac{2t-1}{d_t}(y - tm_*)$$

$$= a_t y + b_t m_*,$$

where the last identity uses $d_t - t(2t-1) = 1 - t$. Proposition 2 now gives

$$\mathcal{L}(v_M) - \mathcal{L}(v_{M_*}) = c_{\mathrm{FM}} \|M - M_*\|^2_{L_2(\mu_{\mathrm{cap}})}, \qquad c_{\mathrm{FM}} = \int_0^1 \frac{(1-t)^2}{((1-t)^2 + t^2)^2} dt = \frac{\pi}{4}. \tag{71}$$

For the integral, substitute $u = 2t - 1$. The odd term vanishes over $[-1, 1]$ and the remainder is $\int_0^1 (1 + u^2)^{-1} du = \pi/4$.

The ODE with field v_M can also be solved explicitly:

$$X_t = tM(x) + \sqrt{d_t}\, E, \qquad E \sim \mathcal{N}(0, \sigma^2). \tag{72}$$

Indeed, differentiating the right side gives $M + (2t-1)E/\sqrt{d_t}$, while substitution in $a_t X_t + b_t M$ gives the same expression because $t a_t + b_t = 1$. Its initial law is the chosen source and its endpoint is $M(x) + E$, so the endpoint mean is $M(x)$.

The covariance assumption ensures that the combined target has the same variance:

$$\mathrm{Var}(Y_r - Y_c \mid x) = \sigma^2 + \sigma^2 - 2(\sigma^2/2) = \sigma^2.$$

Thus (71) applies with the same positive constant to all three training problems. Uniform contexts give $\|M - M_*\|^2_{L_2(\mu_{\mathrm{cap}})} = \|M - M_*\|^2_F/(mn)$.

Optimal rank-constrained fits. Let $N = S + T$. The orthogonality conditions imply that the column spaces of S, T are orthogonal and that their row spaces are orthogonal. In particular,

$$NN^\top = SS^\top + TT^\top.$$

The nonzero singular values of N are the singular values of S together with those of T. Since $\sigma_{r_0}(S) > \|T\|_{\mathrm{op}}$, its largest r_0 singular directions are exactly those of S.

Consider any matrix M of rank at most r_0 and let P be the orthogonal projection onto its column space. Orthogonality gives

$$\begin{aligned} \|N - M\|_F^2 &= \|(I - P)N\|_F^2 + \|PN - M\|_F^2 \\ &\geq \|N\|_F^2 - \mathrm{tr}(PNN^\top). \end{aligned}$$

If $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ are the eigenvalues of $NN^\top$, then

$$\mathrm{tr}(PNN^\top) = \sum_{i=1}^m \lambda_i p_i, \qquad 0 \leq p_i \leq 1, \quad \sum_i p_i \leq r_0.$$

The sum is at most $\sum_{i=1}^{r_0} \lambda_i$: moving any mass from a smaller eigenvalue to an unfilled larger one can only increase it. Thus $\|N - M\|_F^2 \geq \|T\|_F^2$. Equality is attained by the projection onto the column space of S, with $PN = S$ and $M = S$. The strict singular value gap forces this leading subspace, so the best fitted matrix is $M_r^* = S$. The cost mean S and combined mean T already have rank at most r_0, giving unique zero error fits $M_c^* = S$ and $M_U^* = T$.

Subtracting the component fits gives zero, while the true difference is T. Dividing the squared Frobenius error by mn proves (70).

Learned models. Let the learned mean matrices be $\widehat{M}_U, \widehat{M}_r, \widehat{M}_c$. The direct error is at most δ_U. For the separate predictor,

$$\begin{aligned} \|\widehat{M}_r - \widehat{M}_c - T\|_{L_2(\mu_{\mathrm{cap}})} &\geq \|M_r^* - M_c^* - T\|_{L_2(\mu_{\mathrm{cap}})} \\ &\quad - \|\widehat{M}_r - M_r^*\|_{L_2(\mu_{\mathrm{cap}})} - \|\widehat{M}_c - M_c^*\|_{L_2(\mu_{\mathrm{cap}})} \\ &\geq \frac{\|T\|_F}{\sqrt{mn}} - \delta_r - \delta_c. \end{aligned}$$

The condition in the theorem makes this lower bound strictly larger than δ_U, completing the actual error comparison. □

F.3 THE STRICT GAP PERSISTS UNDER TARGET PERTURBATIONS

Corollary 2 (Robustness of the critic approximation gap)**.** *Let S, T satisfy* (67)*, and set*

$$a_* = \sigma_{r_0}(S), \qquad b_* = \|T\|_{\mathrm{op}}, \qquad c_* = 1 - b_*/a_* > 0, \qquad t_* = \|T\|_{L_2(\mu_{\mathrm{cap}})}.$$

