STOCHASTIC DECISION HORIZONS FOR SURVIVAL-BASED CONSTRAINED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose *stochastic decision horizons* (SDH), a theoretically grounded framework for survival-based constrained reinforcement learning from per-step violation signals. Rather than treating violations as expenditure of a cumulative budget, SDH shortens the effective decision horizon when violations occur, reducing current reward credit and future bootstrapping. We show that this construction represents a geometric-horizon survival constraint and, for a fixed Lagrange multiplier, remains the same SDH problem up to a constant reward offset that acts as an implicit incentive for survival. Using control as inference, we derive two off-policy and replay-compatible, regularized algorithms with post-violation semantics. Absorbing-state ends the decision process, so only surviving decisions pay policy KL, yielding AS-SAC. Virtual termination stops reward credit while the process continues, retaining ordinary discounted KL and yielding VT-MPO. To connect SDH with cumulative-cost CMDPs, we introduce *violation-depth profiles*. Under exponential-tail single-scale profiles, the SDH statistic determines cumulative cost; departures from this form mark where the connection breaks. Experiments validate both the method and this predicted scope. On the 90-muscle H2190 humanoid in Hyfydy, VT-MPO matches the EWA reference in peak gait realism, with its highest observed gait-match checkpoint at 21M environment steps versus 90M for EWA and markedly more stable training. On Safety Gymnasium, violation-depth profiles predict the regimes in which SDH achieves strong reward-violation trade-offs.

1 INTRODUCTION

Many control problems expose limits at every interaction step: gait quality depends on velocity, ground-reaction force, joint torque, and muscle use; robotic and physical systems likewise operate within local ranges. These signals are not always well described as consumption of a transferable budget. The standard expected discounted cumulative-cost CMDP (6) instead asks whether the sum of costs stays below a budget, which is natural for quantities such as fuel or total exposure. Early termination and continuation already provide practical mechanisms for handling such per-step violation signals (41; 11; 10). But what constrained objective, if any, do these continuation mechanisms actually optimize? And if they solve a constrained problem, why can they sometimes avoid the usual Lagrange-dual optimization needed to enforce the constraint?

We propose *stochastic decision horizons* (SDH) as a framework for survival-based constrained reinforcement learning. A state-action continuation model $\alpha(s, a) \in [0, 1]$ maps one or more local violation signals to a model-defined continuation probability. This induces the transformation $\tilde{r} = \alpha r$ and $\tilde{\gamma} = \gamma\alpha$: a violating transition loses current reward credit and reduces future bootstrapping. Thus "shortening the effective horizon" is an intuition; formally, SDH uses a survival-weighted reward and a state-action-dependent discount (48). We show that this construction exactly represents a survival-gated geometric-horizon chance-constrained problem whose objective and constraint share the same survival process. Because the local violation signals enter through the same continuation model, the resulting problem has a single survival constraint and hence a single scalar Lagrange multiplier η. For a fixed multiplier η, we show that its Lagrangian closes within the SDH Bellman objective through the reward shift $r \mapsto r + \eta$. Since constant rewards are survival-weighted, this shift directly incentivizes survival. We further show that, when a bounded control cost f is written as a reward $r = 1 - f$, the constant 1 is no longer an irrelevant reward offset: under SDH, it generates the survival-probability term that appears in the constraint, even at $\eta = 0$. If this fixed objective already

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.

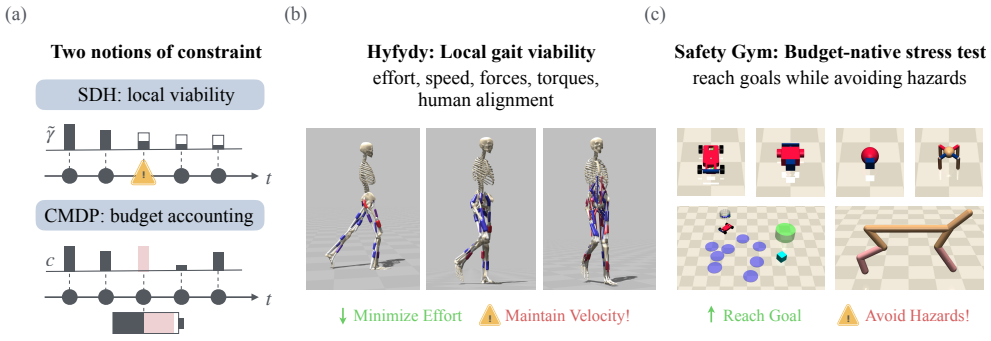


Figure 1: **Stochastic decision horizons vs. cumulative-cost accounting.** Cumulative-cost CMDPs add violations over time. SDH maps per-step violations to continuation probability α and attenuates reward and bootstrapping through $(\tilde{r} := \alpha r, \tilde{\gamma} := \alpha\gamma)$. The right objective depends on the application: Hyfydy local operating-limit signals are naturally modeled through survival, while violation-depth profiles expose when budget-native Safety Gymnasium tasks require additive-cost accounting.

satisfies the survival constraint, no dual optimization is needed; otherwise, the constrained problem can be optimized with a primal-dual method.

To realize this survival objective in high-dimensional continuous control, we derive replay-compatible off-policy algorithms. This requires specifying the post-failure semantics of policy regularization: should survival weighting apply to policy KL as well as reward, or should policy KL remain on the ordinary discounted horizon? Control as inference (CaI) (44; 30; 24) provides the machinery to answer this question. We extend CaI to SDH and show that the resulting inference problem has finite evidence and a well-defined ELBO, and formalize two coherent post-violation semantics. Under *absorbing-state* (AS) semantics, continuation failure ends the controlled decision process, so reward and policy KL are weighted by the same survival process; this yields a variable-discount soft Bellman objective and motivates AS-SAC, built on Soft Actor-Critic (21). Under *virtual termination* (VT), continuation failure stops reward credit but not the decision process, so reward remains survival-weighted while policy KL stays on the ordinary discounted horizon; this yields VT-MPO, which combines the SDH critic with MPO's KL-constrained policy improvement (1).

We next ask whether this local survival formulation can recover cumulative-cost behavior at all, despite never tracking accumulated cost. Exponential continuation makes SDH depend on cumulative violation depth, but through an exponential transform rather than the additive statistic used by cumulative-cost CMDPs. We therefore introduce the *violation-depth profile*: cumulative cost is its unweighted area, whereas SDH reads one Laplace-transform statistic. In general, one does not determine the other; under an exponential-tail single-scale profile, however, we show that the SDH statistic determines cumulative cost. Departures from this single-scale form break this correspondence.

We validate both regimes empirically. Our main application is high-fidelity musculoskeletal control, where velocity, ground-reaction force, torque, and muscle-use violations indicate deterioration of the current gait rather than expenditure of a transferable budget. We introduce a survival-constrained RL formulation of locomotion under such heterogeneous biomechanical operating limits. In Hyfydy (18), locomotion minimizes a bounded *effort* cost f through the reward $r = 1 - f$, so the same survival-probability term that appears in the constraint is already present in the fixed $\eta = 0$ objective, allowing training without dual optimization. On the 90-muscle H2190, VT-MPO matches the EWA reference in gait realism (34) with markedly more stable training; its highest observed gait-match checkpoint occurs at 21M environment steps versus 90M for EWA. On Safety Gymnasium (32), violation-depth profiles predict where SDH excels and where cumulative constraints are required.

Together, these results carry the motivating class of real-world control problems from modeling to solution: SDH provides a survival-constrained formulation and yields two principled off-policy algorithms; we then formulate high-dimensional biomechanical control as an SDH problem and solve it, while the same theory predicts when cumulative-cost control is required.

Related work. Early termination and continuation have been used to handle per-step violation signals (41; 11; 10), while chance-constrained RL studies geometrically weighted prefix-safety probabilities (47). We do not claim continuation or geometric weighting alone as novel. SDH instead introduces a survival-gated geometric-horizon constrained problem in which a survival process jointly gates the task objective and a single survival constraint. This matched structure yields one scalar Lagrange multiplier, fixed-η closure with η acting as a survival incentive, a well-defined CaI ELBO with principled AS/VT off-policy algorithms, and the violation-depth profile scope analysis.

Standard cumulative-cost methods include CPO and trust-region methods (2), primal-dual schemes (32; 40), penalty/interior-point methods (25; 52; 51; 14; 45), and off-policy constrained critics (49; 42; 33; 23; 54). State augmentation can instead carry residual budget in the state (39), capturing residual-budget semantics that state-action-only SDH does not represent.

Hard state-wise safety is a stronger contract (53). Shielding can enforce temporal specifications by intervening on unsafe actions (5); reachability, barrier-function, and projection methods provide related action-level mechanisms (38; 13; 12). Such methods may alter action selection at deployment; SDH instead trains the deployed actor under multiplicative survival weighting and provides no hard action-level guarantee.

Risk-sensitive RL changes how return distributions are evaluated (7), while robust MDPs optimize against transition-model uncertainty (28); SDH instead transforms reward and discount using observed violation signals. This differs from generic reward shaping (27), because continuation changes both reward credit and future discounting. Our regularized algorithms build on CaI (44; 30; 31; 24; 20), SAC (21), and MPO (1). Hyfydy provides musculoskeletal simulation (18); EWA (34) is the domain-specific reference for emergent natural gait.

2 Stochastic Decision Horizons

We consider an infinite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \nu)$. Here $P(\cdot \mid s, a)$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is a bounded non-negative reward function, $\gamma \in [0, 1)$ is the discount, and ν is the initial-state distribution. A policy π induces trajectories $\tau = (s_0, a_0, s_1, \ldots)$. The environment also exposes a non-negative per-step violation magnitude $c(s, a) \geq 0$, with $c(s, a) = 0$ denoting no local violation. We state the finite-state/action case for notation; the same Bellman and trajectory expressions extend to continuous spaces by replacing sums with integrals under the usual measurability and density assumptions.

2.1 Geometric-horizon survival constraint and Bellman equation

Discounted RL admits a geometric-horizon interpretation (44). Let T_γ be independent of the MDP trajectory and continuation variables, with $\Pr(T_\gamma = t) = (1 - \gamma)\gamma^t$ for $t \in \{0, 1, \ldots\}$. We add a model-defined survival process whose continuation probability is constructed from the per-step violation signal. The same survival event gates reward and constraint; this matching yields the SDH Bellman representation and fixed-η Lagrangian closure. App. F gives details, and Figure 16 summarizes its relation to pointwise survival chance constraints and cumulative-cost CMDPs.

Definition 2.1 (Survival-gated geometric-horizon problem)**.** Consider

$$C_t \mid s_t, a_t \sim \text{Bernoulli}(\alpha(s_t, a_t)), \qquad B_t := \prod_{k=0}^{t} C_k, \tag{1}$$

where $\alpha : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a state-action continuation model. For a given survival-probability level $\rho \in (0, 1)$, define

$$\max_{\pi} \ \mathbb{E}_\pi[B_{T_\gamma} r(s_{T_\gamma}, a_{T_\gamma})] \quad \text{subject to} \quad \Pr_\pi(B_{T_\gamma} = 1) \geq \rho. \tag{2}$$

By independence of T_γ, the survival probability is $\Pr_\pi(B_{T_\gamma} = 1) = (1 - \gamma) \sum_{t \geq 0} \gamma^t \Pr_\pi(B_t = 1)$, so the constraint represents geometrically averaged survival over the random horizon, rather than hard pointwise safety. The continuation process terminates with probability $1 - \alpha(s, a)$; this model-defined event should be distinguished from physical system failure.

Marginalizing the continuation variables gives the survival-weighted reward $\tilde{r} = \alpha r$ and state-action-dependent discount $\tilde{\gamma} = \gamma\alpha$. The survival-weighted task critic is the unique fixed point

$$Q^{\pi}_{\mathrm{surv}}(s,a) = \tilde{r}(s,a) + \tilde{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}[Q^{\pi}_{\mathrm{surv}}(s',a')], \tag{3}$$

since $0 \le \tilde{\gamma}(s,a) \le \gamma < 1$ makes the policy-evaluation operator a γ-contraction for bounded rewards. This defines the survival-transformed MDP. For any bounded function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, define the unnormalized survival-weighted occupancy functional

$$\mu^{\pi}_{\alpha}(f) := \mathbb{E}_{\pi}\Big[\sum_{t\ge 0}\gamma^t\Big(\prod_{k=0}^{t}\alpha(s_k,a_k)\Big)f(s_t,a_t)\Big]. \tag{4}$$

Theorem 2.2 (Geometric-horizon representation and survival Lagrangian)**.** *For every bounded function f, the identity $(1-\gamma)\mu^{\pi}_{\alpha}(f) = \mathbb{E}_{\pi}[B_{T_\gamma} f(s_{T_\gamma}, a_{T_\gamma})]$ holds. Consequently, the Lagrangian of* (2) *is*

$$\mathcal{L}(\pi,\eta) = (1-\gamma)\mu^{\pi}_{\alpha}(r+\eta) - \eta\rho, \qquad \eta \ge 0. \tag{5}$$

For fixed η, the Lagrangian closes within the SDH Bellman objective: maximizing $\mathcal{L}$ is equivalent to maximizing an SDH return with reward shifted by η.

The proof is in App. F.3. This closure is made explicit by the identity

$$(1-\gamma)\mu^{\pi}_{\alpha}(r+\kappa) = (1-\gamma)\mu^{\pi}_{\alpha}(r) + \kappa \Pr_{\pi}(B_{T_\gamma}=1), \quad \forall\kappa \ge 0.$$

Thus a constant reward shift adds exactly a survival-probability term. If a bounded control cost f is encoded as a reward $r = 1 - f$, then

$$(1-\gamma)\mu^{\pi}_{\alpha}(1-f) = \Pr_{\pi}(B_{T_\gamma}=1) - (1-\gamma)\mu^{\pi}_{\alpha}(f),$$

therefore the unit reward already provides a survival incentive even at $\eta = 0$. The multiplier η strengthens this incentive through the transformed reward $\alpha(s,a)(r(s,a)+\eta)$; a fixed η selects one Lagrangian operating point, while a given level ρ can be targeted by primal-dual optimization.

2.2 Relation to cumulative-cost CMDPs

The standard expected discounted cumulative-cost CMDP (6) is

$$\max_{\pi} J_r(\pi) := \mathbb{E}_{\pi}\Big[\sum_{t\ge 0}\gamma^t r(s_t,a_t)\Big] \quad \text{subject to} \quad J_c(\pi) := \mathbb{E}_{\pi}\Big[\sum_{t\ge 0}\gamma^t c(s_t,a_t)\Big] \le d.$$

SDH is not a general solver for this additive-cost constraint. Separately, a state-action continuation model cannot reproduce exact pathwise residual-budget semantics without accumulated cost in the state (App. F.5). Figure 16 illustrates the broader comparison developed below.

Let $G_t := \sum_{k=0}^{t} c(s_k,a_k)$ and define the *violation-depth profile* $\Omega_{\pi}(b) := \mathbb{E}_{\pi}[\sum_{t\ge 0}\gamma^t \mathbf{1}\{G_t \ge b\}]$.

Proposition 2.3 (Additive cost and exponential SDH)**.** *At the geometric horizon T_γ, the identity $(1-\gamma)\Omega_{\pi}(b) = \Pr_{\pi}(G_{T_\gamma} \ge b)$ holds. Under exponential continuation $\alpha_\lambda(s,a) = \exp\{-\lambda c(s,a)\}$,*

$$\bar{S}_{\pi}(\lambda) := (1-\gamma)\mu^{\pi}_{\alpha_\lambda}(1) = \Pr_{\pi}(B_{T_\gamma}=1) = \mathbb{E}_{\pi}[e^{-\lambda G_{T_\gamma}}],$$

and

$$J_c(\pi) = \mathbb{E}_{\pi}[G_{T_\gamma}] = (1-\gamma)\int_0^{\infty}\Omega_{\pi}(b)\,db, \qquad \bar{S}_{\pi}(\lambda) = 1 - \lambda(1-\gamma)\int_0^{\infty} e^{-\lambda b}\Omega_{\pi}(b)\,db.$$

For fixed $\lambda_0 > 0$, one exponential statistic does not upper-bound J_c in general: even in a two-state, two-action MDP, $\bar{S}_{\pi}(\lambda_0)$ can remain fixed while $J_c(\pi)$ becomes arbitrarily large (App. F.5).

Thus J_c is the mean of G_{T_γ}, while exponential SDH observes one Laplace-transform statistic of the same distribution. This distinction disappears in the exponential-tail single-scale regime: if $\Omega_{\pi}(b) = (1-\gamma)^{-1}\exp\{-b/\tau_{\pi}\}$, with the required prefactor $(1-\gamma)^{-1}$, then $J_c(\pi) = \tau_{\pi}$ and $\bar{S}_{\pi}(\lambda) = (1+\lambda\tau_{\pi})^{-1}$. Consequently, for any fixed $\lambda > 0$, this gives $J_c(\pi) \le d \iff \bar{S}_{\pi}(\lambda) \ge (1+\lambda d)^{-1}$ (App. F.6). Departures from this single-scale form break this equivalence. Section 5 tests this predicted boundary on Safety Gymnasium (32).

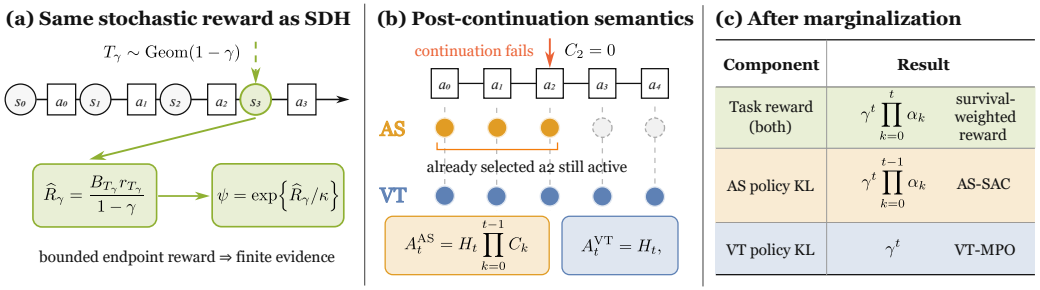


Figure 2: **Control as inference with stochastic decision horizons.** The CaI construction applies a single optimality potential to the SDH endpoint reward $B_{T_\gamma} r_{T_\gamma}/(1-\gamma)$. AS and VT share this reward but differ in policy-KL horizon: AS removes future policy decisions after continuation failure, while VT retains ordinary discounted KL.

3 Control as Inference Formulation

SDH determines which rewards count under the survival process, but regularized off-policy control leaves a second semantic question: which policy decisions should incur policy KL after continuation failure? Control as inference (CaI) (44; 30; 24) makes this choice principled, because the regularizer must arise as the trajectory KL between the corresponding reference- and policy-induced trajectory laws, rather than from an arbitrary weighting of per-step KL terms. By Theorem 2.2, fixing η changes the survival-constrained objective only through a reward shift $r \mapsto r + \eta$ plus a policy-independent constant. Hence, the analysis below applies directly to every fixed-η Lagrangian operating point.

In CaI, the key modeling choice is which trajectory reward defines the optimality potential. For SDH, we use exactly the random-time reward of the survival-constrained problem. Let T_γ be the geometric evaluation time and B_t the survival process, and define $\widehat{R}_\gamma = (1-\gamma)^{-1} B_{T_\gamma} r(s_{T_\gamma}, a_{T_\gamma})$. The unnormalized potential is $\psi = \exp(\widehat{R}_\gamma/\kappa)$ under reference law p. By Theorem 2.2, $\mathbb{E}_\pi[\widehat{R}_\gamma] = \mu_\alpha^\pi(r)$. Crucially, $\widehat{R}_\gamma$ contains one endpoint reward rather than an infinite trajectory sum; hence $|r|_\infty \le R_{\max}$ gives $\psi \le \exp(R_{\max}/[\kappa(1-\gamma)])$ and thus the partition function $Z := \mathbb{E}_p[\psi] < \infty$. Therefore, for every $\kappa > 0$, the construction matches the SDH objective and yields a well-defined CaI ELBO.

What remains is to specify which policy decisions belong to the trajectory law after continuation failure. We consider two natural semantics. Under absorbing-state (AS) semantics, continuation failure sends the process to a policy-trivial absorbing state, so later actions disappear from the trajectory law. Under virtual termination (VT), continuation failure terminates reward credit only; the simulator and controller continue, so later action likelihoods remain active.

Let p_σ and q_σ be the reference- and policy-induced trajectory laws for $\sigma \in \{\mathrm{AS}, \mathrm{VT}\}$, sharing the environment, continuation variables, and geometric horizon, with policies π_0 and π, respectively. Their policy KL is (see Lemma G.5)

$$D_{\mathrm{KL}}(q_\sigma \| p_\sigma) = \mathbb{E}_{q_\sigma}\Big[\sum_{t\ge 0} A_t^\sigma \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}\Big], \qquad A_t^\sigma = \begin{cases} H_t \prod_{k=0}^{t-1} C_k, & \sigma = \mathrm{AS}, \\ H_t, & \sigma = \mathrm{VT}. \end{cases} \tag{6}$$

The one-step offset follows from the event ordering: a_t is selected before C_t is realized. Thus under AS the current action still incurs KL when $C_t = 0$, whereas its reward credit and all subsequent decisions are gated. Write $\mathcal{J}_\sigma$ for the κ-scaled CaI ELBO after marginalizing C, D.