In the same Gaussian flow class, replace the component means by

$$N_r = S + T + H_r, \qquad N_c = S + H_c, \qquad N_U = T + H_r - H_c.$$

Keep the conditional variances and covariance of the construction in Lemma 5. Let P_r, P_c, P_U be any population minimizing mean matrices for these targets. Write $\rho_r = \|H_r\|_{L_2(\mu_{\mathrm{cap}})}$, $\rho_c = \|H_c\|_{L_2(\mu_{\mathrm{cap}})}$, and $\rho_\Delta = \|H_r - H_c\|_{L_2(\mu_{\mathrm{cap}})}$. Then

$$\|P_U - N_U\|_{L_2(\mu_{\mathrm{cap}})} \le \rho_\Delta, \tag{73}$$

$$\|P_r - P_c - N_U\|_{L_2(\mu_{\mathrm{cap}})} \ge \left[t_* - \rho_\Delta - \frac{2\rho_r}{c_*} - 2\rho_c \right]_+ . \tag{74}$$

If trained means have distances at most $\delta_U, \delta_r, \delta_c$ from these minimizers, direct learning has strictly smaller actual value error whenever

$$2\rho_\Delta + \frac{2\rho_r}{c_*} + 2\rho_c + \delta_U + \delta_r + \delta_c < t_*. \tag{75}$$

Proof. The Gaussian flow risk identity (71) still holds, because only the means have changed. Thus population fitting is again best rank r_0 approximation. We first control the sensitivity of the reward fit without assuming uniqueness after perturbation.

Set $N_0 = S + T$. For every matrix M of rank at most r_0, we claim

$$\|N_0 - M\|_F^2 - \|T\|_F^2 \ge c_* \|M - S\|_F^2. \tag{76}$$

To prove this, consider $S + (a_*/b_*)T$. Here $b_* > 0$ because $T \neq 0$. Orthogonality makes its singular values the union of those of S and $(a_*/b_*)T$. All singular values of S are at least a_* and all of the scaled T are at most a_*. The rank approximation argument in the preceding proof therefore shows that S is a best rank r_0 approximation, even if there is a tie at the boundary. Hence

$$\|S + (a_*/b_*)T - M\|_F^2 \ge \|(a_*/b_*)T\|_F^2.$$

Using $\langle S, T \rangle_F = 0$ and expanding gives

$$2\langle T, M - S \rangle_F \le (b_*/a_*)\|M - S\|_F^2.$$

Expand $\|S + T - M\|_F^2$ and apply this inequality to obtain (76).

Since P_r is optimal for $N_0 + H_r$ and S is an admissible competitor,

$$\|N_0 + H_r - P_r\|_F^2 \le \|N_0 + H_r - S\|_F^2.$$

Cancel the $\|H_r\|_F^2$ terms, expand, and use (76):

$$c_* \|P_r - S\|_F^2 \le 2\langle H_r, P_r - S \rangle_F \le 2\|H_r\|_F \|P_r - S\|_F.$$

If the last factor is nonzero, divide by it; the zero case is immediate. Thus $\|P_r - S\|_F \le 2\|H_r\|_F / c_*$. For the cost fit, optimality against S gives $\|S + H_c - P_c\|_F \le \|H_c\|_F$, and the triangle inequality yields $\|P_c - S\|_F \le 2\|H_c\|_F$. Both bounds remain valid after dividing by $\sqrt{mn}$.

The unified target has the admissible competitor T, so $\|P_U - N_U\|_{L_2(\mu_{\mathrm{cap}})} \le \|H_r - H_c\|_{L_2(\mu_{\mathrm{cap}})}$. For separate fitting, write

$$P_r - P_c - N_U = (P_r - S) - (P_c - S) - T - (H_r - H_c).$$

The reverse triangle inequality and the two sensitivity bounds prove (74). The learned unified error is at most $\rho_\Delta + \delta_U$, whereas the learned separate error is at least $t_* - \rho_\Delta - 2\rho_r/c_* - 2\rho_c - \delta_r - \delta_c$. Condition (75) makes the latter strictly larger than the former. □

G Composition of Conditional Policy Flows

We compare direct flow training and additive composition for the regularized actor objective (13).

G.1 Policy targets and additive composition

Fix estimates q_r, q_c of the weighted values $Q^\pi_{r,\omega}, Q^\pi_{c,\omega}$ under a common evaluated policy π. For $q = q_r - q_c$, the reference policy β and temperature $\tau > 0$ define the target

$$\pi_U(da \mid z) = \frac{\exp((q_r(z,a) - q_c(z,a))/\tau)}{Z_U(z)}\beta(da \mid z). \tag{77}$$

where $Z_U(z)$ normalizes the law. By (37), π_U maximizes $\mathcal{J}_z(\cdot; q)$. Direct training fits a flow to π_U; separate training fits reward and safety flows to

$$\pi_r(da \mid z) = \frac{e^{q_r(z,a)/\tau}}{Z_r(z)}\beta(da \mid z), \qquad \pi_c(da \mid z) = \frac{e^{-q_c(z,a)/\tau}}{Z_c(z)}\beta(da \mid z), \tag{78}$$

where Z_r, Z_c are normalizers. Using the same source p_0^π and independent linear interpolation, let v_U, v_r, v_c, v_0 be the population velocity fields for $\pi_U, \pi_r, \pi_c, \beta$, respectively. Additive composition gives

$$v_{\mathrm{mix}} = v_r + v_c - v_0. \tag{79}$$

The subtraction counts the reference field once. Theorem 3 gives the velocity error of this construction, and Corollary 3 computes its endpoint error.

Composition error for general targets. Write the endpoint weights as

$$w_r = e^{q_r/\tau}, \qquad w_c = e^{-q_c/\tau}, \qquad w_U = w_r w_c. \tag{80}$$

For a context distribution μ, the reference sampling law draws $z \sim \mu$, $A_1 \sim \beta(\cdot \mid z)$, independent $A_0 \sim p_0^\pi$, and independent $t \sim \mathrm{Unif}[0,1]$. Set $A_t = (1-t)A_0 + tA_1$, $D = A_1 - A_0$, and $H = (t, A_t, z)$. Lemma 1 gives, for a positive endpoint weight $w = w(z, A_1)$,

$$v_w(H) = \frac{\mathbb{E}_{\mathrm{ref}}[wD \mid H]}{\mathbb{E}_{\mathrm{ref}}[w \mid H]}. \tag{81}$$

Here $\mathbb{E}_{\mathrm{ref}}$ uses the reference law; $w = 1, w_r, w_c, w_U$ give v_0, v_r, v_c, v_U, respectively.

For $w_i = w_i(z, A_1)$, define

$$R_i = \frac{w_i}{\mathbb{E}_{\mathrm{ref}}[w_i \mid H]}, \qquad u_i = v_i - v_0, \quad i \in \{r, c\}, \tag{82}$$

$$\chi = \mathbb{E}_{\mathrm{ref}}[(R_r - 1)(R_c - 1) \mid H],$$
$$h = \mathbb{E}_{\mathrm{ref}}[(D - v_0)(R_r - 1)(R_c - 1) \mid H]. \tag{83}$$

Each R_i has conditional mean one, so $1 + \chi = \mathbb{E}_{\mathrm{ref}}[R_r R_c \mid H] > 0$.

Let ν_U be the law of H with $A_1 \sim \pi_U(\cdot \mid z)$ and the same source, context distribution, and flow time. By (36), excess flow matching risk for this target is $\|v - v_U\|^2_{L_2(\nu_U)}$.

Theorem 3 (Composition error for the combined policy target)**.** *Assume finite second moments of the source and weighted endpoint laws, finite displayed conditional moments and normalizers with positive denominators, and square integrability of the fields under ν_U. Then, ν_U-almost everywhere,*

$$v_U - v_{\mathrm{mix}} = \frac{h - \chi(u_r + u_c)}{1 + \chi}. \tag{84}$$

Set $B = \|v_U - v_{\mathrm{mix}}\|_{L_2(\nu_U)}$. For learned approximations $\widehat{v}_i$ to v_i, $i \in \{U, r, c, 0\}$, define

$$\epsilon_U = \|\widehat{v}_U - v_U\|_{L_2(\nu_U)}, \qquad E = \sum_{i \in \{r,c,0\}} \|\widehat{v}_i - v_i\|_{L_2(\nu_U)}.$$

Their composition $\widehat{v}_{\mathrm{mix}} = \widehat{v}_r + \widehat{v}_c - \widehat{v}_0$ satisfies

$$\|\widehat{v}_{\mathrm{mix}} - v_U\|_{L_2(\nu_U)} \geq [B - E]_+. \tag{85}$$

If $\epsilon_U + E < B$, the learned unified field $\widehat{v}_U$ has strictly smaller excess flow matching risk for π_U than $\widehat{v}_{\mathrm{mix}}$. If $B > 0$ and every field is fitted exactly, their excess risks are zero and B^2, respectively.