Theorem 3.1 (AS and VT CaI objectives)**.** *Let $\alpha_t := \alpha(s_t, a_t)$, $\tilde{r}_t := \alpha_t r(s_t, a_t)$, $\tilde{\gamma}_t := \gamma \alpha_t$, and $u_t := \prod_{k=0}^{t-1} \tilde{\gamma}_k$ with $u_0 = 1$. Then*

$$\mathcal{J}_\sigma(\pi, \pi_0) = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\ge 0} u_t \tilde{r}_t - \kappa \sum_{t\ge 0} v_t^\sigma \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}\Big], \quad v_t^\sigma = \begin{cases} u_t, & \sigma = \mathrm{AS}, \\ \gamma^t, & \sigma = \mathrm{VT}. \end{cases} \tag{7}$$

For a fixed Lagrange multiplier η of the survival constraint, the same CaI objective applies with r replaced by $r + \eta$, up to the policy-independent constant $-\eta\rho/(1-\gamma)$.

The endpoint ELBO is established in App. G.4, and its reduction to the AS and VT objectives is shown in Theorem G.9. AS and VT therefore share the same survival-weighted task objective but induce different ELBO-valid policy-KL horizons. Under AS, reward and policy KL share the same state-action-dependent discount, yielding the KL-regularized Bellman recursion used by AS-SAC. Under VT, policy KL remains ordinarily discounted, so the survival-weighted task critic is paired with KL-constrained policy improvement, yielding VT-MPO.

4 ALGORITHMS

AS and VT share the same survival-weighted task objective but differ in the policy-KL horizon induced by their post-continuation semantics. Replay buffer retains the raw rewards and per-step violation signals needed to construct each critic target. When constructing a target, we compute the model-defined continuation probability $\alpha_\lambda(s,a)$ (in our experiments, exponential $\alpha_\lambda(s,a) = \exp\{-\lambda \sum_i c_i(s,a)\}$), together with $\tilde{r}_\eta(s,a) = \alpha_\lambda(s,a)(r(s,a)+\eta)$ and $\tilde{\gamma}(s,a) = (1-d)\gamma\alpha_\lambda(s,a)$, using the current λ and, during primal-dual optimization, the current η. During λ-annealing, this uses the current value along the ramp. Thus replayed experience is evaluated under the current continuation model α_λ.

AS-SAC. Under AS, reward and policy KL share the same state-action-dependent discount. The KL-regularized Bellman recursion is

$$Q^\pi_{\mathrm{AS}}(s,a) = \tilde{r}_\eta(s,a) + \tilde{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi_{\mathrm{AS}}(s')], \tag{8}$$

$$V^\pi_{\mathrm{AS}}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi_{\mathrm{AS}}(s,a) - \kappa \log \frac{\pi(a \mid s)}{\pi_0(a \mid s)}\right]. \tag{9}$$

For a normalized uniform prior, this yields maximum-entropy AS-SAC. We show that its survival-weighted living cost is policy dependent, unlike the corresponding constant in ordinary SAC, and give a counterexample where dropping it changes the optimizer. For adaptive κ, we use standard SAC-style temperature adaptation at policy improvement: the current regularized AS critic is fixed during the actor/temperature update, while subsequent critic updates use the updated temperature (App. H.2.1).

VT-MPO. Virtual termination is objective-level reward termination without environment reset. Under VT, the task critic remains survival-weighted while policy KL stays on the ordinary discounted horizon. The critic obeys

$$Q^\pi_{\mathrm{surv}}(s,a) = \tilde{r}_\eta(s,a) + \tilde{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}[Q^\pi_{\mathrm{surv}}(s',a')]. \tag{10}$$

We therefore pair Q^π_{surv} with standard MPO relative-entropy-constrained policy improvement:

$$q^\star(a \mid s) \propto \pi_0(a \mid s)\exp\left(\beta^{-1}Q^\pi_{\mathrm{surv}}(s,a)\right), \tag{11}$$

where β is the MPO temperature for the policy-improvement KL budget; η is reserved for the survival constraint. Thus VT-MPO modifies MPO only through the survival-weighted task critic; retaining ordinary actor-side KL constraints is the consequence of the VT CaI semantics, not an approximation.

Primal-dual optimization of the survival constraint. The survival probability is

$$p_{\mathrm{surv}}(\pi) = (1-\gamma)\mu^\pi_\alpha(1) = \Pr_\pi(B_{T_\gamma} = 1). \tag{12}$$

When no specific constraint level is imposed, fixed η selects a Lagrangian operating point without dual optimization; notably, for $r = 1 - f$, the constant 1 already generates the survival-probability term, so even $\eta = 0$ can provide an effective survival incentive. For a given survival-constraint level ρ, an auxiliary survival-mass critic ($r \equiv 1$) can estimate $p_{\mathrm{surv}}(\pi)$ for the dual update. Since the survival constraint is scalar, this requires only one auxiliary survival-mass critic and one scalar dual variable η, regardless of the number of violation signals entering α. For the exponential continuation, λ controls how strongly violations attenuate continuation. Annealing λ therefore changes the continuation model, and hence the SDH objective, making it an optimization curriculum rather than dual optimization of a fixed survival constraint.

5 EXPERIMENTS

The experiments test two linked hypotheses. **H1:** a challenging musculoskeletal-control problem can be formulated as a fixed survival objective built from local operating-limit signals, yielding realistic gait and strong empirical limit adherence. **H2:** violation-depth profiles predict when this survival criterion agrees with, or can diverge from, the standard cumulative-cost CMDP criterion. Hyfydy, our motivating local-operating-regime setting, tests H1; Safety Gymnasium provides a budget-native scope test for H2.

5.1 HYFYDY: MUSCULOSKELETAL OPERATING LIMITS

We evaluate VT-MPO on three Scone-Hyfydy humanoids (17; 18): H0918 (2D, 9 DOFs, 18 muscles), H1622 (3D, 16 DOFs, 22 muscles), and H2190 (3D, 21 DOFs, 90 muscles). The domain reference is Effort Weight Adaptation (EWA) (34), which dynamically reweights effort penalties to resolve the effort-velocity trade-off. We use VT-MPO because constraint events do not reset the physical simulator, its MPO base reduces optimizer confounding with the EWA comparison, and the diagnostic in App. D.4 is more robust across H2190 seeds. An unconstrained low-effort SAC/MPO control is not informative here: removing the gait operating limits makes trivial non-walking behavior optimal, rather than testing how to encode the same gait task.

Formulation and protocol. SDH uses an effort-based task reward and maps target velocity, ground-reaction force (GRF), joint-torque limits, and muscle-activation sparsity into $\alpha = \exp\{-\lambda(c_1 + c_2 + \cdots)\}$; the GRF limit is 1.3 body weights and the remaining model-specific limits are in App. D.2. These are operating-limit violation signals: exceeding a limit indicates deterioration of the current gait, not consumption of a transferable resource budget. For fixed $\eta = 0$, we linearly ramp λ from 0 to 0.2 over the first 5M steps and hold it fixed thereafter. A separate H1622 survival-only diagnostic finds that driving policy improvement with the survival-critic Q_1 alone does not increase Monte-Carlo survival while reducing task return; see App. D.3.1. On the H1622 velocity-only diagnostic, removing this ramp or fixing $\lambda \in \{0.05, \ldots, 1.0\}$ mainly affects early training, while the overall learning trend and limiting behavior remain similar; see App. D.4.1 (Figures 12 and 13). Thresholds were calibrated per model before training and then held fixed, while the reward and continuation schedule were specified in advance; thus, after domain calibration, training proceeds with a fixed protocol, though the method is not tuning-free. Runs use 100M steps and three seeds; every 200k steps, each seed's gait-match is averaged over 10 deterministic rollouts. Following EWA (34), for each method we select the step with the highest across-seed mean and report mean $\pm$ std across seeds at that shared step. H0918/H1622 use 2×256 MLPs and H2190 uses a 2×1024 MLP; MPO uses 20 action samples per state and DEP exploration (35). Approximate runtimes on an NVIDIA RTX 2080 are 21/25/36 hours for H0918/H1622/H2190, respectively (full settings in App. D.2).

Training behavior. Figure 3**a**) shows that VT-MPO reaches stable walking within about 5-10M steps on all three models: episode length approaches the maximum, velocity stays near the target range, and knee over-extension is driven near zero. EWA shows larger oscillations as its adaptive reward weights change during training, and many low-effort checkpoints fail the episode-length stability requirement. Unlike EWA, SDH does not adapt reward weights during training, avoiding this source of target non-stationarity; the observed training curves are correspondingly smoother. Here, $r = 1 - \text{effort}$ already provides a survival incentive at $\eta = 0$, so no Lagrange-dual update is required. Detailed per-limit curves are in App. D.3.

Operating-limit adherence and gait realism. Gait realism is measured by gait-match percentage (17), the fraction of the gait cycle within one standard deviation of experimental human kinematics; neither method optimizes this metric. Best-observed values under this protocol are comparable (Table 4): 75.9% vs. 75.8% on H0918, 70.1% vs. 69.3% on H1622, and 62.9% vs. 64.6% on H2190. Figure 3**b**) shows similar joint-angle and GRF profiles, panel **c**) shows the H2190 gait-match trajectories and selected checkpoints, and panel **d**) compares selected-policy velocity, GRF, and lumbar-torque rollouts against the prescribed operating limits. On H2190, VT-MPO reaches its highest observed gait match at 21M environment steps versus 90M for EWA, giving an earlier best-observed checkpoint under the common evaluation protocol. Model limitations such as missing toes/soft tissues (15; 8) bound attainable realism. In the full problem formulation, fixed-κ AS-SAC

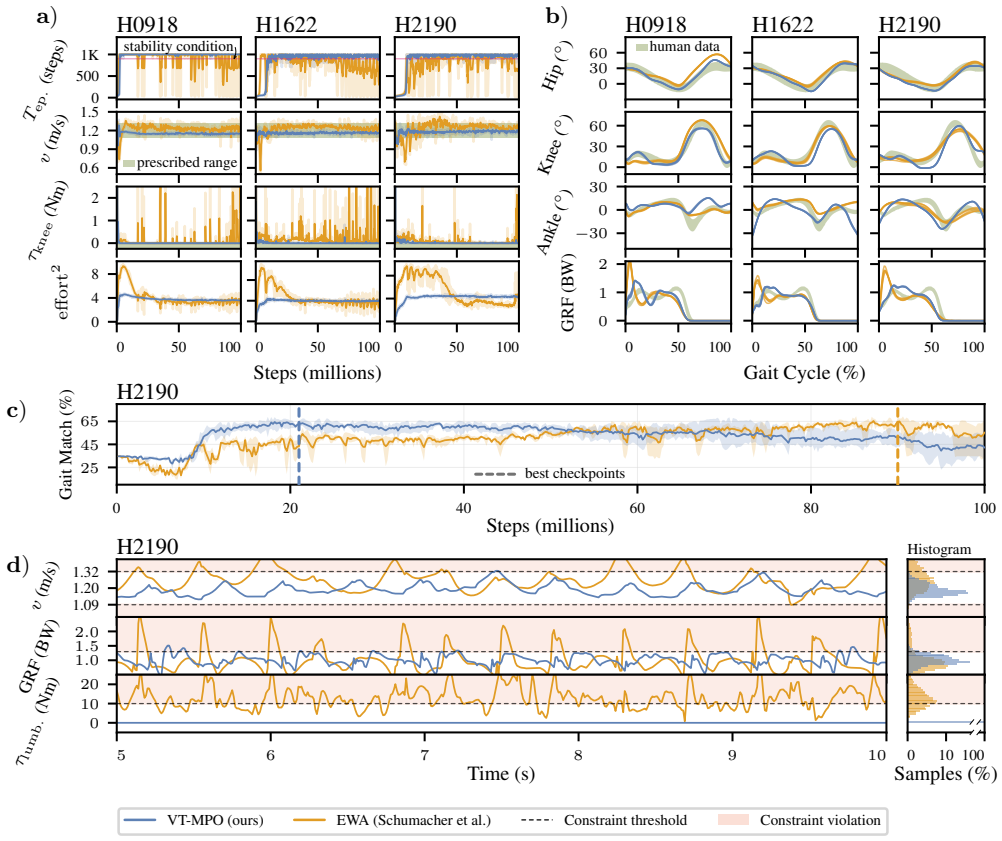


Figure 3: **HyFyDy natural human locomotion.** **a**) Training dynamics for VT-MPO and EWA across three humanoids; VT-MPO reaches stable walking with low knee-limit violation. **b**) Gait-cycle kinematics and GRF at the best-observed gait-match checkpoints. **c**) H2190 gait-match over training; dashed lines mark the selected checkpoints (21M VT-MPO, 90M EWA). **d**) H2190 selected-policy velocity, GRF, and lumbar-torque traces and histograms relative to operating limits.

confirms that it is a viable but more aggressive alternative: successful runs can attain higher human-match values, reaching nearly 90% on H0918, but with greater seed sensitivity (App. D.3); separately, the velocity-only diagnostic includes one failed H2190 seed among three runs (App. D.4.1).

5.2 Safety Gymnasium: violation-depth profiles as a scope test

We evaluate eight Safety Gymnasium tasks (22) with standardized cost signals. AS-SAC and VT-MPO are compared with unconstrained MPO, on-policy CPO and C-TRPO, and off-policy WCSAC and MPO-PID. The benchmark is designed around expected cumulative-cost budgets, whereas SDH uses per-step signals under a survival constraint. We map the native budget d to the per-step threshold $d(1-\gamma)$ used by SDH. Safety Gymnasium also has signed rewards, outside the non-negative-reward assumptions of Sec. 2; we therefore use a fixed positive reward shift η in SDH updates while reporting native rewards. Hyperparameters are in App. E. AS-SAC uses adaptive κ in all Safety Gymnasium results (App. H.2.1). For interpretation, we first assess whether each method satisfies the native cost limit, then compare reward among feasible methods. This benchmark therefore tests this scope boundary between SDH survival and native additive-cost accounting.

Favorable profile geometry. Figure 4 (left) shows that VT-MPO and AS-SAC can substantially reduce violations relative to unconstrained MPO while retaining competitive reward across the benchmark. The full profile sweep across Safety Gymnasium environments (App. E.3) finds the closest single-scale correspondence for `Ant` and `HalfCheetah`. VT-MPO is generally more conservative, consistent with its KL-constrained policy improvement, while maximum-entropy AS-SAC can attain higher reward but is more seed-sensitive.

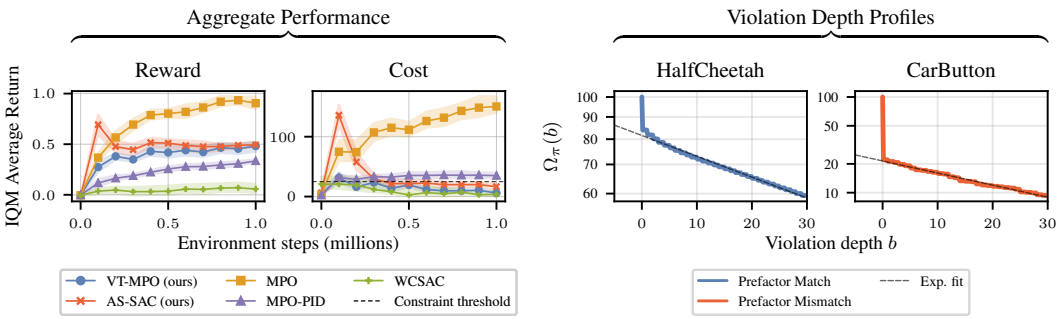


Figure 4: **Safety Gymnasium: aggregate performance and violation-depth profiles.** *Left:* IQM reward and cost across environments and seeds (4). VT-MPO achieves the most favorable aggregate return-violation trade-off; both SDH methods substantially reduce cost relative to unconstrained MPO while retaining competitive reward. *Right:* Violation-depth profiles $\Omega_\pi(b)$ for converged unconstrained SAC policies, with native cost limit $d = 25$. `HalfCheetah` approximately matches the single-scale correspondence, whereas `CarButton` is also nearly exponential but has a strong prefactor mismatch; the environment-wide profile sweep is in App. E.3.

Scope boundary. `CarButton` provides the clearest mismatch: despite an exponential-shaped profile, its prefactor violates the single-scale correspondence, and the fixed-λ AS-SAC ablation either misses the native cost limit or collapses reward (Table 13). `PointGoal` is inconclusive under the profile diagnostic but mixed empirically: neither SDH method is feasible at 1M steps, while at 3M steps AS-SAC becomes feasible and reaches $R = 18.1 \pm 3.3$ with $C = 23.3 \pm 11.4$ (Table 14). Thus, failed single-scale correspondence limits additive-cost recovery without implying SDH failure.

6 Limitations

SDH does not provide hard per-action or state-wise safety certification, a generic additive-cost budget guarantee, or tuning-free continuation design. A low-α action is discouraged through reduced reward credit and future bootstrapping, but it is not prohibited. Thus, when certified exclusion of unsafe actions or temporal guarantees are required, SDH should be combined with or replaced by complementary mechanisms such as shielding, reachability, barrier, or projection methods. We do not benchmark hard pointwise safety methods, which target a stronger safety contract than SDH.

7 Conclusion

SDH formulates survival-based constrained RL from heterogeneous per-step violations through a state-action continuation model. Its geometric-horizon formulation couples the survival-weighted objective to a single survival-probability constraint. For fixed multiplier η, the Lagrangian closes within the same SDH objective through the reward shift $r \mapsto r + \eta$, making η a direct survival incentive. CaI resolves the post-violation semantics of policy regularization from first principles, enabling two new replay-compatible, off-policy algorithms: AS-SAC and VT-MPO.

Violation-depth profiles make the scope relative to expected cumulative-cost CMDPs explicit: under the stated exponential-tail single-scale model, the SDH statistic determines additive cost, whereas profiles that depart from this form generally do not. We then reformulate the motivating high-dimensional biomechanical-control problem in Hyfydy as an SDH survival objective, where heterogeneous operating-limit violations indicate gait deterioration rather than budget expenditure. In Hyfydy, the reward $r = 1 - \text{effort}$ already favors survival through its unit offset, allowing primal-only training with the fixed $\eta = 0$ objective. VT-MPO matches EWA in peak gait realism, trains markedly more stably, and reaches its highest observed H2190 gait-match checkpoint at 21M versus 90M steps. Safety Gymnasium exhibits where SDH excels and where cumulative constraints are required.

Future work includes learned continuation models, full primal-dual validation across survival-constraint levels, offline/pessimistic variants, and evaluation in safety-sensitive applications with complementary certified mechanisms.

AI USE STATEMENT

Generative AI tools were used interactively to assist with mathematical derivations and proofs following approaches developed by the authors, primarily to accelerate intermediate derivations and facilitate checking of arguments. The core conceptual formulation, theoretical insights, and hypotheses were developed by the authors; generative AI was not used to design the experiments, select baselines, interpret the experimental results, or implement the core methods. Generative AI was additionally used for code review and verification, plotting scripts and figure preparation, literature search and summarization, and drafting and editing portions of the manuscript. All AI-assisted mathematical arguments, code, figures, references, and text were reviewed and verified by the authors, who take responsibility for the final content of the work.

ETHICS STATEMENT

This work contributes a technical reinforcement learning objective and accompanying algorithms for simulated constrained control. Its primary expected impact is within machine learning research, safe RL benchmarking, and musculoskeletal control. Deployment in real physical systems would require additional validation, conservative safety mechanisms, and domain-specific oversight beyond the guarantees provided here.

REPRODUCIBILITY STATEMENT

We provide detailed material to reproduce both the theoretical and empirical results. The theoretical appendices contain complete derivations and proofs for the survival-constrained formulation, its relation to cumulative-cost CMDPs, and the control-as-inference objectives underlying AS-SAC and VT-MPO. Appendices A-C give implementation details and implementation-faithful pseudocode for both algorithms, including continuation construction, termination handling, critic targets, replay, and policy updates. The experimental appendices specify the Hyfydy and Safety Gymnasium environments, constraint definitions and thresholds, evaluation metrics, network architectures, hyperparameters, training and evaluation protocols, ablations, and computational resources. We additionally report the number of random seeds and wall-clock compute used for the experiments. The Hyfydy implementation is built within the DEP-RL codebase, allowing VT-MPO and the EWA reference to share simulator settings, preprocessing, action bounds, and exploration machinery.

REFERENCES

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*. OpenReview.net, 2018.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR, 2017.

[3] Marko Ackermann and Antonie J. Van Den Bogert. Optimality principles for model-based prediction of human gait. 43(6):1055–1060.

[4] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2669–2678, 2018.

[6] Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

[7] Nicole Bäuerle and Anna Jaśkiewicz. Markov decision processes with risk-sensitive criteria: An overview. *Mathematical Methods of Operations Research*, 99:141–178, 2024.

[8] Alexandra Buchmann, Simon Wenzler, Lauren Welte, and Daniel Renjewski. The effect of including a mobile arch, toe joint, and joint coupling on predictive neuromuscular simulations of human walking. *Scientific Reports*, 14(1):14879, June 2024.

[9] Elsa K. Bunz, Daniel F. B. Haeufle, Syn Schmitt, and Thomas Geijtenbeek. A spinal network of proprioceptive reflexes can produce a variety of bipedal gaits. 9(1):36.

[10] Elliot Chane-Sane, Joseph Amigo, T. Flayols, Ludovic Righetti, and Nicolas Mansard. Soloparkour: Constrained reinforcement learning for visual locomotion from privileged experience. *Conference on Robot Learning*, 2024.