Appendix G.3 relates the component training errors to the common norm used here.

G.2 Proof of Theorem 3

Proof. All conditional expectations in the first part are under the reference sampling distribution given H. By construction, $\mathbb{E}[R_r \mid H] = \mathbb{E}[R_c \mid H] = 1$. The identity

$$R_r R_c = 1 + (R_r - 1) + (R_c - 1) + (R_r - 1)(R_c - 1)$$

implies $\mathbb{E}[R_r R_c \mid H] = 1 + \chi > 0$. Lemma 1 and $\mathbb{E}[D - v_0 \mid H] = 0$ give

$$\begin{aligned} u_r &= \mathbb{E}[(D - v_0)(R_r - 1) \mid H], \\ u_c &= \mathbb{E}[(D - v_0)(R_c - 1) \mid H], \\ v_U - v_0 &= \frac{\mathbb{E}[(D - v_0) R_r R_c \mid H]}{1 + \chi} = \frac{u_r + u_c + h}{1 + \chi}. \end{aligned}$$

Subtracting $v_{\mathrm{mix}} - v_0 = u_r + u_c$ proves

$$v_U - v_{\mathrm{mix}} = \frac{h - \chi(u_r + u_c)}{1 + \chi}.$$

Thus exact mixing is correct if and only if $h = \chi(u_r + u_c)$ almost everywhere under the unified path law.

For the learned fields, the triangle inequality in $L_2(\nu_U)$ gives

$$\|\widehat{v}_{\mathrm{mix}} - v_{\mathrm{mix}}\|_{L_2(\nu_U)} \le \sum_{i \in \{r,c,0\}} \|\widehat{v}_i - v_i\|_{L_2(\nu_U)} = E.$$

The reverse triangle inequality then yields

$$\begin{aligned} \|\widehat{v}_{\mathrm{mix}} - v_U\|_{L_2(\nu_U)} &\ge \|v_{\mathrm{mix}} - v_U\|_{L_2(\nu_U)} - \|\widehat{v}_{\mathrm{mix}} - v_{\mathrm{mix}}\|_{L_2(\nu_U)} \\ &\ge B - E. \end{aligned}$$

Nonnegativity of the norm proves (85). If $\epsilon_U + E < B$, the right side is strictly larger than ϵ_U. Squaring both sides and using (36) for the unified weight proves the comparison of actual excess risks. □

G.3 Error transfer between path distributions

Let ν_i be a component path law and assume $\nu_U \ll \nu_i$ with $d\nu_U/d\nu_i \le C_i$. For any measurable error field f_i,

$$\|f_i\|^2_{L_2(\nu_U)} = \int \|f_i\|^2 \frac{d\nu_U}{d\nu_i}\, d\nu_i \le C_i \|f_i\|^2_{L_2(\nu_i)}.$$

Applying this to $f_i = \widehat{v}_i - v_i$ proves the bound on E used in Theorem 3.

The constants can be controlled from endpoint density ratios. At a fixed context,

$$\frac{d\pi_U}{d\pi_r}(a) = \frac{w_c(a)}{\mathbb{E}_{\pi_r} w_c}, \qquad \frac{d\pi_U}{d\pi_c}(a) = \frac{w_r(a)}{\mathbb{E}_{\pi_c} w_r}. \tag{86}$$

Define $\mathrm{osc}(f) = \sup_z(\sup_a f(z,a) - \inf_a f(z,a))$. If these oscillations are finite, the ratios are bounded by

$$C_r = e^{\mathrm{osc}(q_c)/\tau}, \qquad C_c = e^{\mathrm{osc}(q_r)/\tau}, \qquad C_0 = e^{\mathrm{osc}(q_r - q_c)/\tau}. \tag{87}$$

For example, if $q_c(z,\cdot) \in [m_c, M_c]$, the numerator in the first ratio of (86) is at most $e^{-m_c/\tau}$ and its denominator is at least $e^{-M_c/\tau}$. The other bounds follow in the same way.

To obtain the interpolation distribution, let $K_t(B \mid a, z)$ be the probability that $(1-t)A_0 + ta \in B$ under the common source. For every measurable set B,

$$p_t^U(B \mid z) = \int K_t(B \mid a, z) \pi_U(da \mid z) \le C_i \int K_t(B \mid a, z) \pi_i(da \mid z) = C_i p_t^i(B \mid z).$$

Integration over the common context distribution and flow time proves the path density ratio bound.

G.4 AN EXACT ENDPOINT COMPARISON

Corollary 3 (An exact separation of generated policies)**.** *Consider a single context with $p_0^\pi = \beta = \mathcal{N}(0,1)$, $\tau = 1$, $q_r(a) = -\eta a^2/2$, and $q_c(a) = \eta a^2/2$, where $\eta > 0$. Under linear interpolation of independently sampled endpoints, the unified target and the law generated by exact component composition are*

$$\pi_U = \mathcal{N}\left(0, \frac{1}{1+2\eta}\right), \qquad \pi_{\mathrm{mix}} = \mathcal{N}\left(0, \frac{1}{(1+\eta)^2}\right). \tag{88}$$

In particular,

$$W_2(\pi_{\mathrm{mix}}, \pi_U) = \frac{1}{\sqrt{1+2\eta}} - \frac{1}{1+\eta} > 0, \tag{89}$$

$$\mathcal{J}(\pi_U; q) - \mathcal{J}(\pi_{\mathrm{mix}}; q) = \tfrac{1}{2}(r_\eta - 1 - \log r_\eta) > 0, \qquad r_\eta = \frac{1+2\eta}{(1+\eta)^2}. \tag{90}$$

A directly fitted realizable flow generates π_U exactly.

PROOF OF COROLLARY 3

Proof. Both component weights are $e^{-\eta a^2/2}$. Multiplying either one by the standard normal density gives $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = (1+\eta)^{-1}$. Multiplying both gives variance $\sigma_U^2 = (1+2\eta)^{-1}$.

For an independent linear interpolation from $A_0 \sim \mathcal{N}(0,1)$ to $A_1 \sim \mathcal{N}(0, s^2)$, set $S_s(t) = (1-t)^2 + t^2 s^2$. Then

$$\mathrm{Var}(A_t) = S_s(t), \qquad \mathrm{Cov}(A_1 - A_0, A_t) = -(1-t) + ts^2 = \tfrac{1}{2} S_s'(t).$$

Gaussian conditional expectation gives

$$v_s(t,a) = k_s(t)a, \qquad k_s(t) = \frac{S_s'(t)}{2S_s(t)}.$$

The denominator is strictly positive on $[0,1]$, so the linear ODE is well defined and

$$\exp\left(\int_0^1 k_s(t)dt\right) = \sqrt{\frac{S_s(1)}{S_s(0)}} = s.$$

The mixed field has coefficient $2k_\sigma - k_1$ and endpoint scale

$$\exp\left(\int_0^1 (2k_\sigma(t) - k_1(t))dt\right) = \sigma^2.$$

It therefore transforms a standard normal source into variance $\sigma^4 = (1+\eta)^{-2}$, proving (88).

For centered scalar Gaussians with standard deviations s_1, s_2, the common standard normal coupling attains mean squared distance $(s_1 - s_2)^2$. Conversely, every coupling satisfies

$$\mathbb{E}(X-Y)^2 = s_1^2 + s_2^2 - 2\mathbb{E}[XY] \geq s_1^2 + s_2^2 - 2s_1 s_2 = (s_1 - s_2)^2$$

by the Cauchy Schwarz inequality. Thus the Wasserstein distance equals $|s_1 - s_2|$. Since $(1+\eta)^2 > 1 + 2\eta$, (89) follows with strict positivity.

For two centered Gaussians with variances s_m^2, s_U^2, integrating the log density ratio gives

$$\mathrm{KL}(\mathcal{N}(0, s_m^2) \| \mathcal{N}(0, s_U^2)) = \tfrac{1}{2}\left(\frac{s_m^2}{s_U^2} - 1 - \log \frac{s_m^2}{s_U^2}\right).$$

Their variance ratio here is $r_\eta \in (0,1)$. The function $r - 1 - \log r$ is strictly positive for $r \neq 1$. Applying (37) with $\tau = 1$ proves (90). The direct population velocity v_{σ_U} generates the correct endpoint, establishing the final claim. □

H ILLUSTRATIVE DRIVING ROLLOUTS

We present recorded examples of speed response, overtaking, and behavior in the presence of a vehicle ahead. The recordings are grouped by the cost-allowance settings reported in the evaluation. They provide qualitative observations that complement the task-level reward and cost estimates.