[11] Elliot Chane-Sane, Pierre-Alexandre Leziart, Thomas Flayols, Olivier Stasse, Philippe Souères, and Nicolas Mansard. Cat: Constraints as terminations for legged locomotion reinforcement learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13303–13310. IEEE, 2024.

[12] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[13] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[14] Sumanta Dey, Pallab Dasgupta, and Soumyajit Dey. P2BPO: Permeable Penalty Barrier-Based Policy Optimization for Safe RL. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21029–21036, 2024.

[15] Lars D'Hondt, Friedl De Groote, and Maarten Afschrift. A dynamic foot model for predictive simulations of human gait reveals causal relations between foot structure and whole-body mechanics. *PLOS Computational Biology*, 20(6):e1012219, June 2024.

[16] Antoine Falisse, Gil Serrancolí, Christopher L. Dembia, Joris Gillis, Ilse Jonkers, and Friedl De Groote. Rapid predictive simulations with complex musculoskeletal models suggest that diverse healthy and pathological human gaits can emerge from similar control strategies. 16(157):20190402.

[17] Thomas Geijtenbeek. Scone: Open source software for predictive simulation of biological motion. *Journal of Open Source Software*, 4(38):1421, 2019.

[18] Thomas Geijtenbeek. The Hyfydy simulation software, 11 2021. `https://hyfydy.com`.

[19] Thomas Geijtenbeek and Pierre Schumacher. sconegym: Gym environments for SCONE. `https://github.com/tgeijten/sconegym`, 2023. GitHub repository.

[20] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

[21] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[22] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[23] Prajwal Koirala, Zhanhong Jiang, Soumik Sarkar, and Cody Fleming. Feasibility informed advantage weighted regression for persistent safety in offline reinforcement learning. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 5837–5844, 2025.

[24] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *CoRR*, abs/1805.00909, 2018.

[25] Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-Point Policy Optimization under Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:4940–4947, 04 2020.

[26] Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into RL: A new take on trust region methods. In *International Conference on Machine Learning*. OpenReview.net, 2025.

[27] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann, 1999.

[28] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[29] Henry J. Ralston. Energy-speed relation and optimal speed during level walking. *Internationale Zeitschrift für angewandte Physiologie einschließlich Arbeitsphysiologie*, 17:277–283, 1958.

[30] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. Approximate inference and stochastic optimal control. *CoRR*, abs/1009.3958, 2010.

[31] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

[32] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[33] Julien Roy, Roger Girgis, Joshua Romoff, Pierre-Luc Bacon, and Christopher J. Pal. Direct behavior specification via constrained reinforcement learning. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2022.

[34] Pierre Schumacher, Thomas Geijtenbeek, Vittorio Caggiano, Vikash Kumar, Syn Schmitt, Georg Martius, and Daniel F.B. Haeufle. Emergence of natural and robust bipedal walking by learning from biologically plausible objectives. *iScience*, page 112203, March 2025.

[35] Pierre Schumacher, Daniel F. B. Haeufle, Dieter Büchler, Syn Schmitt, and Georg Martius. DEP-RL: embodied exploration for reinforcement learning in overactuated and musculoskeletal systems. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[36] Pierre Schumacher, Daniel F.B. Haeufle, Dieter Büchler, Syn Schmitt, and Georg Martius. DEP-RL: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems. `https://github.com/martius-lab/depRL`, 2023. GitHub repository.

[37] Ajay Seth, Michael Sherman, Jeffrey A. Reinbolt, and Scott L. Delp. Opensim: a musculoskeletal modeling and simulation framework for in silico investigations and exchange. *Procedia IUTAM*, 2:212–232, 2011. IUTAM Symposium on Human Body Dynamics.

[38] Oswin So and Chuchu Fan. Solving stabilize-avoid via epigraph form optimal control using deep reinforcement learning. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023.

[39] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pages 20423–20443. PMLR, 2022.

[40] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

[41] Hao Sun, Ziping Xu, Zhenghao Peng, Meng Fang, Taiyi Wang, Bo Dai, and Bolei Zhou. Constrained mdps can be solved by early-termination with recurrent models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[42] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *ICLR*, 2019.

[43] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[44] Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pages 945–952, 2006.

[45] Ilnura Usmanova, Yarden As, Maryam Kamgarpour, and Andreas Krause. Log Barriers for Safe Black-box Optimization with Application to Safe Reinforcement Learning. *Journal of Machine Learning Research*, 25(171):1–54, 2024.

[46] K. Veerkamp, N.F.J. Waterval, T. Geijtenbeek, C.P. Carty, D.G. Lloyd, J. Harlaar, and M.M. Van Der Krogt. Evaluating cost function criteria in predicting healthy gait. 123:110530.

[47] Nolan C. Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10630–10640. PMLR, 2021.

[48] Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3742–3750. PMLR, 2017.

[49] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10639–10646, 2021.

[50] Karl E. Zelik and Arthur D. Kuo. Human walking isn't all hard work: evidence of soft tissue contributions to energy dissipation and return. 213(24):4257–4264.

[51] Baohe Zhang, Yuan Zhang, Hao Zhu, Shengchao Yan, Thomas Brox, and Joschka Boedecker. Constrained reinforcement learning with smoothed log barrier function. *Trans. Mach. Learn. Res.*, 2025.

[52] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Xueqian Wang, Bo Yuan, and Dacheng Tao. Penalized Proximal Policy Optimization for Safe Reinforcement Learning. In *Thirty-First International Joint Conference on Artificial Intelligence*, pages 3719–3725, 07 2022.

[53] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6814–6822, 2023.

[54] 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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Appendix

Algorithmic Details

A Overview

This section collects only the engineering/implementation points that are shared across methods or are not already spelled out in App. B (VT-MPO details) or App. H (AS-SAC and uniform-prior structure).

Unifying viewpoint (what changes and what does not). Across all SDH-based variants, constraints enter learning solely through a continuation probability $\alpha(s,a) \in [0,1]$ computed from constraint signals, inducing shaped quantities

$$\tilde{r}(s,a) = \alpha(s,a)\,(r(s,a)+\eta), \qquad \tilde{\gamma}(s,a) = \gamma\,\alpha(s,a),$$

which appear *only* in critic targets. The replay buffer retains the raw quantities needed to construct these targets, which are transformed on demand using the current λ and η; target networks and actor-side updates (SAC or MPO) are otherwise unchanged. AS vs. VT is a semantic distinction of the CaI objective that affects how information costs are accumulated (main text), but the continuation model α and survival gate used for shaping are shared.

A.1 Environment termination and `done` handling

In our experiments, constraint events do *not* reset the simulator; "termination" is objective-level through $\alpha(s,a)$. Episodes end only when the environment returns `done`.

True terminals vs. time limits. For critic bootstrapping we follow standard off-policy practice:

- for true terminal transitions, we disable bootstrapping with $(1-\texttt{done})$ (so $\tilde{\gamma} = (1-\texttt{done})\gamma\alpha$);
- for time-limit truncations, we allow bootstrapping via the usual timeout mask (absorbed into `done` in notation).

A.2 Continuation models used in the experiments

The theory requires only $\alpha(s,a) \in [0,1]$. In practice, the choice of α is a modeling decision that controls how aggressively violations shorten the effective horizon and can materially affect off-policy stability. We use an *exponential continuation model* throughout our experiments. Given per-step violation magnitudes $c_i(s,a) \geq 0$, we use the smooth attenuation

$$\alpha_{\exp}(s,a) := \exp\Big(-\lambda \sum_{i\in I} c_i(s,a)\Big) \in (0,1], \tag{13}$$

which avoids discontinuities and yields stable critic targets under replay. This continuation is used for both AS-SAC and VT-MPO on Safety Gymnasium and for the Schumacher et al. style formulation on Hyfydy.

A.3 Practitioner's Guide

In practice, we recommend VT-MPO with exponential continuation (Exp) as the default configuration, because it gives the most reliable overall reward/cost trade-off: across Safety Gymnasium it is the

more conservative and stable option, consistently achieving feasible or near-feasible behavior with good returns and lower variance, while AS-SAC is better viewed as the high-reward alternative when some additional instability or tighter use of the cost budget is acceptable. Concretely, AS-SAC attains the highest rewards on several easier locomotion-style tasks (e.g., Ant, HalfCheetah, HumanoidVelocity, RacecarCircle), but VT-MPO is stronger when robust feasibility matters more, including tasks such as Hopper and PointPush, and it degrades more gracefully under a single shared hyperparameter setting. We suggest exponential continuation as the default because it is simple and empirically robust, with fixed-λ results and no-schedule ablations showing that performance is largely unchanged once λ is above a small threshold. CaT (11) uses a different continuation construction for legged locomotion; we do not use that construction here. All experiments in this paper use the exponential continuation in Eq. 13.

B VT-MPO: MPO WITH A SURVIVAL-SHAPED CRITIC

This section specifies the only deviation from standard MPO: we keep MPO's *actor-side* E/M updates unchanged, but replace the critic by the *survival critic* induced by SDH (Sec. 3). Concretely, constraints affect learning *only* through critic targets via

$$\tilde{r}(s,a) = \alpha(s,a)\,(r(s,a)+\eta), \qquad \tilde{\gamma}(s,a) = \gamma\,\alpha(s,a).$$

All KL budgets, dual updates, and policy-fitting details are exactly as in MPO.

What is different from original MPO? Relative to MPO, VT-MPO changes only:

1. the **score** used in the E-step (use Q_{surv} instead of the standard return critic), and
2. the **critic training target** (use shaped reward $\tilde{r}$ and variable discount $\tilde{\gamma}$).

Everything else (E/M steps, KL constraints, trust region options, action sampling) is standard MPO.

B.1 SURVIVAL CRITIC INDUCED BY SDH

Define the (unregularized) survival action-value function of a policy π as

$$Q^{\pi}_{\mathrm{surv}}(s,a) \;\coloneqq\; \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0} u_t\,\tilde{r}(s_t,a_t)\;\Big|\; s_0=s, a_0=a\Big], \qquad u_t \coloneqq \prod_{k=0}^{t-1}\tilde{\gamma}(s_k,a_k), \tag{14}$$

which is the unique fixed point of the variable-discount Bellman recursion

$$Q^{\pi}_{\mathrm{surv}}(s,a) = \tilde{r}(s,a) + \tilde{\gamma}(s,a)\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[\mathbb{E}_{a'\sim\pi(\cdot|s')}\big[Q^{\pi}_{\mathrm{surv}}(s',a')\big]\Big]. \tag{15}$$

This is the critic used by VT-MPO.

B.2 POLICY IMPROVEMENT (UNCHANGED MPO E/M STEPS)

Fix an update state distribution d (estimated from replay) and a behavior/reference policy π_0 (typically the current actor, as in MPO). Given a critic estimate $Q^{\pi_\theta}_{\mathrm{surv}}$, the nonparametric MPO improvement problem becomes

$$\max_{q}\;\; \mathbb{E}_{s\sim d}\Big[\mathbb{E}_{a\sim q(\cdot|s)}Q^{\pi_\theta}_{\mathrm{surv}}(s,a)\Big] \quad \text{s.t.} \quad \mathbb{E}_{s\sim d}\Big[\mathrm{KL}\big(q(\cdot\mid s)\,\|\,\pi_0(\cdot\mid s)\big)\Big] \leq \varepsilon, \tag{16}$$

(with $\int q(\cdot\mid s)\,da = 1$ for each s). The solution is the standard Boltzmann form

$$q^{\star}(a\mid s) \;\propto\; \pi_0(a\mid s)\,\exp\Big\{\beta^{-1}Q^{\pi_\theta}_{\mathrm{surv}}(s,a)\Big\}, \tag{17}$$

where the dual temperature $\beta>0$ is chosen to satisfy the KL budget (exactly as in MPO). The M-step then fits the parametric actor to $q^{\star}$ using the same objective/parameterization as MPO (e.g. weighted maximum-likelihood or reverse KL), optionally with an additional trust region to the previous actor. *No actor-side changes are required beyond substituting Q_{surv} for the usual critic.*

B.3 OFF-POLICY CRITIC LEARNING WITH VARIABLE DISCOUNT

VT-MPO learns $Q_\phi \approx Q^{\pi_\theta}_{\text{surv}}$ off-policy using replay and a target network ϕ^-. The only modification to standard off-policy regression is that targets use $(\tilde{r}, \tilde{\gamma})$.

Virtual termination and `done` handling. Constraint events do *not* reset the simulator; "termination" is purely objective-level through $\alpha(s, a)$. Environment termination is handled normally via `done`: for true terminal transitions we disable bootstrapping by multiplying the bootstrap term by $(1 - \texttt{done})$. (Time-limit truncations may still bootstrap, using the standard timeout mask.)

TD(n) replay. The replay buffer retains the raw transitions needed to construct the n-step target. For a sampled sequence of length $m \le n$,

$$\{(s_{t+j}, a_{t+j}, r_{t+j}, c_{t+j}, s_{t+j+1}, d_{t+j})\}_{j=0}^{m-1},$$

where terminal sequences are truncated, we compute at critic-update time, using the current λ and η,

$$\alpha_{t+j} = \exp\Big\{-\lambda \sum_i c_{i,t+j}\Big\}, \qquad \tilde{r}_{\eta,t+j} = \alpha_{t+j}(r_{t+j} + \eta), \qquad \tilde{\gamma}_{t+j} = (1 - d_{t+j})\gamma\alpha_{t+j}. \tag{18}$$

Define

$$R^{(m)}_{t,\tilde{\gamma}} \coloneqq \sum_{k=0}^{m-1} u_{t,k}\, \tilde{r}_{\eta,t+k}, \qquad u_{t,k} := \prod_{j=0}^{k-1} \tilde{\gamma}_{t+j}, \qquad u_{t,m} := \prod_{j=0}^{m-1} \tilde{\gamma}_{t+j}. \tag{19}$$

The TD(n) target is

$$y_t = R^{(m)}_{t,\tilde{\gamma}} + u_{t,m}\, Q_{\phi^-}(s_{t+m}, a'), \qquad a' \sim \pi_\theta(\cdot \mid s_{t+m}). \tag{20}$$

Remark (off-policy corrections). Original MPO suggests sequence-based corrections (e.g. Retrace) for off-policy evaluation. We do not apply them. Empirically, TD(n) with survival-shaped $(\tilde{r}, \tilde{\gamma})$ was stable in our settings.

C ALGORITHMIC DESCRIPTIONS OF VT-MPO AND AS-SAC

We present the VT-MPO and AS-SAC implementation-faithful pseudocode in Algorithms 1 and 2, respectively. Given nonnegative per-constraint violation magnitudes $c_{i,t}$, both algorithms use the exponential continuation $\alpha_t = \exp\{-\lambda \sum_i c_{i,t}\}$, reward $\tilde{r}_{\eta,t} = \alpha_t(r_t + \eta)$, and discount $\tilde{\gamma}_t = \gamma\alpha_t$.

Algorithm 1 VT-MPO (termination=Virtual; actor=MPO, critic=unregularized-survival-TD(n))

Initialize actor π_θ, prior π_0, critic Q_ϕ, target critic $Q_{\phi^-} \leftarrow Q_\phi$, replay buffer $\mathcal{B}$
Initialize rolling window of length n for raw transitions $(s, a, r, c, s', \texttt{done})$
while training **do**
  Collect data: run policy π_0 in the environment; observe $(s_t, a_t, r_t, c_t, s_{t+1}, \texttt{done})$
  Append $(s_t, a_t, r_t, c_t, s_{t+1}, \texttt{done}_t)$ to the rolling window
  If the rolling window has length n, store the raw window in $\mathcal{B}$; at episode end also store the remaining shorter windows
  Critic update: sample a raw sequence of length $m \le n$ from $\mathcal{B}$. Using the current λ and η:
  for $j = \{0, \dots, m-1\}$ **do**
    $\alpha_{t+j} = \exp\{-\lambda \sum_i c_{i,t+j}\}$, $\tilde{r}_{\eta,t+j} = \alpha_{t+j}(r_{t+j} + \eta)$, and $\tilde{\gamma}_{t+j} = (1 - \texttt{done}_{t+j})\gamma\alpha_{t+j}$
  end for
  Set $R^{(m)}_{t,\tilde{\gamma}} = \sum_{k=0}^{m-1}(\prod_{j=0}^{k-1} \tilde{\gamma}_{t+j})\tilde{r}_{\eta,t+k}$ and $u_{t,m} = \prod_{j=0}^{m-1} \tilde{\gamma}_{t+j}$
  Regress $Q_\phi(s_t, a_t)$ towards $y_t = R^{(m)}_{t,\tilde{\gamma}} + u_{t,m} Q_{\phi^-}(s_{t+m}, a')$, where $a' \sim \pi_\theta(\cdot \mid s_{t+m})$
  Actor update (standard MPO): perform MPO E/M steps using Q_{ϕ^-} as score
    (Implementation: MPO regression may use stacked states $\{s_t\} \cup \{s_{t+n}\}$ for improved coverage.)
  Periodically update target networks $\phi^- \leftarrow \phi$, and optionally a slow-moving actor
end while

Algorithm 2 AS-SAC (termination=AbsorbingState; actor=SAC, critic=regularized-survival-TD(0))

Initialize actor π_θ (tanh-Gaussian), critics Q_{ϕ_1}, Q_{ϕ_2} and targets $\phi_i^- \leftarrow \phi_i$; replay $\mathcal{B}$ stores $(s, a, r, c, s', \texttt{done})$
Fix a normalized uniform reference policy $\log \pi_0(a \mid s) = -\ell_c$ and set $b_{\mathrm{AS}} := \ell_c$
Initialize temperature $\kappa > 0$ (fixed or autotuned), SAC entropy target $\bar{\mathcal{H}}$ and parameter β_κ
while training **do**
  Collect transition: choose a_t randomly for $t < t_{\text{learn}}$, otherwise sample $a_t \sim \pi_\theta(\cdot \mid s_t)$; execute a_t and observe $(r_t, c_t, s_{t+1}, \texttt{done}_t)$
  Store $(s_t, a_t, r_t, c_t, s_{t+1}, \texttt{done}_t)$ in $\mathcal{B}$
  if $t \ge t_{\text{learn}}$ **then**
    Critic: sample batch from $\mathcal{B}$; using the current λ and η, compute $\alpha = \exp\{-\lambda \sum_i c_i\}$, $\tilde{r}_\eta = \alpha(r + \eta)$, and $\tilde{\gamma} = (1 - \texttt{done})\gamma\alpha$; sample $(a', \log \pi_\theta(a'|s'))$;
    set $\widehat{V}^-(s') = \min_i Q_{\phi_i^-}(s', a') - \mathrm{sg}(\kappa)(\log \pi_\theta(a'|s') + b_{\mathrm{AS}})$, where $a' \sim \pi_\theta(\cdot|s')$;
    set $y = \tilde{r}_\eta + \tilde{\gamma}\, \widehat{V}^-(s')$.
    Update ϕ_i by MSE to target y (twin critics)
    Periodically update
      Actor: sample $a_\pi \sim \pi_\theta(\cdot|s)$, optimize $\max_\theta \mathbb{E}[\min_i Q_{\phi_i}(s, a_\pi) - \mathrm{sg}(\kappa) \log \pi_\theta(a_\pi|s)]$
      if `autotune`, estimate $\widehat{\mathcal{H}}$ on sampled replay batch; update $\kappa \leftarrow [\kappa + \beta_\kappa(\bar{\mathcal{H}} - \widehat{\mathcal{H}})]_+$
      update target critics: $\phi_i^- \leftarrow \tau\phi_i + (1 - \tau)\phi_i^-$
  end if
end while

Experimental Results

Table of Contents

D Hyfydy Experiments **18**
D.1 Environments, metrics, and evaluation protocol 18
D.2 Schumacher-style constrained gait formulation 20
D.3 Constrained gait learning results 22
D.4 Diagnostic Variants 26
D.5 Hyperparameters 29
D.6 Compute and runtime 29

E Experiments on Safety Gymnasium **31**
E.1 Details on the Experimental Setup 31
E.2 Main results on Safety Gymnasium 31
E.3 Violation-depth profiles across Safety Gymnasium 32
E.4 Fixed-λ ablations 34
E.5 Increased step budget 34
E.6 Evaluated AS baselines and ablations 34
E.7 Compute and runtime 36

D Hyfydy Experiments

This section details the Hyfydy experiments: full-formulation results achieving state-of-the-art realism, followed by diagnostic variants.

D.1 Environments, metrics, and evaluation protocol

Musculoskeletal environments. We evaluate on three humanoid musculoskeletal models simulated with Scone-Hyfydy (17; 18). The models increase in dimensionality and actuation redundancy from H0918 to H2190, with H2190 being the hardest setting utilizing 90 muscle actuators. The models are among the most realistic options for studying gait with reinforcement learning, as they provide high simulation speed compared to e.g. OpenSim (37), while being significantly more realistic than, e.g., MuJoCo (43; 34). See Table 1 for an overview of key model properties. We generally use the `v1` version of the sconegym (19) environments.

Metrics for gait characterization. We consider the following metrics in our humanoid gait experiments:

- **Forward velocity.** Non-zero velocity is defining for the state of locomotion and thus essential for judging gait dynamics. When we speak about velocity, we refer to the x-component of the center of mass velocity of the model throughout this document. Note that we are targeting natural *walking*, not running, in this work. This means that our goal is *not* to maximize velocity, but rather to bring it to a range that is natural for human walking ($\approx 1.2\,\mathrm{m\,s^{-1}}$ (29)).
- **Effort.** We define p-effort (with $p \geq 2$) to be the muscle averaged p-th power of the muscle activation a:

$$\mathrm{effort}^p = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} a_m^p, \tag{21}$$

  where $\mathcal{M}$ is the model's set of muscles. The metric has a tight relationship to the metabolic energy consumption, but is notably *not* equivalent. Effort penalizes high activations more than metabolic consumption rates would suggest, and reflects the disproportionate preference of humans for low

Table 1: Hyfydy model overview. The models we use for our experiments span different complexity levels, ranging from 2D restricted models with only 9 muscles and neglecting trunk muscles, to full range of motion with 90 muscles and trunk muscles being modeled.

Model	Motion restriction	DOFs	Muscle count	Trunk muscles	Appearance
H0918	Sagittal plane only (2D)	9	18	✗	
H1622	None	16	22	✗	
H2190	None	21	90	✓	

activations to prevent muscle fatigue (3). Effort definitions often vary in which exponent of the effort ($p = 2, 3, 4, \dots$) they use, and it is debated which exponent provides the best optimization target. In order to stay consistent with Schumacher et al. (34), who also optimize 3-effort but report 2-effort, we report effort with $p = 2$ in our visualizations.

- **Episode length.** Number of simulated control steps (each with a time duration of $0.025\,\mathrm{s}$) before environment termination or timeout (1000 steps). Higher values indicate stable sustained gait. We consider a gait to be stable, if the average attained episode length is higher than 900 steps ($22.5\,\mathrm{s}$).
- **Ground Reaction Forces (GRF).** The force the ground exerts back on the foot when the foot is pushed into the ground. Humans choose motion patterns that lack peaks in ground reaction forces, making it a key metric for gait naturalness (46; 16). Gait simulations, in contrast, tend to produce unrealistically high ground reaction forces (34; 9). One hypothesis for this is that soft tissues, which are mostly not modeled in musculoskeletal simulations, help dampen peaks in the force signal (50), making peaks more likely in models which lack them. Notably, our models lack soft tissues entirely.
- **Joint limit violation.** The total torque exerted by joint limits. Model joints are constrained to anatomically plausible ranges of motion; when a joint reaches its limit, the simulator applies a restoring force to prevent further excursion. Humans select movement patterns that avoid forceful contact with their joint limits, as repeated or excessive limit forces risk joint damage.
- **Human gait alignment.** We report the gait match percentage as calculated in SCONE Studio's gait analysis (17). For each evaluated joint, detected gait cycles are time-normalized and averaged into a single mean trajectory, which is then compared against a reference norm band derived from healthy human walking data (see green band in Figure 3). At each sample point, the excess outside the norm envelope is computed and normalized by the band width; the per-joint score is $100 \times \max(1 - \bar{e},\, 0)$, where $\bar{e}$ denotes the mean normalized excess. The overall alignment score is the arithmetic mean across all evaluated joints (Hip, Knee, Ankle) and the ground reaction force, yielding a value in $[0, 100]$ where higher values indicate closer agreement with human gait kinematics. We report values from our own Python reimplementation of the score. Numerical differences to SCONE Studio's implementation arise primarily from gait cycle detection and time discretization (we use uniform step sizes, whereas SCONE uses variable-step simulation time points). Per-joint scores can differ by up to $\sim 15\%$, though overall alignment scores are consistent in magnitude. We note that the metric is not highly discriminative — it is best understood as a coarse indicator of gait realism rather than a precise ranking tool.

Training and evaluation protocol. We train each method for 100 million environment steps and report mean and standard deviation over three random seeds (in selected plots mean, min and max).

Evaluation is performed every 200000 steps using 10 deterministic rollouts. Our algorithms were implemented within the DEP-RL code base (36), thus sharing many structural components with EWA, which is also part of the code base, and thus ensuring a maximally fair comparison. EWA runs were performed with the original implementation of Schumacher et al.. All methods use the same simulator settings, time limits, observation preprocessing, action bounds, and DEP exploration settings. The exact hyperparameters used are listed in Table 7.

D.2 SCHUMACHER-STYLE CONSTRAINED GAIT FORMULATION

We cast the problem solved by EWA (34) into our instantaneous constraint setting. We first recapitulate Schumacher et al.'s reward and the EWA reweighting procedure, which provides relevant context for RL based gait learning and motivates our continuation and reward design.

Schumacher et al.'s reward function. Schumacher et al.'s reward is used in the standard MDP setting with cumulative discounted returns. It consists of three components, each targeting a different aspect of natural gait:

$$r = r_{\text{vel}} - c_{\text{effort}} - c_{\text{pain}} \tag{22}$$

where

$$r_{\text{vel}} = \begin{cases} \exp\left[-(v_{\text{com}} - v_{\text{target}})^2\right] & \text{if} \quad v < v_{\text{target}} \\ 1 & \text{otherwise} \end{cases} \tag{23}$$

$$c_{\text{effort}} = \alpha(t)\,\text{effort}^3 + w_1 \sum_{m \in \mathcal{M}} (u - u_{\text{prev}})^2 + w_2 N_{\text{active}} \qquad c_{\text{pain}} = w_3 \sum_{j \in \mathcal{J}} \tau_j^{\text{lim}} + w_4 \sum_{l \in \mathcal{L}} F_l^{\text{GRF}}$$

Here, r_{vel} is the velocity objective, c_{effort} combines several effort related objectives (3-effort, activation smoothness and activation sparseness), and c_{pain} targets joint limit violations and ground reaction force, labeled jointly as "pain". Individual symbols are defined in Table 2. Note that the weights w_1 to w_4 were found via extensive hyperparameter search in the original publication, targeting maximal human gait alignment.

Table 2: Symbol definitions for reward and constraint specifications.

Symbol	Definition
v_{com}	x-velocity (forward direction) of the center of mass.
v_{target}	Target x-velocity.
$\alpha(t)$	Dynamically reweighted effort coefficient.
effort^3	3-effort (see Equation (21)) with third power activation exponent.
w_1, w_2, w_3, w_4	Objective weights, each representing a hyperparameter of the reward formulation.
$\mathcal{M}, \mathcal{J}, \mathcal{L}$	The sets of muscles, joints and legs of a model.
$\mathcal{H}, \mathcal{K}, \mathcal{A}$	Sets of hip, knee and ankle joints (containing the left and right joint each).
$u_m, u_{m,\text{prev}}$	Current and previous muscle excitation (policy output signal) for muscle m. This is *not* the activation a, which is distinct from the control signal u as excitations first get translated into activations via the activation dynamics of the model.
N_{active}	Number of currently active muscles. A muscle is considered to be active if its activation a_m is greater than 0.15.
τ_j^{lim}	Joint torque produced by model joint limit constraints of joint j.
F_l^{GRF}	Ground reaction force produced by leg l measured in terms of body weight: $F/(M \cdot g)$.

Effort Weight Adaptation (EWA). EWA is a heuristic, adaptive reward scheduling procedure that resolves the tension between effort minimization and other gait objectives. Since effort optimality is trivially attained by a non-acting policy, it conflicts with every other objective. EWA balances these

Algorithm 3 Effort Weight Adaptation

Input: threshold θ, smoothing β, adaptation rate $\Delta\alpha$, decay rate λ
$R_{\text{mean}} \leftarrow 0, \alpha \leftarrow 0, c_{\text{mean}} \leftarrow 0$
for episodes $t = 0, 1, 2, \dots$ **do**
  $R \leftarrow$ `train_episode()` // Train and get return R
  $R_{\text{mean}} \leftarrow \beta R_{\text{mean}} + (1-\beta)R$
  if $R_{\text{mean}} > \theta$ **and** $c_{\text{mean}} < 0.5$ **then**
    $\Delta\alpha \leftarrow \lambda\Delta\alpha$ // Convergence decay
  else if $R_{\text{mean}} > \theta$ **and** $c_{\text{mean}} > 0.5$ **then**
    $\alpha(t+1) \leftarrow \alpha(t) + \Delta\alpha$ // Increase effort pressure
  else
    $\alpha(t+1) \leftarrow \alpha(t) - \Delta\alpha$ // Relax effort pressure
  end if
  $c_{\text{target}} \leftarrow \mathbb{I}(R_{\text{mean}} > \theta)$
  $c_{\text{mean}} \leftarrow \beta c_{\text{mean}} + (1-\beta)c_{\text{target}}$
end for

by dynamically adjusting the effort weight $\alpha(t)$ based on a moving average of the task return R_{mean} (Algorithm 3).

We categorize EWA informally as being a "Lagrange style" constrained RL approach. This is because the return-based update logic of a sub-objective weight is reminiscent of the dual update in Lagrangian constrained RL in the CMDP setting. Notably, the adaptation logic of the "Lagrange multiplier" $\alpha(t)$ is more complicated than vanilla Lagrange dual updates, but modifications to Lagrange multiplier updates are commonly found in constrained RL literature, for example in PID Lagrangian RL (40), where they smooth training dynamics. We interpret the logic in Algorithm 3 as serving a similar purpose.

Interpretation of Schumacher et al.'s gait learning procedure. The structural similarity of EWA to primal-dual optimization suggests the following interpretation: EWA implicitly solves a *constrained problem* where pain, velocity and non-p-effort terms form the *constraint*, and p-effort is the *objective*. In other words, behavior is constrained to resemble a natural gait, and the optimizer seeks the lowest-effort solution within this feasible set.

SDH reformulation. We now cast Schumacher et al.'s procedure into our SDH framework. According to our interpretation, effort must be the objective, which is why we choose the following reward function:

$$r = 1 - \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{a_m^3}{a_{\text{max}}^3}, \tag{24}$$

where a_{max} is the maximal possible activation, enforced on the model side to be 0.5. Note that this ensures $r \in [0, 1]$, with 0 being achieved with maximal activation and 1 for zero activation. Other gait objectives are encoded as constraints:

Constraint terms. Let $c_i(s, a) \geq 0$ denote nonnegative violation magnitudes. We map each of the non p-effort objectives to one or multiple of these violation magnitudes, summarized in Table 3.

Similarly to Schumacher et al. (34), we performed an extensive hyperparameter search to find constraint thresholds which yield the highest human match percentage. Our problem formulation is the most sensitive to the knee torque limit. H1622 is uniquely sensitive to the lumbar torque limit. Unlike reward-based formulations, where penalty terms implicitly adapt to model morphology by summing over existing joints, explicit constraint thresholds must reflect model-specific biomechanics — e.g., models with active trunk muscles require fundamentally different lumbar torque limits than those without. The resulting thresholds are, however, physically interpretable (e.g., Nm of joint torque) rather than abstract reward weights.

Table 3: Hyfydy constraint violation definitions with thresholds and normalizations. See Table 2 for definitions of the individual symbols. Due to some models being able to control the lumbar torque, while others do not (see Table 1), we use different limit torques depending on run and model. For H2190 we always use $10\,\mathrm{N\,m}$. For H1622 and H0918 we used $80\,\mathrm{N\,m}$, with the exception for our highest gait match H1622 run, which after tuning resulted in a $40\,\mathrm{N\,m}$ threshold, but showed slightly diminished sample efficiency.

Constraint	Violation $c_i(s,a)$	Threshold	Normalization
Minimum velocity	$[v_{\text{min}} - v]_+$	$1.085\,\mathrm{m\,s^{-1}}$	v_{min}
Maximum velocity	$[v - v_{\text{max}}]_+$	$1.315\,\mathrm{m\,s^{-1}}$	v_{max}
GRF	$\left[\max_{l\in\mathcal{L}} F_l^{\text{GRF}} - F_{\text{max}}\right]_+$	1.3	1.0
Muscle count	$[N_{\text{active}}/\lvert\mathcal{M}\rvert - p_{\text{max}}]_+$	0.33	1.0
Hip violation	$\left[\max_{h\in\mathcal{H}} \tau^{\text{lim}}_{\text{hip},h} - \tau^{\text{lim}}_{\text{hip,max}}\right]_+$	$10\,\mathrm{N\,m}$	$\tau^{\text{lim}}_{\text{hip,max}}$
Knee violation	$\left[\max_{k\in\mathcal{K}} \tau^{\text{lim}}_{\text{knee},k} - \tau^{\text{lim}}_{\text{knee,max}}\right]_+$	$80\,\mathrm{N\,m}$	$\tau^{\text{lim}}_{\text{knee,max}}$
Ankle violation	$\left[\max_{a\in\mathcal{A}} \tau^{\text{lim}}_{\text{ankle},a} - \tau^{\text{lim}}_{\text{ankle,max}}\right]_+$	$20\,\mathrm{N\,m}$	$\tau^{\text{lim}}_{\text{ankle,max}}$
Lumbar torque	$\left[\tau^{\text{lim}}_{\text{lumbar}} - \tau^{\text{lim}}_{\text{lumbar,max}}\right]_+$	$\{10, 40, 80\}\mathrm{N\,m}$	$\tau^{\text{lim}}_{\text{lumbar,max}}$

Continuation construction. Given the per-constraint violations, we construct the SDH continuation probabilities α_i, using the following exponential continuation design:

$$\alpha_i(s,a) = \exp(-\lambda\,\bar{c}_i(s,a)), \tag{25}$$

where $\bar{c}_i$ is the normalized violation and λ is a hyperparameter. We explore the influence of λ on our solutions in Section D.4.1. As a default we choose $\lambda = 0.2$, which works reliably across models and experiments. The individual continuations are aggregated via a product:

$$\alpha(s,a) = \prod_{i=1}^{m} \alpha_i(s,a) = \exp\left(-\lambda \sum_i \bar{c}_i(s,a)\right), \tag{26}$$

allowing us to simply sum individual violations, to obtain a global one.

Critic target. For every replay transition, the critic target uses

$$\tilde{r}_t = \alpha(s_t,a_t) r(s_t,a_t), \qquad \tilde{\gamma}_t = (1-d_t)\gamma\alpha(s_t,a_t), \tag{27}$$

with standard timeout handling for time-limit truncations. VT-MPO then uses the survival critic and the standard MPO E/M update, as described in Section B. We choose VT-MPO due to the higher stability observed in our ablations (Section D.4).

D.3 Constrained gait learning results

We now analyze the results of the Schumacher-style constrained gait formulation, focusing on constraint satisfaction, gait realism, and comparison with EWA.

Constraint satisfaction and objective maximization. The training dynamics of selected gait metrics are visualized in Figure 3**a**) and compared with EWA. The most notable features of our method are the *speed* and *stability* with which we converge to a solution. Effort reaches comparable results on the smaller models. On H2190 we see a slight gap in effort minimization. A possible explanation is that the solution EWA recovers is infeasible within our problem formulation — Figure 3 shows that the EWA policy violates our constraints. There might not exist a feasible policy with equal effort as the EWA solution.

The sample efficiency curves of all cost and reward terms are visualized in Figure 5. The curves show that average per-step violations quickly converge to near zero across all models, while the effort reaches a plateau at a low effort level.

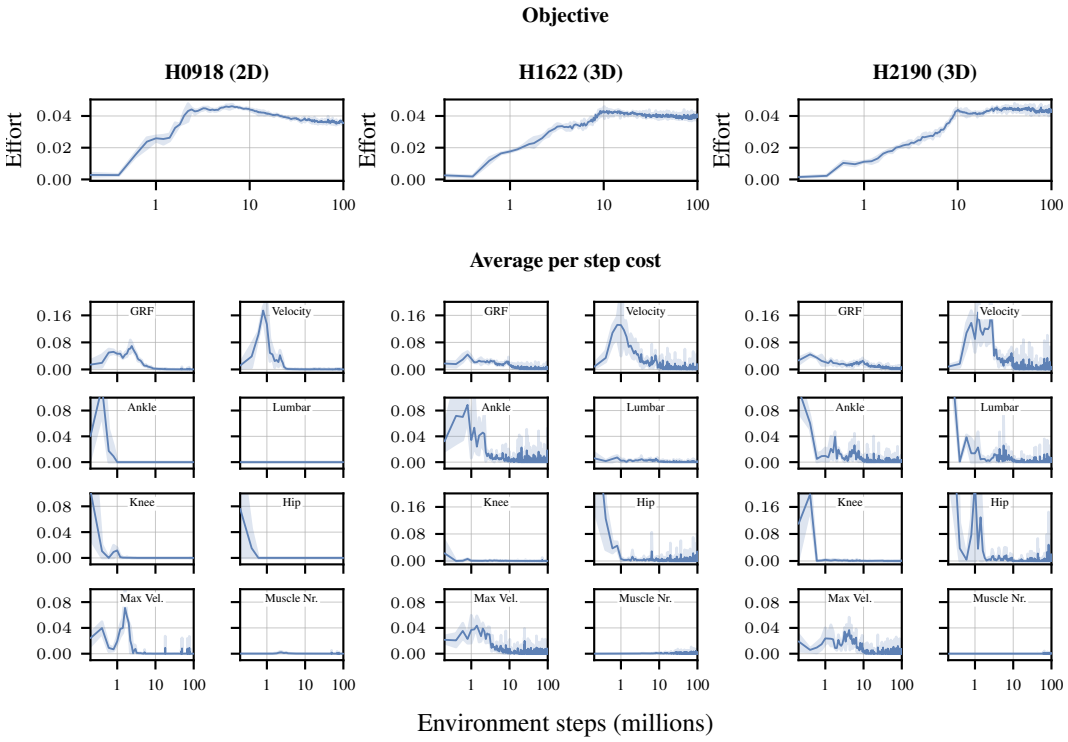


Figure 5: Objective and costs over the training for all three models. The plot shows mean and standard deviation over three seeds. We plot the x-axis logarithmically to emphasize early training dynamics, as later during the training costs and objective remain relatively constant. We see that the effort objective converges to low values, and the costs are reduced to near zero.

Human gait alignment. Both EWA and SDH aim to produce human-like gait without demonstrations. We evaluate gait realism using gait-match percentage. Every 200k environment steps, each of the three training seeds is evaluated using ten deterministic rollouts, yielding one seed-level mean gait-match percentage. These values are averaged across seeds at each evaluation step. Separately for each method, the best-observed checkpoint is the evaluation step with the highest across-seed mean. Figure 6 shows the full training trajectories, while Table 4 reports the mean $\pm$ standard deviation across seeds at the selected checkpoint.

Table 4: **Best-observed gait match and corresponding checkpoint.** Values are mean $\pm$ standard deviation across three seeds at the selected evaluation step.

	VT-MPO (ours)		EWA	
Model	Gait Match (%)	Checkpoint (M steps)	Gait Match (%)	Checkpoint (M steps)
H0918 (2D)	75.9 ± 2.5	4	75.8 ± 7.1	20
H1622 (3D)	70.1 ± 1.7	31	69.3 ± 4.0	41
H2190 (3D)	62.9 ± 2.1	21	64.6 ± 4.7	90

The best-observed gait-match values are similar across the two methods on all three models, whereas the selected VT-MPO checkpoint occurs earlier in each case. The largest separation occurs on H2190, at 21M environment steps for VT-MPO versus 90M for EWA. This timing difference is descriptive of the observed trajectories and should not be interpreted as an estimate of convergence time or sample complexity. Because gait match is not part of either training objective, its highest observed value need not occur late in training, consistent with the decline after the selected checkpoints in Figure 6.

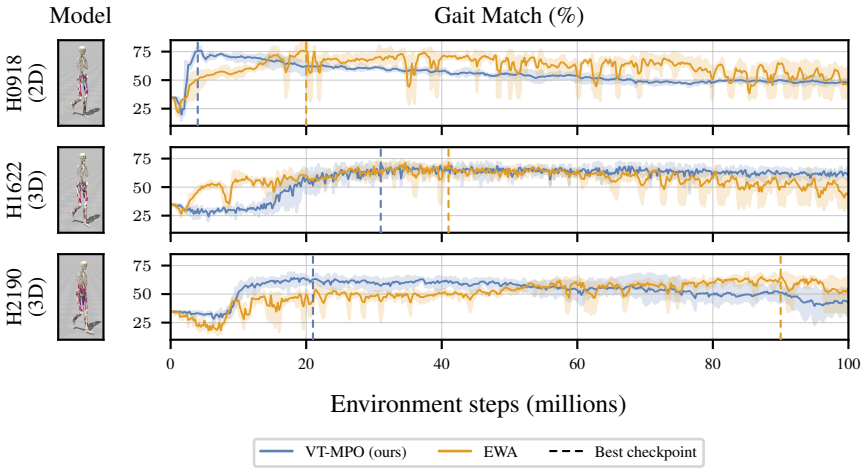


Figure 6: **Gait match over training.** VT-MPO and EWA on the three Hyfydy models. Dashed vertical lines mark the best-observed checkpoint selected separately for each method according to the protocol described above. Gait match is an external evaluation metric and is not optimized by either method.

Full-formulation AS-SAC comparison. To test whether AS-SAC provides a viable alternative under the same full Schumacher-style reformulation, we additionally evaluate fixed-κ AS-SAC using the complete set of local operating-limit violation signals from Table 3. This experiment is distinct from the velocity-only diagnostic in Section D.4.1. As shown in Figure 7, AS-SAC can attain higher human-match values on successful runs, most clearly on H0918 where it approaches 90%. However, its training is more seed-sensitive, with greater variability in gait acquisition and episode length, particularly on H2190. VT-MPO therefore remains our default for the full-formulation comparison because it provides the more robust optimization behavior, while AS-SAC is a viable but more aggressive alternative.