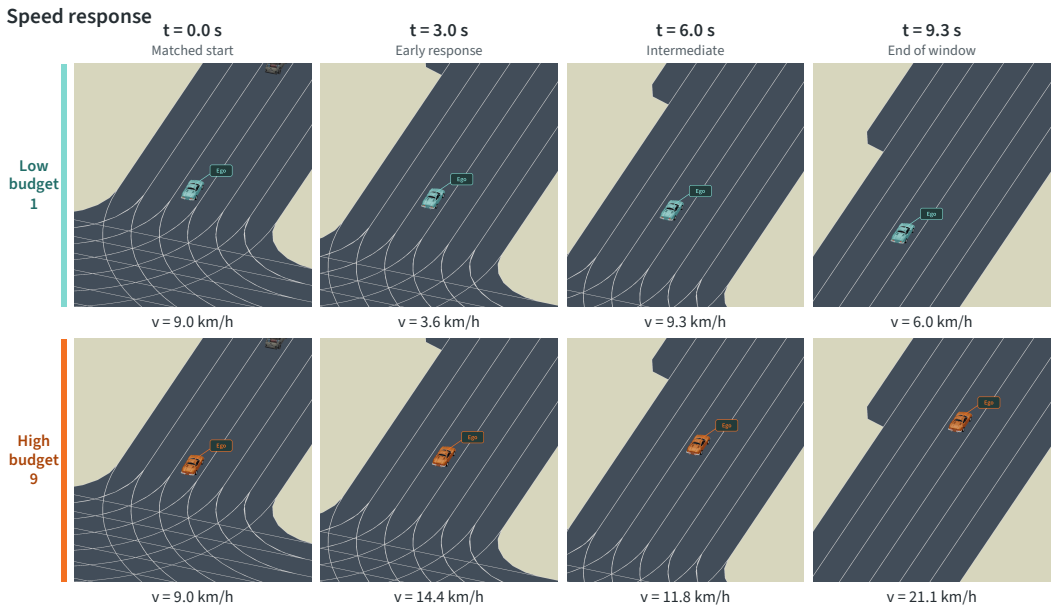


Figure 3: **Recorded speed response.** Reported cost allowances are 1 (top, cyan) and 9 (bottom, orange). Columns show $t = 0.0$, 3.0, 6.0, and 9.3 s. From a matched starting state, the low-budget vehicle initially slows down, while the high-budget vehicle accelerates.

Comparison setup. Within each scenario, we compare two recorded rollouts of the same trained policy, labeled by different initial cost allowances. The rollouts start from matched physical states and environment observations and use matched inference-noise seeds. The low-budget and high-budget ego vehicles are shown in cyan and orange, respectively.

Visualization. Rows correspond to budget settings, and columns show synchronized simulation times; v denotes instantaneous ego speed. Keyframes are rendered from the recorded trajectories using the same setup as the videos, and timestamps exclude the opening and closing video holds. At each timestamp, both rows share the same camera pose and spatial scale. The camera remains fixed in the overtaking scenario; in the other two scenarios, it follows the midpoint of the two ego positions while maintaining a fixed orientation and scale. The corresponding videos are available on our anonymous project page.

H.1 Speed Response

Figure 3 compares initial budgets of 1 and 9. Starting from approximately 9.0 km/h, the low-budget vehicle initially decelerates, while the high-budget vehicle accelerates. At $t = 3.0$ s, their speeds are 3.6 and 14.4 km/h, respectively; by $t = 9.3$ s, they reach 6.0 and 21.1 km/h. The recordings show distinct speed profiles, with the low-allowance rollout traveling more slowly at all displayed post-initial times.

H.2 Overtaking

Figure 4 compares the overtaking behavior in the two recorded allowance settings discussed in the main text. With an initial budget of 1, the ego vehicle decelerates and does not complete a pass during the displayed interval. With an initial budget of 10, the same policy first completes a pass of the original lead vehicle at $t = 5.6$ s. The frames capture the matched starting state, the approach, the first completed pass, and the end of the displayed interval.

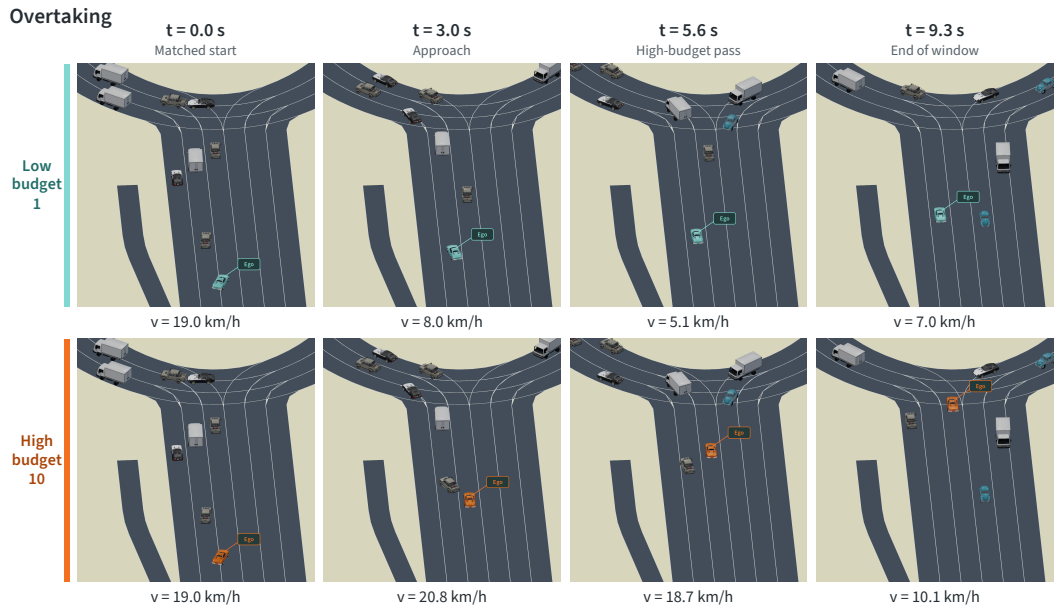


Figure 4: **Recorded overtaking behavior.** Reported cost allowances are 1 (top, cyan) and 10 (bottom, orange). Columns show $t = 0.0$, 3.0, 5.6, and 9.3 s. The high-budget vehicle first passes the original lead vehicle at $t = 5.6$ s; the low-budget vehicle does not complete a pass within the displayed interval.

H.3 Response to a Vehicle Ahead

Figure 5 compares initial budgets of 1 and 9 in a scene with a vehicle ahead. Both rollouts begin at approximately 21.2 km/h. At $t = 3.0$ s, the low-budget vehicle has slowed to 9.4 km/h, while the high-budget vehicle is traveling at 20.9 km/h. By $t = 9.3$ s, their speeds are 4.8 and 12.7 km/h, respectively. Both vehicles reduce their speed over the displayed sequence, with a larger reduction under the lower budget at each displayed post-initial time.

The tracked original lead vehicle remains ahead throughout the displayed interval but changes lanes at $t = 3.4$ s, between the second and third columns. This sequence records two responses to a vehicle ahead; it does not show continuous following in the same lane.

I Detail Experiment Results

I.1 Main Results

Table 2 provides the detailed results of UniQ on all nine MetaDrive tasks. The normalized cost remains below the safety threshold of 1 in every task, with small standard deviations across three random seeds. Normalized reward varies across environments, with stronger performance on the Medium tasks and lower rewards on the more challenging Hard tasks. These results demonstrate consistent constraint satisfaction across different task difficulties and cost-density settings.

I.2 Number of cumulative-cost samples.

We vary the number of cumulative-cost draws per transition, $N_u \in \{8, 16, 24\}$, while keeping the remaining settings fixed (Table 3). All nine configurations remain feasible, with normalized costs ranging from 0.90 to 0.99, showing that the safety behavior is stable across the tested sample counts. Reward also changes only moderately with N_u. For EasySparse, increasing N_u from 8 to 24 improves reward from 0.49 to 0.55, whereas EasyMean favors the smaller setting, with reward decreasing from 0.55 at $N_u = 8$ to 0.50 at $N_u = 24$. EasyDense exhibits a similarly small variation, with rewards between 0.60 and 0.65. These results show no consistent monotonic benefit from