Interpretation of gait cycle curves in Figure 3b). We consider knee and hip kinematics to be equally well matched by both SDH and EWA on all models. For the GRF and ankle dynamics, we see a different picture: EWA generally possesses better ankle kinematics, while the ground reaction forces are more realistic for SDH. We hypothesize that there is a trade-off between realistic ankle dynamics, and realistic ground reaction force patterns in the considered models, possibly representing a model limitation.

As we constrain the ground reaction force for SDH, we hypothesize that the only feasible gait remains to be a toe first ground contact, which does not constitute a proper roll-off. This reduces the realism of the ankle kinematics. As discussed previously, a possible explanation is the missing soft tissue in the models, especially in the foot (15; 8).

Visual evaluation. We provide comparative videos on our project page, and static rollout visualizations in Figure 8. Both EWA and SDH yield visually comparable policies. For the H2190 model, we observe slightly less postural stability for EWA.

D.3.1 Survival-only diagnostic of the fixed $\eta = 0$ objective

For the Hyfydy reward $r = 1 - f$, the fixed $\eta = 0$ SDH objective already contains the survival-probability term:

$$(1-\gamma)\mu_\alpha^\pi(1-f) = p_{\text{surv}}(\pi) - (1-\gamma)\mu_\alpha^\pi(f).$$

We therefore ask whether removing the effort term entirely reveals an operating point with materially higher survival. On H1622, we train two otherwise identical VT-MPO agents using

$$Q_{\text{actor}} = w_r Q_r + (1-w_r)Q_1, \qquad w_r \in \{0,1\},$$

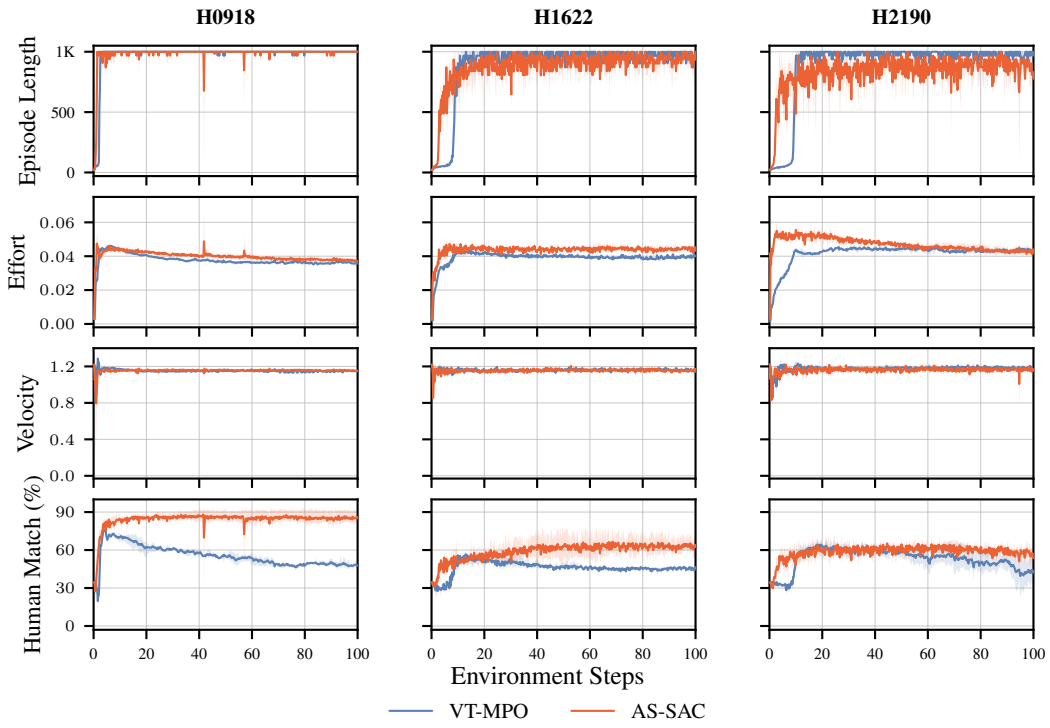


Figure 7: **Full-formulation AS-SAC versus VT-MPO on Hyfydy.** Training behavior under the same Schumacher-style reformulation and local operating-limit violation signals used in the main Hyfydy experiment. AS-SAC can attain higher human-match values on successful runs, especially on H0918, but exhibits greater seed sensitivity; VT-MPO is more robust on the H2190 model.

Table 5: **Task versus survival-only optimization on H1622.** Mean $\pm$ s.d. over three seeds, averaged over the final 5M environment steps. $\hat{p}_{\mathrm{MC}}$ is computed from evaluation rollouts; $\hat{p}_{Q_1}$ is the corresponding estimate from the learned survival-mass critic.

	Q_r only ($\eta = 0$)	Q_1 only
MC survival $\hat{p}_{\mathrm{MC}} \uparrow$	0.691 ± 0.007	0.683 ± 0.014
Critic estimate $\hat{p}_{Q_1}$	0.598 ± 0.006	0.596 ± 0.013
Task reward $\uparrow$	0.822 ± 0.006	0.794 ± 0.003
Violation magnitude / step $\downarrow$	0.0816 ± 0.0111	0.1114 ± 0.0073
Episode length $\uparrow$	922.4 ± 12.2	935.0 ± 19.4

where $w_r = 1$ is the task objective used by the main method (Q_r, $\eta = 0$), whereas $w_r = 0$ performs policy improvement using only the survival-mass critic Q_1. Both critics are trained in both runs; only the score used for MPO policy improvement changes. The continuation specification is fixed and no survival target or dual update is used.

The rollout estimate gives essentially the same survival at the two endpoints (Table 5), and the training curves show that this similarity persists after convergence rather than arising from a single checkpoint (Figure 9). In contrast, removing the effort term decreases task reward and increases the aggregate per-step violation magnitude. Thus, under this H1622 training protocol, direct optimization of survival mass does not uncover additional attainable survival beyond that obtained by the fixed $\eta = 0$ objective. This supports the practical role predicted by the decomposition above: the unit offset in $r = 1 - f$ supplies substantial survival pressure without requiring a separately optimized survival multiplier. We interpret this as an H1622 endpoint diagnostic, not as evidence that the resulting policy globally maximizes survival.

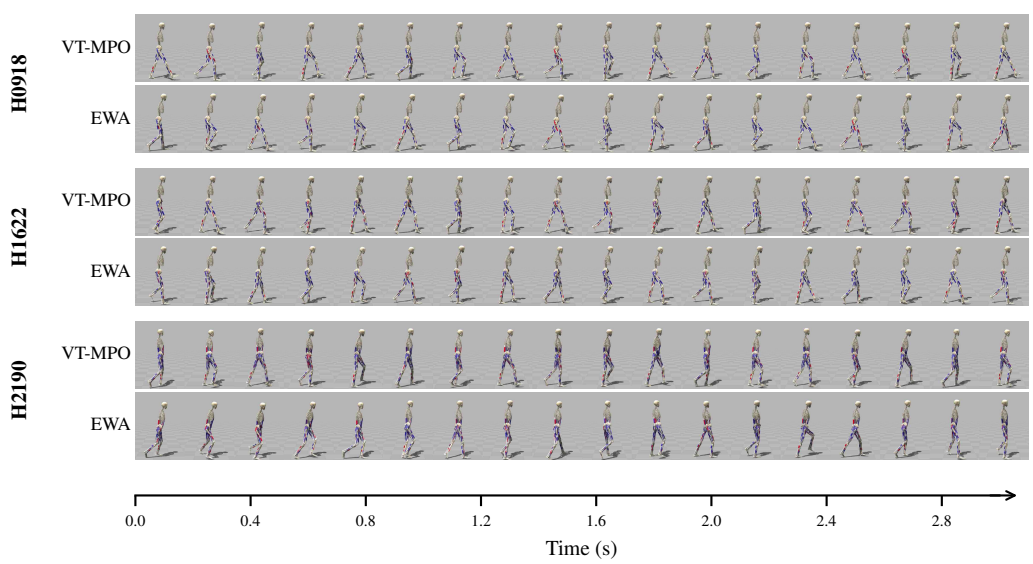


Figure 8: Representative full-formulation rollouts for H0918, H1622, and H2190 of SDH and EWA. Videos are available on the project page. We see no significant gap in visual realism between both approaches.

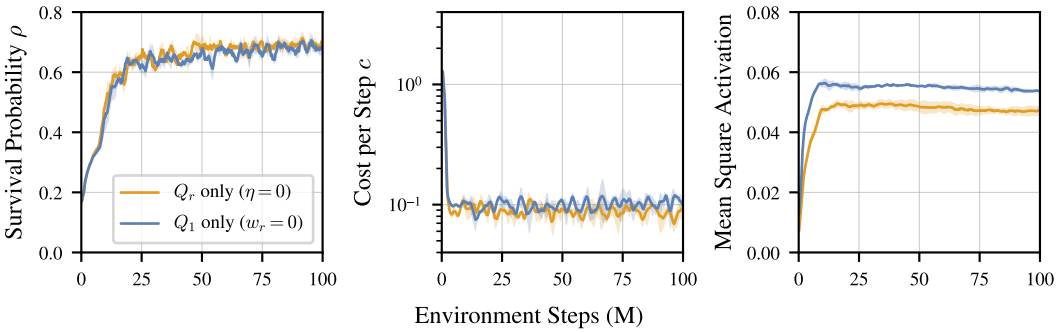


Figure 9: **Training dynamics for task versus survival-only policy improvement on H1622.** Optimizing Q_1 alone does not reveal a higher-survival regime: survival remains comparable to the fixed $\eta = 0$ task objective, while aggregate operating-limit violation and muscle effort are higher. Curves show mean and standard deviation over three seeds.

D.4 Diagnostic Variants

The following experiments are diagnostic and not part of the main comparison to Schumacher et al. (34). They isolate specific algorithmic properties of SDH under simplified problem formulations.

D.4.1 Velocity-only diagnostic setting

The velocity-only formulation uses a single instantaneous constraint, omitting the additional biomechanical constraints (GRF, joint limits, muscle count) of the full formulation.

While we retain the reward formulation in Equation (24), we modify the cost to only consist of the velocity violation:

$$c_{\text{vel}} = [v_{\min} - v]_+. \tag{28}$$

This setting focuses on the core tension of gait optimization: the balancing of effort and velocity.

AS-SAC vs. VT-MPO on Hyfydy. We compare AS-SAC and VT-MPO on the velocity-only diagnostic formulation to separate the effect of the base algorithm from the problem design. The main

empirical pattern is that VT-MPO is more robust on H2190, while AS-SAC can be more aggressive on easier models. AS-SAC fails learning to perform a walking gait on the H2190 model for one of the three seeds. See Figure 10 for a comparison of sample efficiency curves.

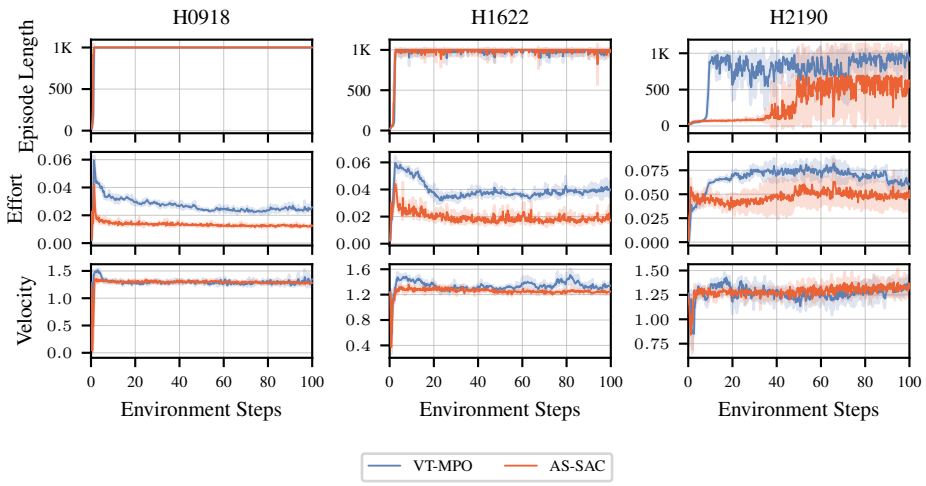


Figure 10: Sample efficiency curves for VT-MPO and AS-SAC on the velocity-only diagnostic problem formulation. Shown are mean and standard deviation over three seeds.

Table 6: Velocity-only diagnostic results at 100 million environment steps. Results reflect common knowledge about the tradeoff between MPO and SAC: SAC can optimize more aggressively, but is potentially unstable. MPO is more stable but peak performance (effort minimization) is lacking.

Model	Method	Effort ↓	Episode length ↑	Velocity
H0918	VT-MPO	0.026 ± 0.004	1000.00 ± 0.00	1.352 ± 0.076
H0918	AS-SAC	0.012 ± 0.001	1000.00 ± 0.00	1.279 ± 0.020
H1622	VT-MPO	0.040 ± 0.003	974.03 ± 36.72	1.346 ± 0.023
H1622	AS-SAC	0.017 ± 0.002	1000.00 ± 0.00	1.251 ± 0.010
H2190	VT-MPO	0.061 ± 0.007	901.43 ± 83.88	1.318 ± 0.040
H2190	AS-SAC	0.054 ± 0.009	622.93 ± 394.02	1.322 ± 0.050

An interesting side observation of these experiments is that effort reaches *worse* values with the simplified, velocity-only formulation, compared to the full problem formulation (see Figure 5). This suggests that added constraints are synergistic with the effort objective, in the sense that they seem to steer the training in the right direction. Final values at 100 million environment steps are given in Table 6.

To ensure fair comparison, and decouple algorithmic learning properties (SAC vs. MPO) from the influence of the problem design (EWA vs. SDH), we opted for VT-MPO for our final full problem formulation experiments, as EWA is also MPO based. While SAC might yield even faster convergence, we also expect a stability loss, which seems to be a downside on the H2190 model.

PID-Lagrangian comparison. For diagnostic comparisons, we evaluate SAC-PID-Lagrangian and MPO-PID-Lagrangian variants (40). These baselines adapt a multiplier-like parameter using PID feedback on constraint violations. The goal of the comparison is to illustrate that training instability is a general downside of objective reweighting based approaches, such as Lagrange multiplier based ones. We implemented PID Lagrange ourselves in the DEP-RL codebase, and tuned PID parameters on the considered environments. For the PID baseline we designed a similar effort-velocity problem setting in the CMDP framework, where the reward equals Equation (24), and the constraint is given

by:

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t c_t\right] \leq 0 \qquad \text{where} \qquad c = v_{\text{target}} - v. \tag{29}$$

As we have a cumulative constraint, not an instantaneous one, we choose $v_{\text{target}} = 1.2\,\mathrm{m\,s^{-1}}$, the midpoint of the instantaneous velocity band used in our full formulation. The sample efficiency curves are contrasted with SDH based approaches in Figure 11.

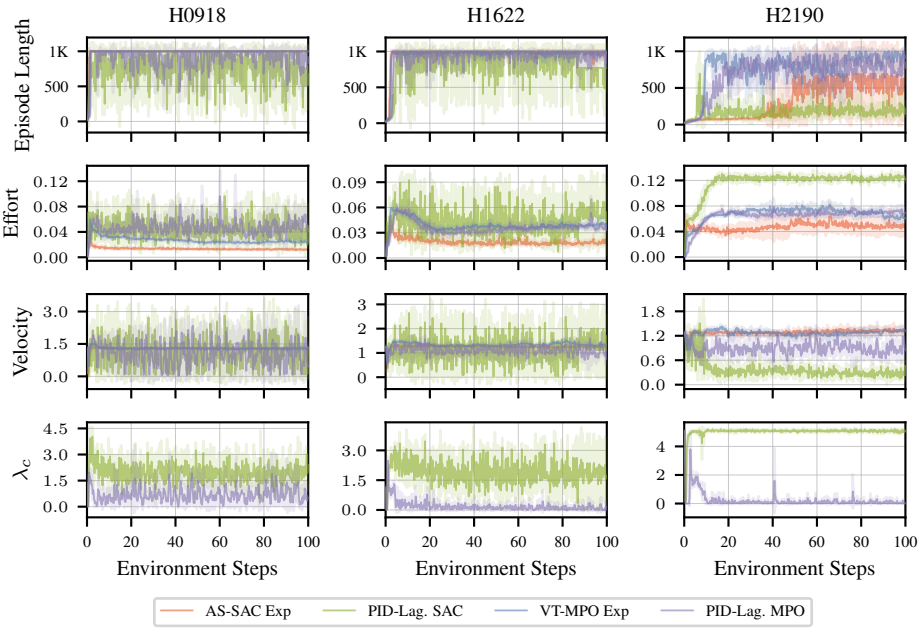


Figure 11: Comparison of PID Lagrangian based approaches (PID SAC, PID MPO) with SDH based approaches (AS-SAC, VT-MPO). Shown are mean and standard deviation for three seeds over 100 million environment steps.

There are three general patterns: PID SAC is more unstable than PID MPO, PID MPO and VT-MPO follow similar mean curve trajectories, and, most importantly: compared to SDH methods, even after tuning, PID methods still show significantly larger objective oscillations. This illustrates the difficulty of getting Lagrange based methods to work properly, clearly highlighting the practical benefits of our method.

Continuation schedule and strength ablations. We report fixed-strength and no-schedule ablations to characterize sensitivity to λ. The purpose is to support the narrower claim that SDH replaces adaptive dual-variable feedback with a continuation-design choice; it is not tuning-free.

In our standard setup we use a linear rampup of the λ parameter in the exponential continuation over the first 5 million (out of 100 million) environment steps. We ablated this design choice on a test run on H1622. The training curves are shown in Figure 12. Additionally, we evaluate the effects of choosing different λ values for the exponential continuation. The results are shown in Figure 13.

The results show that SDH is generally not sensitive to the exact choice of continuation-related hyperparameters. This robustness is a clear upside of our method.

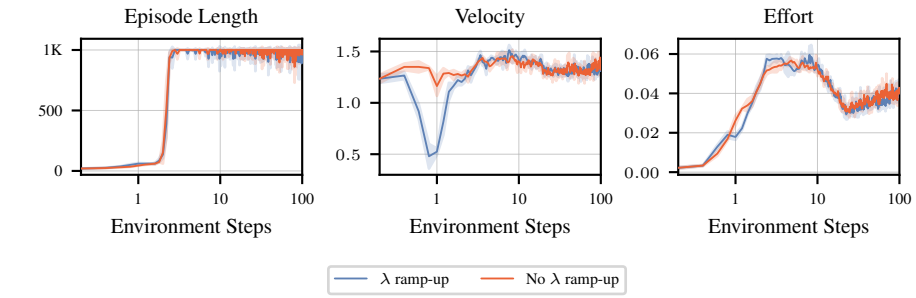


Figure 12: Ablation of initial λ ramp-up. The λ ramp-up has negligible effects on the training in the limit, only showing moderate influence in the very beginning of training.

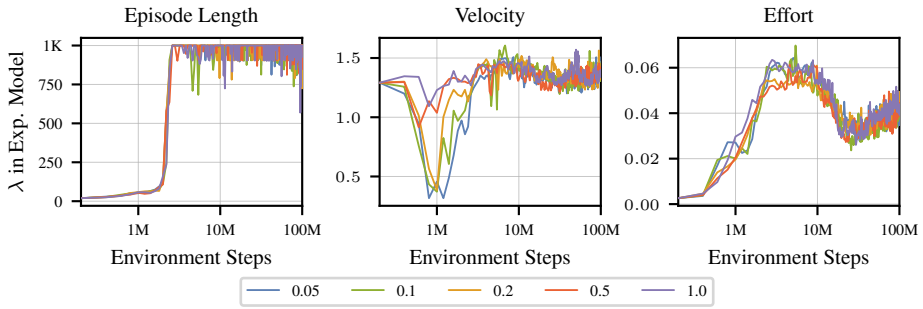


Figure 13: Runs on the velocity-only constraint formulation with SDH, for different choices of λ in the exponential continuation on H1622. We see that training dynamics are only affected in the initial phases of training, with the limit behavior being remarkably robust to λ.

D.5 Hyperparameters

D.6 Compute and runtime

SDH does not introduce additional environment interaction or policy rollouts relative to the underlying actor-critic method. The only algorithmic change is the critic target through $(\tilde{r}, \tilde{\gamma})$ and the computation of $\alpha(s, a)$ from constraint signals. For transparency, we report the compute used in Table 9; we do not claim a systematic wall-clock advantage.

Table 7: VT-MPO hyperparameters for Hyfydy.

Hyperparameter	Value
Discount γ	0.99
Replay buffer size	10^6
Batch size	256
Batch iterations	30
n-step returns	1
Warmup steps	2×10^5
Steps between batches	1 000
Total training steps	10^8
Parallel environments	20
Sequential steps	10
Network architecture	$\mathrm{MLP}(256, 256)$ / $\mathrm{MLP}(1024, 1024)^\dagger$
Actor learning rate	3×10^{-4} / $3.53 \times 10^{-5\dagger}$
Critic learning rate	10^{-3} / $6.08 \times 10^{-5\dagger}$
Dual variable learning rate	10^{-2} / $2.13 \times 10^{-3\dagger}$
Action samples per state	20
Continuation family	Exponential
λ schedule	$[0,\ 0.2]$
η	0
Schedule horizon	5×10^6

† H2190 uses $\mathrm{MLP}(1024, 1024)$ and tuned learning rates (second value); H0918 and H1622 use $\mathrm{MLP}(256, 256)$ with defaults (first value).

Table 8: DEP hyperparameters for Hyfydy.

Hyperparameter	Value
DEP intervention probability	0.00371
DEP intervention length	8
DEP bias rate	0.002
DEP buffer size	200
DEP κ	1 000
DEP regularization	32
DEP τ_{DEP}	40

Table 9: Hyfydy compute summary. All runs use NVIDIA RTX 2080 on our cluster. SDH modifies only the critic target and incurs no additional runtime overhead.

Resource / setting	Value
Hardware	NVIDIA RTX 2080
Parallel environments	20
Total environment steps	10^8
Throughput H0918 (steps/s)	$\sim$1386
Throughput H1622 (steps/s)	$\sim$1278
Throughput H2190 (steps/s)	$\sim$870
Wall-clock H0918 (h)	$\sim$21
Wall-clock H1622 (h)	$\sim$25
Wall-clock H2190 (h)	$\sim$36

E EXPERIMENTS ON SAFETY GYMNASIUM

E.1 DETAILS ON THE EXPERIMENTAL SETUP

We evaluate on a subset of 8 continuous-control tasks from Safety Gymnasium (22), which provide locomotion and navigation-style environments with standardized cost signals (e.g., hazards). The benchmark consists of 4 locomotion and 4 navigation environments. The full list can be seen in Table 11. We report average episodic returns of reward and cost per episode in the form of sample efficiency curves measured against the number of environment interactions. Each training run is repeated for 5 seeds and the sample efficiency results are aggregated across seeds for each environment. We compare against unconstrained MPO (1), constrained off-policy baselines MPO-PID (40) and WCSAC (49) as well as constrained on-policy baselines CPO (2) and C-TRPO (26). Hyperparameters are taken from the respective publications, as they evaluate on the same environments. VT-MPO and AS-SAC introduce one additional hyperparemeter λ that controls the mapping from constraint violations to continuation probabilities.

For both SDH methods, we use the same linear schedule $\lambda \in [0, 0.9]$ over the first $500k$ environment steps. Replay retains the raw reward and violation information, and the SDH transformations are constructed at critic-update time using the current λ, as described in Sec. 4. All hyperparameters are summarized in Table 10.

Table 10: Hyperparameters for the off-policy algorithms on Safety Gymnasium.

Hyperparameter	**AS-SAC**	**WCSAC**	**MPO-PID**	**VT-MPO**
General				
Discount factor γ	0.99	0.99	0.99	0.99
Replay buffer size	10^6	10^6	10^6	10^6
Batch size	256	256	256	256
Warmup steps	5 000	5 000	1 000	1 000
Network architecture	MLP$(256, 256)$	MLP$(256, 256)$	MLP$(256, 256)$	MLP$(256, 256)$
SAC				
Actor learning rate	3×10^{-4}	3×10^{-4}	—	—
Critic learning rate	10^{-3}	10^{-3}	—	—
Target smoothing τ	5×10^{-3}	5×10^{-3}	—	—
Actor update frequency	2	2	—	—
Target network update frequency	1	1	—	—
Entropy coefficient α	adaptive	0	—	—
MPO				
Policy/critic learning rate	—	—	3×10^{-4}	3×10^{-4}
Dual variable learning rate	—	—	10^{-2}	10^{-2}
n-step returns	—	—	4	4
ε (non-parametric KL)	—	—	0.1	0.1
ε_μ (policy mean KL)	—	—	0.01	0.01
ε_σ (policy std KL)	—	—	10^{-6}	10^{-6}
ε_{pen} (action limit KL)	—	—	10^{-3}	10^{-3}
Target/policy update period	—	—	100	100
Action samples per state	—	—	20	20
Gradient norm clip	—	—	40	40
Constraint				
Cost limit	25.0	25.0	25.0	25.0
λ-schedule (0-500k steps)	$[0, 0.9]$	—	—	$[0, 0.9]$
SDH survival bonus η	0.1	—	—	0.1
CVaR confidence level	—	0.9	—	—
Dual variable β learning rate	—	5×10^{-4}	—	—
PID gains $(K_P,\, K_I,\, K_D)$	—	—	$(0.1,\, 0.01,\, 0.5)$	—

E.2 MAIN RESULTS ON SAFETY GYMNASIUM

Aggregate performance. Figure 4 shows that VT-MPO achieves the most favorable aggregate return-violation trade-off. Compared to unconstrained MPO, both VT-MPO and AS-SAC substantially

reduce cost while retaining competitive reward. Compared to the on-policy baselines, they reach similar or lower violation levels with substantially higher return.

Per-environment behavior. Figure 15 shows that this aggregate trend is consistent across most individual environments. VT-MPO and AS-SAC reliably suppress cost towards the end of training while improving reward rapidly early on, whereas CPO and C-TRPO are typically conservative early on and improve reward slowly, reflecting their on-policy nature. Overall, AS-SAC tends to achieve higher asymptotic return in environments where constraint satisfaction is maintained throughout training. VT-MPO, however, is more robust across tasks and seeds, which we attribute to MPO's KL-regularized policy improvement under variable discounting. Finally, in two of the eight environments (`CarButton` and `PointGoal`), neither VT-MPO nor AS-SAC produce a feasible policy.

Effect of survival weighting. Across environments, decreasing the scale of the continuation probability α over training improves the return-violation trade-off. Early in learning, higher α allows exploration of infeasible regions and reuse of infeasible transitions. Later, reduced survival weighting attenuates long-horizon credit assignment through infeasible states, guiding convergence toward feasible behavior. This supports interpreting SDH as a continuous relaxation of hard constraints rather than a binary feasibility mechanism.

Table 11: Expected returns (R) and costs (C) on Safety Gymnasium after one million environment steps. The cost limit is 25.0 mean cost return in all environments (22). Violating entries are marked red. The algorithm with the highest inter quartile mean return (25%) among the feasible ones in an environment is marked in bold. The larger side of the confidence interval is denoted by $\pm$, rounded to one decimal. In `CarButton` and `PointGoal`, no algorithm yields a feasible policy in the prescribed number of environment interactions.

		Ant	Half Cheetah	Humanoid	Hopper	Car Button	Point Goal	Racecar Circle	Point Push
VT-MPO	R	829 ± 99	2021 ± 71	2683 ± 531	1293 ± 186	15.5 ± 2.4	15.6 ± 2.0	1.0 ± -0.0	**1.8 ± 0.7**
(ours)	C	1.3 ± 0.1	1.6 ± 0.6	0.3 ± 0.7	0.2 ± 0.2	303 ± 27	59.6 ± 7.9	7.3 ± 6.8	13.4 ± 5.0
AS-SAC	R	**2779 ± 64**	**2608 ± 37**	**5549 ± 235**	681 ± 289	0.1 ± 3.4	25.6 ± 0.6	**26.6 ± 1.9**	-0.3 ± 0.4
(ours)	C	0.4 ± 0.2	0.2 ± 0.1	0.0 ± 0.0	3.5 ± 3.1	311 ± 170	57.6 ± 8.1	22.6 ± 11.6	80.7 ± 44.7
C-TRPO	R	1634 ± 198	1363 ± 182	857 ± 198	**1364 ± 65**	-0.5 ± 0.4	8.8 ± 4.9	1.4 ± 2.4	0.2 ± 0.2
	C	22.3 ± 9.7	17.4 ± 4.5	6.3 ± 2.7	8.4 ± 11.1	32.1 ± 35.7	28.8 ± 10.3	24.5 ± 15.9	25.7 ± 19.1
CPO	R	1522 ± 90	1163 ± 97	881 ± 79	1115 ± 28	-0.8 ± 0.4	8.2 ± 3.7	1.6 ± 2.1	0.4 ± 0.5
	C	20.0 ± 5.8	18.2 ± 9.5	7.2 ± 1.8	18.8 ± 32.2	41.0 ± 6.8	32.8 ± 5.3	21.5 ± 32.4	18.3 ± 6.9
PPO	R	44.6 ± 6.9	1671 ± 758	506 ± 7	491 ± 174	16.8 ± 1.2	25.2 ± 0.7	35.0 ± 2.7	0.7 ± 0.3
	C	1.4 ± 0.3	151 ± 266	1.3 ± 1.4	135 ± 30	333 ± 25	50.6 ± 6.1	194 ± 46	52.8 ± 36.4
MPO	R	1136 ± 185	8654 ± 126	2395 ± 1088	1643 ± 548	26.0 ± 0.9	28.0 ± 0.4	1.1 ± -0.0	6.3 ± 2.6
	C	254 ± 74	979 ± 0	1.2 ± 3.7	410 ± 172	250 ± 46	56.5 ± 9.1	10.8 ± 13.2	36.3 ± 13.3
MPO-PID	R	543 ± 64	2498 ± 14	2500 ± 74	1269 ± 222	0.5 ± 0.4	12.9 ± 1.9	1.0 ± 0.0	2.4 ± 0.5
	C	33.7 ± 5.1	47.9 ± 2.7	1.1 ± 0.2	76.7 ± 14.2	61.8 ± 17.2	29.5 ± 13.6	9.8 ± 13.3	30.0 ± 21.9
WCSAC	R	-2301.4 ± 843.1	1824 ± 677	1153 ± 1609	885 ± 529	-1.3 ± 0.1	**-0.1 ± 0.1**	25.8 ± 16.0	-1.1 ± 0.9
	C	0.1 ± 0.2	1.3 ± 0.6	0.4 ± 0.7	0.0 ± 0.0	47.1 ± 20.0	0.0 ± 28.1	140 ± 92	46.8 ± 60.8
SafePO Best	R	3297.3	3336.8	6620.7	1716.4	0.1	19.0	64.8	4.4
Alg. @10M.	C	23.6	1.1	0.0	5.4	11.9	22.9	20.2	23.5

E.3 Violation-depth profiles across Safety Gymnasium

To test the violation-depth profile scope claim, we use the same diagnostic in every Safety Gymnasium environment. We compute each profile from the converged unconstrained SAC policy, rather than selecting environments based on AS-SAC or VT-MPO outcomes. The profile analysis is therefore independent of SDH performance. Since Ω_π depends on the policy, this is not a policy-uniform statement. Instead, it is the same method-independent probe applied across environments.

The single-scale correspondence assumption states that

$$\Omega_\pi(b) = (1-\gamma)^{-1} \exp\{-b/\tau_\pi\} \qquad \text{for some } \tau_\pi > 0.$$

Empirically, this requires two features of the profile. First, the positive-depth profile must be well described by an exponential, which appears as an approximately linear profile on the logarithmic scale. Second, the fitted constant factor must match the value $1/(1-\gamma)$ required by this single-scale form. Figure 14 and Table 12 report both factors.

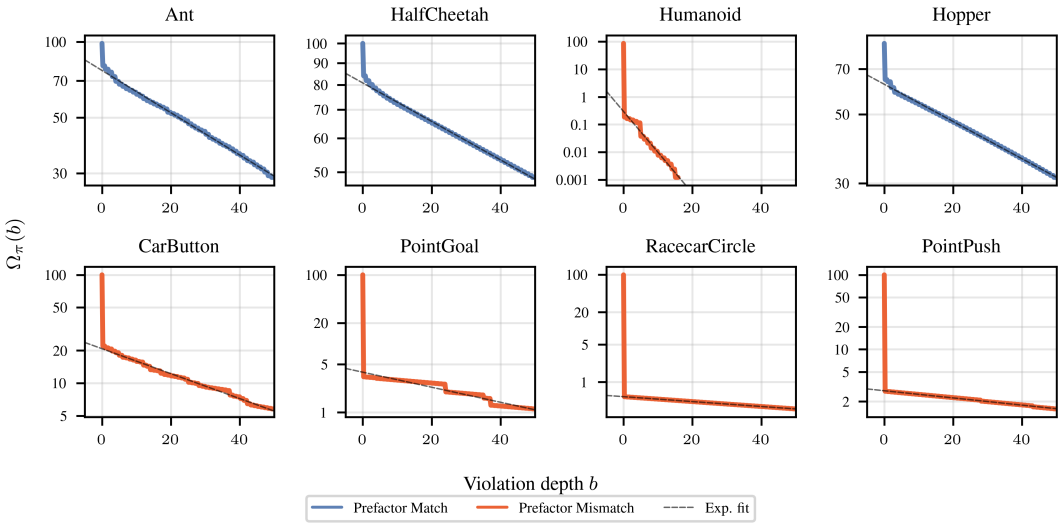


Figure 14: **Violation-depth profiles across Safety Gymnasium.** Profiles $\Omega_\pi(b)$ are computed from converged unconstrained SAC policies using the same procedure in every environment. Dashed lines show exponential fits over positive violation depths. `Ant`, `HalfCheetah`, `Hopper`, and `CarButton` have nearly log-linear positive-depth profiles, but the single-scale form additionally requires the fitted constant factor to equal $1/(1-\gamma)$. This second condition is approximately matched for `Ant` and `HalfCheetah`, but not for `Hopper` or `CarButton`. For `Humanoid`, `PointGoal`, and `PointPush`, positive violations occur in only 18%, 13%, and 11% of episodes, respectively, making the diagnostic inconclusive.

Table 12: **Violation-depth profile diagnostic.** R^2 measures the exponential fit of the positive-depth profile; A is the fitted prefactor; with $\gamma = 0.99$, the single-scale form requires $A = (1-\gamma)^{-1} = 100$; J_c is the discounted cumulative cost and τ is the fitted exponential scale. "Single-scale" denotes a near-exponential positive-depth profile. The final column reports whether the fitted prefactor approximately matches the value required by the single-scale form.

Environment	Type	τ	R^2	A	J_c	Prefactor
Ant	Velocity	51.6	0.997	76.9	38.4	Match
HalfCheetah	Velocity	96.5	0.997	80.8	80.1	Match
Humanoid	Velocity	2.9	0.977	0.3	0.0	Mismatch
Hopper	Velocity	72.4	0.998	62.5	36.7	Match
CarButton	Navigation	37.7	0.992	20.8	10.6	Mismatch
PointGoal	Navigation	39.7	0.942	3.8	4.2	Mismatch
RacecarCircle	Navigation	94.8	1.000	0.5	1.4	Mismatch
PointPush	Navigation	89.9	0.997	2.8	4.8	Mismatch

The sweep sharpens the empirical characterization of the boundary predicted by the theory. Log-linearity alone is not enough: `Hopper` and `CarButton` have almost perfectly exponential positive-depth profiles, yet their fitted constant factors differ substantially from the value required by the single-scale form. `Ant` and `HalfCheetah` are closest to satisfying both conditions. The difference is also reflected in the relation between J_c and the fitted τ: they are considerably closer for `Ant` and `HalfCheetah` than for `Hopper` and `CarButton`. For `Humanoid`, `PointGoal`, and `PointPush`, the unconstrained SAC probe rarely enters positive violation depths, so the profile test is inconclusive.

Accordingly, we use the violation-depth profile as a scope diagnostic for when the specific correspondence of Sec. 2.2 is empirically supported. Failure of this diagnostic does not imply that an SDH optimizer must fail or violate the native cost limit; it means only that one SDH Laplace statistic is not justified as a surrogate for additive cost by the stated single-scale result.

E.4 FIXED-λ ABLATIONS

We demonstrate additional fixed-λ ablations (no annealing) for exponential continuation with AS-SAC on a subset of environments in Table 13. The results demonstrate that SDH undergoes a "phase transition" in the environments `CarButton` and `PointGoal`, where the converged policy either satisfies the constraints with close-to-zero reward or achieves high reward by ignoring the constraint

Table 13: Expected returns (R) and costs (C) on Safety Gymnasium for **fixed-λ exponential continuation** with AS-SAC after one million environment steps. The cost limit is 25.0 mean cost return in all environments (22). Violating entries are marked red. The algorithm with the highest inter quartile mean return (25%) among the feasible ones in an environment is marked in bold. The larger side of the confidence interval is denoted by $\pm$, rounded to one decimal.

		HalfCheetahVelocity	CarButton	PointGoal
AS-SAC (lambda=0.005)	R	8797 ± 2271	11.3 ± 1.9	26.9 ± 0.4
	C	975 ± 4	232 ± 75	45.7 ± 3.1
AS-SAC (lambda=0.05)	R	**2918 ± 36**	-0.1 ± 0.3	**-0.3 ± 0.3**
	C	2.2 ± 0.5	77.4 ± 84.4	22.9 ± 6.8
AS-SAC (lambda=0.1)	R	2868 ± 64	**0.0 ± 0.0**	-0.3 ± 0.1
	C	1.2 ± 0.6	0.0 ± 0.0	34.7 ± 20.5
AS-SAC (lambda=0.2)	R	2817 ± 36	0.5 ± 0.4	-0.1 ± 0.2
	C	0.6 ± 0.2	143 ± 35	55.9 ± 80.6
AS-SAC (lambda=0.5)	R	2766 ± 57	0.1 ± 0.3	-0.1 ± 0.2
	C	0.3 ± 0.1	137 ± 108	69.4 ± 4.1

E.5 INCREASED STEP BUDGET

Table 14 shows performance on the most challenging Safety Gymanasium tasks from our benchmark for an increased step budget of 3 million steps. Both algorithms have substantially improved performance after 3 million steps on `PointGoal`, while reward remains close to zero on *CarButton* for our default setting.

Table 14: Expected returns (R) and costs (C) on the challenging *CarButton* and *PointGoal* environments in Safety Gymnasium with an **increased step budget of 3 million steps**. The cost limit is 25.0 mean cost return in all environments (22). Violating entries are marked red. The algorithm with the highest inter quartile mean return (25%) among the feasible ones in an environment is marked in bold. The larger side of the confidence interval is denoted by $\pm$, rounded to one decimal.

		CarButton	PointGoal
VT-MPO (ours)	R	0.8 ± 1.1	17.7 ± 1.9
	C	58.8 ± 9.4	30.3 ± 17.9
AS-SAC (ours)	R	**-1.6 ± 0.6**	**18.1 ± 3.3**
	C	4.0 ± 9.6	23.3 ± 11.4

E.6 EVALUATED AS BASELINES AND ABLATIONS

We distinguish three AS-SAC implementation variants: (i) **AS-SAC (adaptive κ)**, our main implementation, which uses the AS living-cost-aware critic together with standard SAC-style temperature adaptation (App. H.2.1), (ii) **AS-SAC (fixed κ)**, which uses the same critic but keeps κ constant, and (iii) **AS-SAC (no living cost)**, which *incorrectly* removes the prior-normalization term and the induced horizon regularization from AS policy evaluation. The "no living cost" ablation directly tests the theoretical prediction that, under AS, the uniform-prior term is not a removable constant, while the comparison between adaptive and fixed κ tests the effect of temperature adaptation. Table 15 shows that AS-SAC with adaptive κ achieves the highest feasible return among these AS-SAC variants on 3 out of 8 environments. We hypothesize that this variant strikes a favorable balance between the theoretically justified AS critic and optimization stability from standard SAC-style temperature adaptation with a constant entropy target. This is also the version we use in the main benchmark.

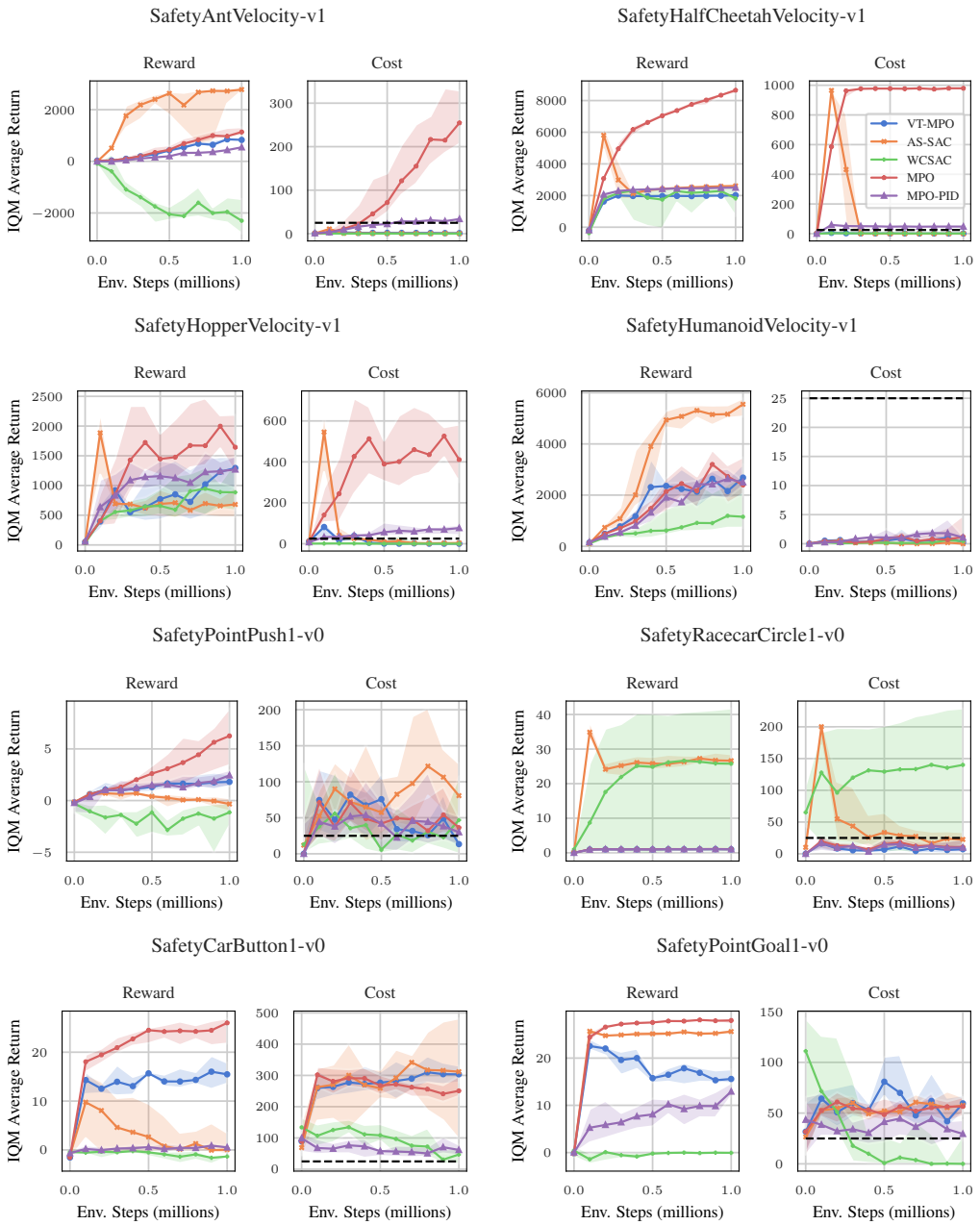


Figure 15: Results on the Safety Gymnasium benchmark for individual environments.