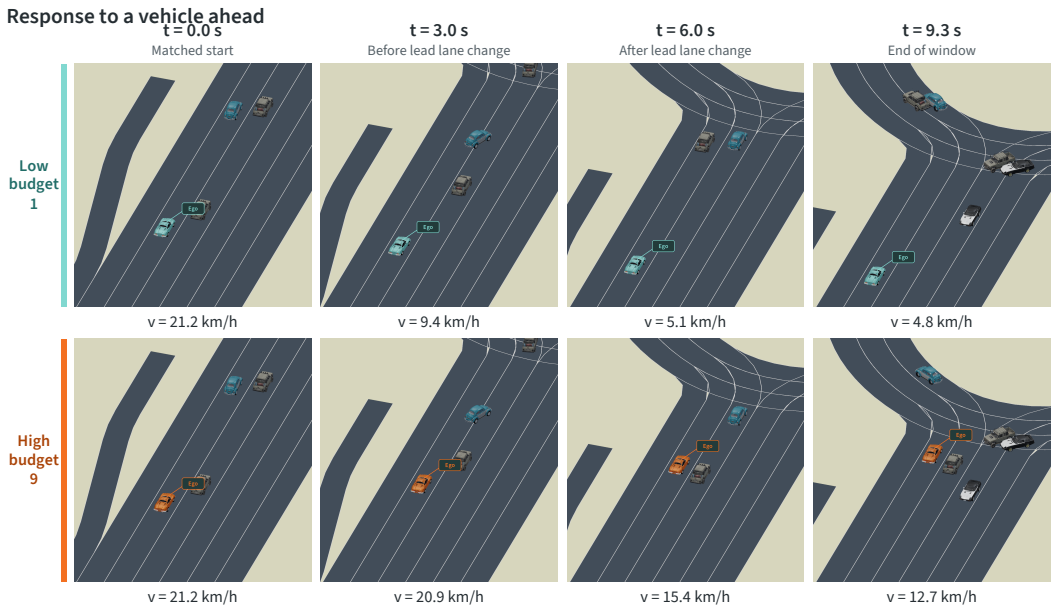


Figure 5: **Recorded response to a vehicle ahead.** Reported cost allowances are 1 (top, cyan) and 9 (bottom, orange). Columns show $t = 0.0$, 3.0, 6.0, and 9.3 s. The low-budget vehicle exhibits a larger speed reduction. The original lead vehicle remains ahead but changes lanes at $t = 3.4$ s, between the second and third columns.

Table 2: Main results of UNIQ on the MetaDrive benchmark. We report the normalized reward and normalized cost over nine tasks, where higher reward is better and lower cost is better. Results are shown as mean $\pm$ standard deviation over 3 random seeds.

Task	Normalized Reward ↑	Normalized Cost ↓
EasySparse	0.54 ± 0.07	0.91 ± 0.03
EasyMean	0.52 ± 0.06	0.90 ± 0.02
EasyDense	0.60 ± 0.03	0.91 ± 0.02
MediumSparse	0.78 ± 0.11	0.93 ± 0.04
MediumMean	0.75 ± 0.03	0.93 ± 0.02
MediumDense	0.83 ± 0.02	0.97 ± 0.01
HardSparse	0.47 ± 0.01	0.94 ± 0.01
HardMean	0.43 ± 0.01	0.96 ± 0.01
HardDense	0.39 ± 0.04	0.91 ± 0.01

increasing the number of cumulative-cost samples. Overall, UNIQ maintains similar reward and safety performance across $N_u = 8$–24, suggesting that its performance is not highly sensitive to this hyperparameter within the tested range.

I.3 CUMULATIVE-COST SAMPLING RATIO.

We vary the fraction of budget samples drawn from $u \in [0, 10]$, denoted by ρ, while drawing the remaining samples from $u > 10$. The total number of samples and the sampling distribution within each region are kept fixed. Table 4 shows that all tested ratios, $\rho \in \{0.25, 0.6, 0.75\}$, remain feasible on the three Easy tasks, with normalized costs ranging from 0.84 to 0.99. Reward is also relatively stable across the tested mixtures, varying by at most 0.07 within each task. On EasySparse, reward increases from 0.51 at $\rho = 0.25$ to 0.57 at $\rho = 0.75$, whereas EasyMean achieves its highest reward of 0.59 at $\rho = 0.25$. EasyDense shows even smaller variation, with rewards between 0.60 and

Table 3: Number of budget samples per training instance. The sampling ratio is kept fixed. Reward and Cost denote normalized reward and normalized cost, respectively; the same notation is used in the following tables.

	$n_u = 8$		$n_u = 16$		$n_u = 24$	
Task	Reward ↑	Cost ↓	Reward ↑	Cost ↓	Reward ↑	Cost ↓
EasySparse	0.49	0.95	0.54	0.91	0.55	0.99
EasyMean	0.55	0.97	0.52	0.90	0.50	0.99
EasyDense	0.65	0.91	0.60	0.91	0.64	0.92

Table 4: Budget sampling ratio. A fraction ρ of the budget samples is drawn from $[0, 10]$, and the remaining fraction from $u > 10$. The total number of budget samples is kept fixed.

	$\rho = 0.25$		$\rho = 0.6$		$\rho = 0.75$	
Task	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓
EasySparse	0.51	0.94	0.54	0.91	0.57	0.97
EasyMean	0.59	0.96	0.52	0.90	0.56	0.99
EasyDense	0.64	0.88	0.60	0.91	0.60	0.84

0.64. Thus, no single sampling ratio consistently dominates across tasks, and changing the mixture primarily induces modest task-dependent variations rather than a systematic improvement. Overall, these results suggest that UNIQ is reasonably robust to the budget sampling ratio within the tested range. This experiment evaluates sensitivity to the resampling mixture only and does not compare resampling against training directly on logged cumulative costs.