Table 15: Expected returns (R) and costs (C) on Safety Gymnasium for the AS-SAC ablations after one million environment steps with cost limit 25.0 mean cost return (22). The algorithm with the highest inter quartile mean return (25%) among the feasible ones in an environment is marked in bold.

		Ant	HalfCheetah	Humanoid	Hopper	CarButton	PointGoal	RacecarCircle	PointPush
AS-SAC (adaptive κ)	R	**2779 ± 64**	2608 ± 37	**5549 ± 235**	681 ± 283	0.1 ± 3.4	25.6 ± 0.6	**26.6 ± 1.9**	-0.3 ± 0.5
	C	0.4 ± 0.2	0.2 ± 0.1	0.0 ± 0.0	3.5 ± 3.1	311 ± 170	57.6 ± 8.1	22.6 ± 11.6	80.7 ± 42.4
AS-SAC (fixed κ)	R	2553 ± 18	2699 ± 17	5384 ± 343	**1009 ± 50**	-3.4 ± 1.5	-0.7 ± 0.3	0.1 ± 0.1	-0.6 ± 0.3
	C	0.5 ± 0.2	0.3 ± 0.0	0.0 ± 0.2	1.1 ± 9.2	270 ± 160	115 ± 66	0.3 ± 2.3	27.9 ± 22.0
AS-SAC (no-living-cost)	R	1895 ± 838	**2722 ± 15**	5314 ± 430	794 ± 196	3.3 ± 7.2	25.6 ± 0.3	26.2 ± 1.1	-0.6 ± 0.2
	C	0.3 ± 0.1	0.4 ± 0.4	0.0 ± 0.4	3.1 ± 22.9	305 ± 67	64.0 ± 6.2	22.4 ± 36.9	86.9 ± 31.1

E.7 Compute and runtime

SDH does not introduce additional environment interaction or policy rollouts relative to the underlying actor-critic method. The only algorithmic change is the critic target through $(\tilde{r}, \tilde{\gamma})$ and the computation of $\alpha(s, a)$ from constraint signals. The only extra critic appears in the survival-only diagnostic (App. D.3.1), where the survival critic Q_1 is trained alongside the task critic Q_r; no Lagrange dual update is used. For transparency, we report the compute used in Table 16; we do not claim a systematic wall-clock advantage.

Table 16: Safety Gymnasium compute summary. All runs use NVIDIA A100 on our cluster. SDH modifies only the critic target and incurs no additional runtime overhead.

Resource / setting	Value
Hardware	NVIDIA A100
Parallel environments	10 (1 for MPO)
Total environment steps	10^6
Throughput (steps/s)	$\sim$90
Wallclock time (h)	$\sim$3

Theoretical Foundations

Table of Contents

F Survival Occupancy and Stochastic Survival Chance Constraints **37**
F.1 Survival processes and chance constraints 38
F.2 The objective mismatch 39
F.3 Survival occupancy and the exact SDH chance problem 40
F.4 Two interpretable continuation models 40
F.5 Why local SDH is not an additive-budget CMDP solver 41
F.6 A positive regime: single-scale survival 42
F.7 Interpretation and scope 43

G Control as Inference with Stochastic Decision Horizons **44**
G.1 The ELBO constraint 44
G.2 Stochastic horizon, continuation, and endpoint reward 45
G.3 Semantic trajectory laws 45
G.4 The SDH ELBO 46
G.5 Reducing the endpoint and decision gates 47
G.6 AS/VT classification theorem 47
G.7 Consequences for dynamic programming 48
G.8 Stability of the variable-discount critic 49
G.9 Summary: ELBO-valid SDH regularization 49

H Absorbing State: Uniform Prior and AS-SAC **51**
H.1 AS objective under a normalized uniform prior 51
H.2 AS admits a SAC-style Bellman recursion 52

F Survival Occupancy and Stochastic Survival Chance Constraints

This appendix formalizes the chance-constrained interpretation of stochastic decision horizons. The purpose is not to reinterpret SDH as an ordinary additive-budget CMDP. Rather, we identify the precise constrained-control problem for which SDH is exact: a survival-gated chance problem evaluated at the same random time that represents discounting.

Figure 16 gives the conceptual map for the theory. The central distinction is that SDH treats violations as instantaneous-(in)feasibility signals rather than entries in a residual-budget ledger. This changes both the objective and the constraint: reward is counted only on prefixes that survive to the geometrically sampled evaluation time, and feasibility is summarized by the same survival event. The resulting formulation is therefore exact for a stochastic survival-chance problem, but not for arbitrary additive-budget CMDPs. The violation-depth profile makes this boundary precise: additive cost depends on the full area under the profile, whereas exponential SDH observes one Laplace-weighted slice. Thus SDH is predictive when violation geometry is effectively single-scale, and structurally mismatched when budget use is broad, history-dependent, or multi-scale.

The section has four steps. First, we relate SDH to previously studied survival chance constraints. Second, we show that SDH changes not only the constraint but also the objective: reward is counted only on surviving prefixes. Third, we prove the exact Lagrangian closure property of SDH. Finally, we separate this formulation from additive-budget CMDPs and identify regimes where SDH is exact or informative.

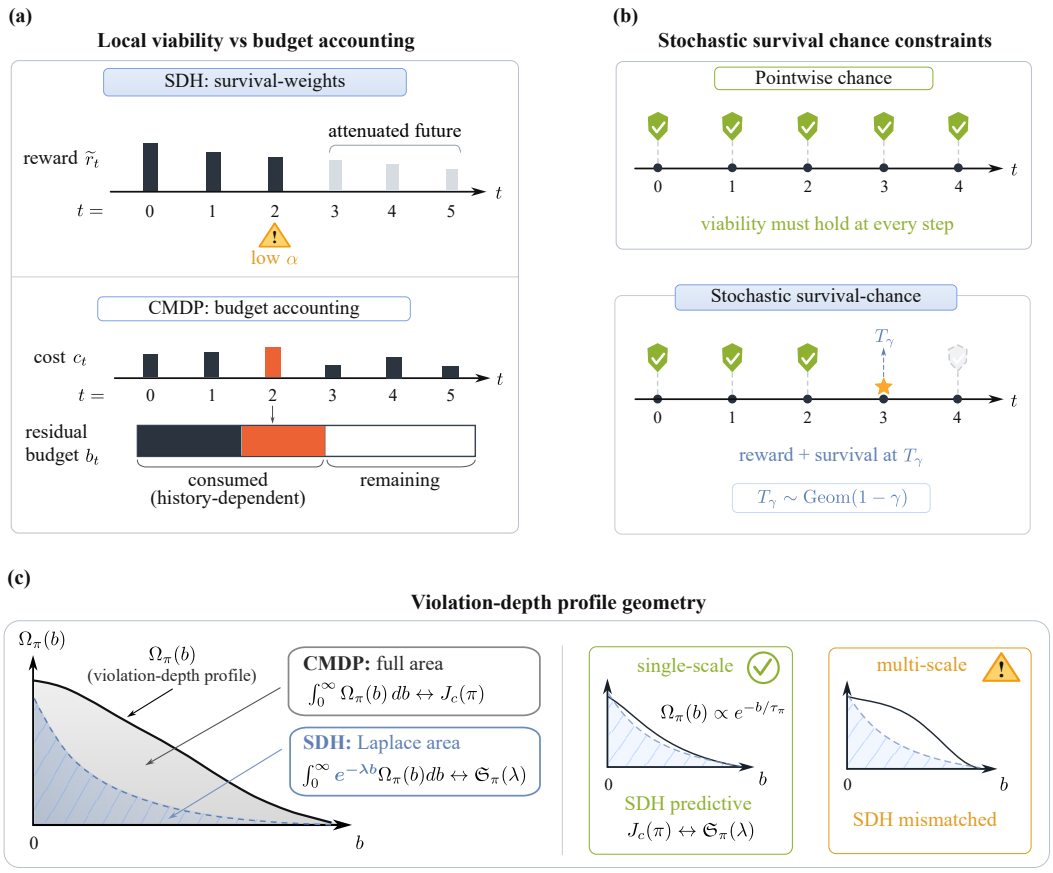


Figure 16: **Stochastic decision horizons as instantaneous-feasibility signals. (a)** CMDPs treat violations as additive cost in a history-dependent residual-budget ledger; SDH treats them as local continuation signals that attenuate reward and bootstrapping through $(\tilde{r} = \alpha r, \tilde{\gamma} = \gamma\alpha)$. **(b)** Unlike pointwise chance constraints, SDH evaluates reward and feasibility at the same geometrically sampled time T_γ, yielding a stochastic survival-chance objective. **(c)** The violation-depth profile explains the scope: additive cost is the full area under $\Omega_\pi(b)$, whereas exponential SDH observes one Laplace-weighted slice. Thus SDH is predictive in single-scale local-viability regimes, but not a general replacement for budget accounting.

F.1 Survival processes and chance constraints

We start with a formal survival process.

Definition F.1 (Continuation model and survival process)**.** Consider an infinite-horizon MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with $\gamma \in (0, 1)$ and bounded reward. A continuation model is a measurable function $\alpha : \mathcal{S} \times \mathcal{A} \to [0, 1]$. For a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$, define independent Bernoulli variables

$$C_t \mid \tau \sim \text{Bernoulli}(\alpha(s_t, a_t)), \qquad B_t := \prod_{k=0}^{t} C_k. \tag{30}$$

The event $B_t = 1$ indicates the stochastic survival process has survived through the reward at time t.

For a policy π, write $s_t(\pi) := \Pr_\pi(B_t = 1)$. A direct survival chance-constrained control problem requires these probabilities to remain high at every horizon.

Definition F.2 (Pointwise survival chance problem)**.** Given tolerances $(\delta_t)_{t\geq 0}$, the pointwise survival chance problem is

$$\begin{aligned} \max_{\pi} \quad & J_\gamma(\pi) := \mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t r(s_t,a_t)\right] \\ \text{s.t.} \quad & \Pr_\pi(B_t=1)\geq 1-\delta_t, \qquad t=0,1,2,\ldots. \end{aligned} \tag{31}$$

This formulation is multi-scale: it attaches a distinct feasibility requirement to every time horizon. Discounted control, however, has a natural random-time representation.

Definition F.3 (Discounted random time)**.** Let T_γ be independent of the trajectory and survival process, with

$$\Pr(T_\gamma=t)=(1-\gamma)\gamma^t, \qquad t=0,1,2,\ldots. \tag{32}$$

The random-time survival probability is exactly the geometrically weighted average of the pointwise survival probabilities.

Proposition F.4 (Pointwise survival implies random-time survival)**.** *Let $w_t=(1-\gamma)\gamma^t$. If $\Pr_\pi(B_t=1)\geq 1-\delta_t$ for every $t\geq 0$, then*

$$\Pr_\pi(B_{T_\gamma}=1)\geq 1-\sum_{t\geq 0} w_t\delta_t. \tag{33}$$

Thus the pointwise constraints imply the random-time survival constraint $\Pr_\pi(B_{T_\gamma}=1)\geq\rho$ whenever $\rho\leq 1-\sum_{t\geq 0}(1-\gamma)\gamma^t\delta_t$. The converse does not hold in general.

Proof. Independence of T_γ gives $\Pr_\pi(B_{T_\gamma}=1)=\sum_{t\geq 0} w_t\Pr_\pi(B_t=1)$. The lower bound follows immediately from the pointwise constraints.

For non-converseness, fix m and consider a deterministic one-action trajectory with $\alpha(s_t,a_t)=1$ for $t<m$ and $\alpha(s_m,a_m)=0$. Then survival is certain before m and impossible from m onward. Hence the pointwise constraints can fail for all $t\geq m$, while $\Pr_\pi(B_{T_\gamma}=1)=1-\gamma^m$, which can be arbitrarily close to one. □

This proposition identifies the constraint-side simplification made by SDH: it replaces infinitely many horizon-wise survival requirements by one discounted survival requirement. This is not yet the full SDH formulation, because the objective is also changed.

F.2 The objective mismatch

The pointwise chance problem in Def. F.2 maximizes ordinary discounted reward. In random-time form, ordinary discounted reward evaluates $r(s_{T_\gamma},a_{T_\gamma})$. SDH instead evaluates reward only if the trajectory has survived until T_γ.

Definition F.5 (Survival-gated random-time reward)**.** The survival-gated random-time reward of a policy π is

$$J_{\mathrm{surv}}(\pi):=\mathbb{E}_\pi[B_{T_\gamma}r(s_{T_\gamma},a_{T_\gamma})]. \tag{34}$$

Proposition F.6 (Objective mismatch)**.** *For every policy π,*

$$(1-\gamma)J_\gamma(\pi)-J_{\mathrm{surv}}(\pi)=\mathbb{E}_\pi[(1-B_{T_\gamma})r(s_{T_\gamma},a_{T_\gamma})]. \tag{35}$$

If $0\leq r(s,a)\leq R$ for every state-action pair (s,a), then

$$0\leq(1-\gamma)J_\gamma(\pi)-J_{\mathrm{surv}}(\pi)\leq R\Pr_\pi(B_{T_\gamma}=0). \tag{36}$$

Proof. The random-time identity for discounted reward gives $(1-\gamma)J_\gamma(\pi)=\mathbb{E}_\pi[r(s_{T_\gamma},a_{T_\gamma})]$. Subtracting Def. F.5 proves (35). If $0\leq r\leq R$, then $(1-B_{T_\gamma})r(s_{T_\gamma},a_{T_\gamma})$ is bounded between zero and $R(1-B_{T_\gamma})$, giving (36). □

Thus SDH is not merely a relaxation of pointwise chance constraints. It makes a matched pair of changes: survival is evaluated at the discounted random time, and reward is gated by the same survival event. This matching is exactly what produces the Lagrangian closure property below.

F.3 Survival occupancy and the exact SDH chance problem

We now define the invariant optimized by SDH.

Definition F.7 (Survival-weighted occupancy)**.** For a bounded measurable test function f, define

$$\mu_\pi^\alpha(f) := \mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t\left(\prod_{k=0}^{t}\alpha(s_k,a_k)\right)f(s_t,a_t)\right]. \tag{37}$$

The SDH return is $\mu_\pi^\alpha(r)$.

The corresponding Bellman representation uses shaped reward $\tilde{r}(s,a)=\alpha(s,a)r(s,a)$ and variable discount $\tilde{\gamma}(s,a)=\gamma\alpha(s,a)$. Since $\tilde{\gamma}(s,a)\leq\gamma$, the induced policy-evaluation operator remains a γ-contraction for bounded rewards.

Theorem F.8 (Survival occupancy identity)**.** *For every bounded measurable f,*

$$(1-\gamma)\mu_\pi^\alpha(f)=\mathbb{E}_\pi[B_{T_\gamma}f(s_{T_\gamma},a_{T_\gamma})]. \tag{38}$$

In particular, $(1-\gamma)\mu_\pi^\alpha(1)=\Pr_\pi(B_{T_\gamma}=1)$.

Proof. Conditioned on the trajectory, the survival variables are independent and $\mathbb{E}[B_t \mid \tau]=\prod_{k=0}^{t}\alpha(s_k,a_k)$. Averaging over the geometric time T_γ gives

$$\mathbb{E}_\pi[B_{T_\gamma}f(s_{T_\gamma},a_{T_\gamma})]=(1-\gamma)\mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t\left(\prod_{k=0}^{t}\alpha(s_k,a_k)\right)f(s_t,a_t)\right],$$

which is exactly (38). Taking $f\equiv 1$ gives the survival probability identity. □

The exact constrained problem associated with SDH can now be stated.

Definition F.9 (Survival-gated random-time chance problem)**.** Given a survival threshold $\rho\in[0,1]$, the SDH chance problem is

$$\begin{aligned}\max_{\pi}\quad & \mathbb{E}_\pi[B_{T_\gamma}r(s_{T_\gamma},a_{T_\gamma})]\\ \text{s.t.}\quad & \Pr_\pi(B_{T_\gamma}=1)\geq\rho.\end{aligned} \tag{39}$$

This formulation evaluates feasibility and reward at the same random decision time. That shared timing is the structural reason the Lagrangian remains inside the SDH objective class.

Theorem F.10 (Exact SDH Lagrangian)**.** *The Lagrangian of Def. F.9 is*

$$\mathcal{L}(\pi,\eta)=(1-\gamma)\mu_\pi^\alpha(r+\eta)-\eta\rho,\qquad \eta\geq 0. \tag{40}$$

Consequently, for fixed η, maximizing the Lagrangian over policies is equivalent to maximizing an SDH return with shifted reward $r+\eta$.

Proof. The Lagrangian is $\mathbb{E}_\pi[B_{T_\gamma}r(s_{T_\gamma},a_{T_\gamma})]+\eta(\Pr_\pi(B_{T_\gamma}=1)-\rho)$. Applying Thm. F.8 with $f=r$ and with $f\equiv 1$ gives $\mathbb{E}_\pi[B_{T_\gamma}r_{T_\gamma}]=(1-\gamma)\mu_\pi^\alpha(r)$ and $\Pr_\pi(B_{T_\gamma}=1)=(1-\gamma)\mu_\pi^\alpha(1)$. Linearity of μ_π^α then yields (40). □

The dual variable η is therefore a survival bonus, not an additive cost penalty. At the Bellman level, the Lagrangian critic uses $\tilde{r}_\eta(s,a)=\alpha(s,a)(r(s,a)+\eta)$ and $\tilde{\gamma}(s,a)=\gamma\alpha(s,a)$. This is the SDH Lagrange closure property.

F.4 Two interpretable continuation models

The previous results hold for any continuation model. We now record two important special cases constructed from a per-step violation magnitude $c(s,a)\geq 0$. Throughout the remainder of this appendix, we write

$$G_t:=\sum_{k=0}^{t}c(s_k,a_k)$$

for cumulative violation depth. We reserve C_t for the Bernoulli continuation indicator, see (30).

Definition F.11 (Hard continuation)**.** For a binary violation signal $c : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$, hard continuation is the model

$$\alpha_\infty(s, a) = \mathbf{1}\{c(s, a) = 0\}. \tag{41}$$

Proposition F.12 (Hard SDH is exact for no-violation chance constraints)**.** *Under hard continuation,* $B_t = \mathbf{1}\{G_t = 0\}$*. Therefore, for every fixed* t*,* $\Pr_\pi(B_t = 1) \geq 1 - \delta_t$ *is equivalent to* $\Pr_\pi(\exists k \leq t : c(s_k, a_k) = 1) \leq \delta_t$*. At the discounted random time,* $\Pr_\pi(B_{T_\gamma} = 1) = \Pr_\pi(G_{T_\gamma} = 0)$*.*

Proof. Under (41), survival at time k is deterministic given the trajectory and occurs exactly when $c(s_k, a_k) = 0$. Therefore $B_t = 1$ exactly when no violation occurs through time t. □

Definition F.13 (Exponential continuation)**.** For a nonnegative violation magnitude $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ and scale $\lambda > 0$, exponential continuation is

$$\alpha_\lambda(s, a) = \exp(-\lambda c(s, a)). \tag{42}$$

Proposition F.14 (Exponential continuation is cumulative-depth survival)**.** *Under exponential continuation,*

$$\Pr_\pi(B_{T_\gamma} = 1) = \mathbb{E}_\pi[e^{-\lambda G_{T_\gamma}}]. \tag{43}$$

Proof. Conditioned on the trajectory, $\mathbb{E}[B_t \mid \tau] = \prod_{k=0}^{t} \exp(-\lambda c(s_k, a_k)) = \exp(-\lambda G_t)$. Evaluating at T_γ and taking expectation proves the claim. □

Thus exponential SDH observes a Laplace transform of cumulative violation depth at the discounted random time. This statistic gives a one-sided chance certificate.

Proposition F.15 (Cumulative-depth chance diagnostic)**.** *Let* $\bar{S}_\pi(\lambda) := \mathbb{E}_\pi[e^{-\lambda G_{T_\gamma}}]$*. For every* $b > 0$*,*

$$\Pr_\pi(G_{T_\gamma} \geq b) \leq \frac{1 - \bar{S}_\pi(\lambda)}{1 - e^{-\lambda b}}. \tag{44}$$

Consequently, $\bar{S}_\pi(\lambda) \geq 1 - \delta(1 - e^{-\lambda b})$ *implies* $\Pr_\pi(G_{T_\gamma} \geq b) \leq \delta$*.*

Proof. Let $Y = 1 - e^{-\lambda G_{T_\gamma}}$. If $G_{T_\gamma} \geq b$, then $Y \geq 1 - e^{-\lambda b}$. Markov's inequality gives the result since $\mathbb{E}[Y] = 1 - \bar{S}_\pi(\lambda)$. □

The certificate is useful but limited: one exponential scale does not determine an arbitrary cumulative-cost profile. This is the first place where SDH separates from additive-budget CMDPs.

F.5 Why local SDH is not an additive-budget CMDP solver

The first is history dependence. While CMDPs constrain expected cumulative cost, exact budget accounting is Markov only after augmenting the state with the spent cost or remaining budget. Thus any local continuation rule that tries to represent pathwise budget feasibility must depend on residual-budget information, whereas $\alpha(s, a)$ is history-blind.

Theorem F.16 (History-aliasing obstruction)**.** *There exists a finite deterministic constrained-control problem with nonnegative costs and budget* $d > 0$ *such that no history-blind continuation model* $\alpha : \mathcal{S} \times \mathcal{A} \to [0, 1]$ *can agree with the exact residual-budget continuation rule*

$$\alpha^\star_d(h_t, a) := \mathbf{1}\{G_{t-1}(h_t) + c(s_t, a) \leq d\} \tag{45}$$

on all reachable histories.

Proof. Consider a deterministic two-stage problem. From s_0, action a_L has cost zero and action a_H has cost one; both lead to the same state x. At x, action a_{risk} has cost one. Set the budget to $d = 1$.

There are two histories reaching x: one with spent cost zero and one with spent cost one. The residual-budget rule allows a_{risk} after the first history and forbids it after the second. Both decisions correspond to the same state-action pair (x, a_{risk}), so no single history-blind value $\alpha(x, a_{\text{risk}})$ can match both. □

The second obstruction is functional. Additive cost and exponential survival probe different summaries of the same cumulative-depth profile.

Definition F.17 (Violation-depth profile)**.** For a fixed policy π, define

$$\Omega_\pi(b) := \mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t \mathbf{1}\{G_t \geq b\}\right], \qquad b \geq 0. \tag{46}$$

Proposition F.18 (Profile identities)**.** *For a policy π, define the expected discounted cumulative cost and the exponential survival statistic by*

$$J_c(\pi) := \mathbb{E}_\pi\Big[\sum_{t\geq 0}\gamma^t c(s_t,a_t)\Big], \qquad \mathfrak{S}_\pi(\lambda) := \mathbb{E}_\pi\Big[\sum_{t\geq 0}\gamma^t e^{-\lambda G_t}\Big].$$

Then

$$\frac{J_c(\pi)}{1-\gamma} = \int_0^\infty \Omega_\pi(b)\,db, \qquad \frac{1}{\lambda}\left(\frac{1}{1-\gamma} - \mathfrak{S}_\pi(\lambda)\right) = \int_0^\infty e^{-\lambda b}\Omega_\pi(b)\,db. \tag{47}$$

Proof. The first identity follows from the tail-integral representation $x = \int_0^\infty \mathbf{1}\{x \geq b\}\,db$ applied to G_t, together with $\sum_{t\geq 0}\gamma^t G_t = (1-\gamma)^{-1}\sum_{t\geq 0}\gamma^t c(s_t,a_t)$. The second follows similarly from $(1-e^{-\lambda x})/\lambda = \int_0^\infty e^{-\lambda b}\mathbf{1}\{x \geq b\}\,db$. □

Thus additive CMDP cost is the total area under the violation-depth profile, whereas exponential SDH observes a Laplace-weighted slice of the same profile.

Proposition F.19 (One exponential statistic does not control arbitrary additive cost)**.** *Fix $\gamma \in (0,1)$, $\lambda_0 > 0$, and $s \in (0,1)$. For every $K > 0$, there exists a finite MDP with nonnegative costs and a stationary randomized policy π such that*

$$\bar{S}_\pi(\lambda_0) := \mathbb{E}_\pi[e^{-\lambda_0 G_{T_\gamma}}] = s, \qquad J_c(\pi) > K. \tag{48}$$

Hence one exponential SDH statistic cannot in general determine or upper-bound expected discounted cumulative cost.

Proof. Consider an initial state s_0 with two actions. Action a_0 has cost zero and action a_M has cost $M > 0$; both transition to an absorbing zero-cost state. Let

$$\pi_M(a_M \mid s_0) = q_M := \frac{1-s}{1-e^{-\lambda_0 M}},$$

where $M \geq \lambda_0^{-1}\log(1/s)$ so that $q_M \in [0,1]$. Since all cost is incurred at time zero, $G_{T_\gamma} \in \{0, M\}$ with probabilities $1-q_M$ and q_M. Therefore

$$\bar{S}_{\pi_M}(\lambda_0) = 1 - q_M(1-e^{-\lambda_0 M}) = s.$$

On the other hand,

$$J_c(\pi_M) = q_M M = \frac{(1-s)M}{1-e^{-\lambda_0 M}} \geq (1-s)M.$$

Choosing $M > \max\{\lambda_0^{-1}\log(1/s), K/(1-s)\}$ gives $J_c(\pi_M) > K$ while $\bar{S}_{\pi_M}(\lambda_0) = s$. □

These results explain the scope of SDH. Local continuation cannot reproduce residual-budget semantics without augmenting the state, and one exponential scale cannot recover an arbitrary additive-cost profile.

F.6 A POSITIVE REGIME: SINGLE-SCALE SURVIVAL

The preceding obstruction is a worst-case statement: one exponential SDH scale cannot identify an arbitrary violation-depth profile. The obstruction disappears when the profile has only one effective depth parameter. In that case, the additive CMDP cost, the random-time chance tail, and the exponential SDH statistic are all different views of the same scalar.

Theorem F.20 (Single-scale recoverability). *Fix a policy π. Suppose that, for some $\tau_\pi > 0$,*

$$\Omega_\pi(b) = \frac{1}{1-\gamma} e^{-b/\tau_\pi}, \qquad b \geq 0. \tag{49}$$

Equivalently, $\Pr_\pi(G_{T_\gamma} \geq b) = e^{-b/\tau_\pi}$. Then, for every $\lambda > 0$,

$$J_c(\pi) = \tau_\pi, \qquad \mathfrak{S}_\pi(\lambda) = \frac{1}{(1-\gamma)(1+\lambda\tau_\pi)}, \qquad \bar{S}_\pi(\lambda) = \frac{1}{1+\lambda\tau_\pi}. \tag{50}$$

Consequently, for any fixed $\lambda_0 > 0$,

$$J_c(\pi) = \frac{1}{\lambda_0}\left(\frac{1}{(1-\gamma)\mathfrak{S}_\pi(\lambda_0)} - 1\right) = \frac{1}{\lambda_0}\left(\frac{1}{\bar{S}_\pi(\lambda_0)} - 1\right). \tag{51}$$

Moreover, for any $b > 0$ and $\delta \in (0,1)$, $\Pr_\pi(G_{T_\gamma} \geq b) \leq \delta$ is equivalent to

$$\bar{S}_\pi(\lambda) \geq \frac{1}{1 + \lambda b / \log(1/\delta)}. \tag{52}$$

Proof. The equivalence between (49) and the random-time tail follows from

$$\Pr_\pi(G_{T_\gamma} \geq b) = (1-\gamma)\Omega_\pi(b).$$

Using Prop. F.18,

$$\frac{J_c(\pi)}{1-\gamma} = \int_0^\infty \Omega_\pi(b)\, db = \frac{\tau_\pi}{1-\gamma},$$

so $J_c(\pi) = \tau_\pi$. The same profile identity gives

$$\frac{1}{\lambda}\left(\frac{1}{1-\gamma} - \mathfrak{S}_\pi(\lambda)\right) = \int_0^\infty e^{-\lambda b}\Omega_\pi(b)\, db = \frac{\tau_\pi}{(1-\gamma)(1+\lambda\tau_\pi)},$$

which rearranges to the expression for $\mathfrak{S}_\pi(\lambda)$. Since $\bar{S}_\pi(\lambda) = (1-\gamma)\mathfrak{S}_\pi(\lambda)$, this also gives the expression for $\bar{S}_\pi(\lambda)$ and its inversion in (51). Finally, $e^{-b/\tau_\pi} \leq \delta$ is equivalent to $\tau_\pi \leq b/\log(1/\delta)$; substituting this into $\bar{S}_\pi(\lambda) = (1+\lambda\tau_\pi)^{-1}$ yields (52). □

This theorem is the positive counterpart to Prop. F.19. For arbitrary profiles, one exponential slice is only one Laplace-weighted summary. For the single-scale family (49), the total area under the profile, the random-time chance tail, and the exponential SDH statistic are all controlled by the same depth parameter τ_π. Exponential SDH is therefore exact in this regime not because it solves additive-budget CMDPs in general, but because the task's violation geometry has collapsed to one effective scale.

F.7 Interpretation and scope

The appendix establishes the following formal picture. Standard survival chance constraints are horizon-wise. SDH replaces them by survival at the discounted random time T_γ. This is a weaker, geometrically averaged feasibility condition. The objective is also changed: reward is survival-gated at the same random time.

Because both reward and feasibility are represented by the same survival occupancy measure, the Lagrangian of the survival-gated random-time chance problem closes exactly as an SDH objective with shifted reward. This closure is the precise constrained-control interpretation of SDH.

The same analysis also clarifies scope. SDH is exact for hard no-violation chance constraints and for the survival-gated random-time formulation. It is not an exact additive-budget CMDP solver on the original state space, because budget feasibility is history-dependent and additive costs depend on the full violation-depth profile. Exponential SDH is therefore best understood as a single-scale survival model: limited for arbitrary profiles, but exactly informative when violation depth is governed by one effective scale.

G CONTROL AS INFERENCE WITH STOCHASTIC DECISION HORIZONS

The preceding survival-constrained formulation identifies the control problem for which SDH is exact. Its Lagrangian is

$$\mathcal{L}(\pi, \eta) = (1-\gamma)\mu_\alpha^\pi(r+\eta) - \eta\rho.$$

After division by $1-\gamma$, fixed-η policy optimization is SDH with reward shift $r \mapsto r+\eta$, plus the policy-independent constant $-\eta\rho/(1-\gamma)$. Thus η remains the Lagrange multiplier of the survival constraint; through the constant reward shift it acts as a uniform survival incentive in the transformed reward.

For the CaI construction we fix an arbitrary bounded reward and write it as r. All statements below extend pointwise to any fixed finite η by replacing r with $r+\eta$. Crucially, we apply inference to the same geometric evaluation time and survival event (Definition 2.1) used in the primal problem (2). The pathwise energy is the scaled endpoint reward

$$\widehat{R}_\gamma = \frac{B_{T_\gamma} r_{T_\gamma}}{1-\gamma}.$$

This is not a pathwise rewriting of a geometrically stopped prefix return: the two random variables differ. What is preserved exactly is the policy-level SDH reward functional. The endpoint construction has expectation $\mu_\alpha^\pi(r)$ and is the direct variational representation of the random-horizon reward in Definition 2.1.

The remaining CaI question is therefore entirely about policy information cost. A weighted log-ratio is legitimate only when it is the trajectory KL between actual probability laws. Under absorbing-state semantics, continuation failure ends the controlled decision process. Under virtual termination (objective-level reward termination without environment reset), the controller continues to act after reward credit has terminated. The two semantics share the endpoint optimality potential and differ only in which action likelihood factors remain active.

G.1 THE ELBO CONSTRAINT

We use the Gibbs variational principle in a form that keeps the trajectory measures explicit. Let X collect the random variables of the trajectory model, let p be the reference trajectory law, and let $q \ll p$ be a variational trajectory law. For the unnormalized optimality potential

$$\psi(X) = \exp\Big(\frac{R(X)}{\kappa}\Big), \qquad \kappa > 0,$$

define the evidence $Z := \mathbb{E}_p[\psi(X)]$. Assume $0 < Z < \infty$, $\mathbb{E}_q[|R(X)|] < \infty$, and $D_{\mathrm{KL}}(q\|p) < \infty$. The corresponding Gibbs posterior is the probability law

$$\frac{dp^\star}{dp}(X) = \frac{1}{Z}\exp\Big(\frac{R(X)}{\kappa}\Big).$$

Since its density with respect to p is strictly positive, $q \ll p$ implies $q \ll p^\star$. Expanding $D_{\mathrm{KL}}(q\|p^\star)$ then gives

$$D_{\mathrm{KL}}(q\|p^\star) = D_{\mathrm{KL}}(q\|p) - \frac{1}{\kappa}\mathbb{E}_q[R(X)] + \log Z,$$

or, equivalently,

$$\kappa \log Z = \mathbb{E}_q[R(X)] - \kappa D_{\mathrm{KL}}(q\|p) + \kappa D_{\mathrm{KL}}(q\|p^\star).$$

By nonnegativity of KL divergence,

$$\kappa \log Z \geq \mathbb{E}_q[R(X)] - \kappa D_{\mathrm{KL}}(q\|p).$$

Thus the CaI objective is the standard variational lower bound on the log evidence: reward favors trajectories of high optimality potential, while the trajectory KL measures their departure from the reference law.

G.2 Stochastic horizon, continuation, and endpoint reward

Consider an infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, with $\gamma \in (0,1)$, bounded reward, reference policy $\pi_0(a \mid s)$, and candidate policy $\pi(a \mid s)$. For $\tau = (s_0, a_0, s_1, a_1, \ldots)$, write $r_t := r(s_t, a_t)$ and $\alpha_t := \alpha(s_t, a_t)$.

Definition G.1 (Geometric evaluation horizon)**.** Let T_γ be independent of the MDP trajectory and continuation variables, with

$$\Pr(T_\gamma = t) = (1-\gamma)\gamma^t, \qquad t \in \{0, 1, \ldots\},$$

and define $H_t := \mathbf{1}\{T_\gamma \geq t\}$. Then $\mathbb{E}[H_t] = \gamma^t$.

Definition G.2 (Continuation process)**.** Given the model-defined continuation probability $\alpha : \mathcal{S} \times \mathcal{A} \to [0,1]$, let

$$C_t \mid s_t, a_t \sim \mathrm{Bernoulli}(\alpha(s_t, a_t)), \qquad B_t := \prod_{k=0}^{t} C_k.$$

Thus $B_t = 1$ iff the model-defined survival process continues through the state-action pair at time t.

Definition G.3 (Geometric-endpoint reward)**.** Define

$$\widehat{R}_\gamma := \frac{B_{T_\gamma} r_{T_\gamma}}{1-\gamma}, \qquad r_t := r(s_t, a_t).$$

The factor $(1-\gamma)^{-1}$ is a normalization convention chosen so that the expected reward term is exactly the unnormalized SDH return $\mu_\alpha^\pi(r)$ used by the Bellman equations. Omitting this factor would merely require the corresponding rescaling of the CaI temperature. With the present convention, Theorem 2.2 gives $\mathbb{E}_\pi[\widehat{R}_\gamma] = \mu_\alpha^\pi(r)$ under the ordinary π-controlled trajectory law.

Definition G.4 (Active decision gates)**.** The active decision indicators are

$$A_t^{\mathrm{AS}} := H_t \prod_{k=0}^{t-1} C_k, \qquad A_t^{\mathrm{VT}} := H_t,$$

with the empty product equal to one. Under AS semantics, no policy decision is active after the first continuation failure. Under VT semantics, policy decisions remain active until the geometric evaluation horizon, even after reward credit has terminated at the objective level.

Notably, there is a one-step offset between B_t and A_t^{AS}. The action a_t is selected before C_t is realized. Thus C_t enters B_t and determines whether an endpoint at time t or later receives reward credit, but it does not gate the likelihood factor for the already selected action a_t.

G.3 Semantic trajectory laws

Choose a fixed inactive action distribution $\lambda_\perp$ on $\mathcal{A}$, shared by the variational and prior laws. Define the semantic variational action kernel

$$K_{q,\sigma,t}(\cdot \mid s_t, A_t^\sigma) := \begin{cases} \pi(\cdot \mid s_t), & A_t^\sigma = 1, \\ \lambda_\perp(\cdot), & A_t^\sigma = 0, \end{cases}$$

and the semantic prior action kernel

$$K_{p,\sigma,t}(\cdot \mid s_t, A_t^\sigma) := \begin{cases} \pi_0(\cdot \mid s_t), & A_t^\sigma = 1, \\ \lambda_\perp(\cdot), & A_t^\sigma = 0. \end{cases}$$

The semantic laws q_σ and p_σ first sample the same T_γ and $s_0 \sim \nu$. For each $t = 0, \ldots, T_\gamma$, they sample action a_t from the corresponding kernel above and then sample continuation $C_t \sim \mathrm{Bernoulli}(\alpha(s_t, a_t))$; when $t < T_\gamma$, they also sample next-state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. After an AS continuation failure, both laws use the same inactive action kernel for the remaining bookkeeping variables. Those variables cannot affect the endpoint return because then $B_{T_\gamma} = 0$.

Thus the two laws agree on the geometric-horizon distribution, initial-state distribution, environment dynamics, continuation kernels, and inactive action kernel. They differ only at active decision times. Assume $\pi(\cdot \mid s) \ll \pi_0(\cdot \mid s)$ on active states and

$$\mathbb{E}_{q_\sigma}\Big[\sum_{t \geq 0} A_t^\sigma \Big| \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}\Big|\Big] < \infty.$$

Lemma G.5 (Semantic trajectory KL). *For either semantics $\sigma \in \{\mathrm{AS}, \mathrm{VT}\}$,*

$$D_{\mathrm{KL}}(q_\sigma \| p_\sigma) = \mathbb{E}_{q_\sigma}\Big[\sum_{t\geq 0} A_t^\sigma \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}\Big].$$

Proof. Condition on $T_\gamma = m$. The two laws have the same horizon mass, initial-state law, transition kernels, continuation kernels, and inactive action kernel. These factors cancel in the likelihood ratio. At each $t \leq m$, the action factor also cancels when $A_t^\sigma = 0$; when $A_t^\sigma = 1$, its ratio is $\pi(a_t \mid s_t)/\pi_0(a_t \mid s_t)$. Hence

$$\log \frac{q_\sigma}{p_\sigma} = \sum_{t=0}^{m} A_t^\sigma \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}.$$

Taking expectation over the shared geometric-horizon law gives the result. The stated absolute-integrability assumption justifies the expectation of the random finite sum. □

Crucially, the gated log-ratio is not an adhoc regularizer: it is exactly the trajectory KL induced by semantic laws that differ only at active decision times.

G.4 THE SDH ELBO

For either semantics define the endpoint optimality potential and evidence by

$$\psi_\sigma := \exp\Big\{\frac{\widehat{R}_\gamma}{\kappa}\Big\}, \qquad Z_\sigma := \mathbb{E}_{p_\sigma}[\psi_\sigma].$$

Theorem G.6 (Semantic endpoint SDH ELBO). *Assume $\|r\|_\infty \leq R_{\max} < \infty$. For either semantics $\sigma \in \{\mathrm{AS}, \mathrm{VT}\}$ and every $\kappa > 0$,*

$$0 < Z_\sigma \leq \exp\Big\{\frac{R_{\max}}{\kappa(1-\gamma)}\Big\} < \infty,$$

and

$$\kappa \log Z_\sigma \geq \mathcal{J}_\sigma(\pi, \pi_0),$$

where

$$\mathcal{J}_\sigma(\pi, \pi_0) := \mathbb{E}_{q_\sigma}\Big[\widehat{R}_\gamma - \kappa \sum_{t\geq 0} A_t^\sigma \log \frac{\pi(a_t \mid s_t)}{\pi_0(a_t \mid s_t)}\Big].$$

Proof. Since $B_{T_\gamma} \in \{0, 1\}$,

$$\widehat{R}_\gamma \leq \frac{\|r\|_\infty}{1-\gamma} \leq \frac{R_{\max}}{1-\gamma}$$

pathwise. Therefore $0 < \psi_\sigma \leq \exp\{R_{\max}/[\kappa(1-\gamma)]\}$, which proves the evidence bound. Applying the generic variational identity to $R = \widehat{R}_\gamma$ gives

$$\kappa \log Z_\sigma \geq \mathbb{E}_{q_\sigma}[\widehat{R}_\gamma] - \kappa D_{\mathrm{KL}}(q_\sigma \| p_\sigma).$$

Lemma G.5 yields the stated objective. □

For any fixed finite $\eta \geq 0$, replace r by $r + \eta$. Then

$$Z_{\sigma,\eta} \leq \exp\Big\{\frac{\|r\|_\infty + \eta}{\kappa(1-\gamma)}\Big\} < \infty.$$

After the endpoint reduction below, subtracting $\eta\rho/(1-\gamma)$ gives exactly the scaled fixed-η Lagrangian from Theorem 2.2, together with the semantic policy-KL cost.

G.5 REDUCING THE ENDPOINT AND DECISION GATES

Let $\mathbb{P}_\pi$ denote the joint law generated by the policy π, the environment dynamics, the independent geometric horizon T_γ, and continuation variables that are conditionally independent given the state-action trajectory τ, with $C_t \mid \tau \sim \mathrm{Bernoulli}\alpha(s_t, a_t))$. Expectations over its state-action marginal are denoted by $\mathbb{E}_{\tau\sim\pi}$.

Lemma G.7 (Endpoint and active-decision transfer)**.** *Let f_t be an integrable function of the trajectory prefix up to (s_t, a_t). For either semantics $\sigma \in \{\mathrm{AS}, \mathrm{VT}\}$,*

$$\mathbb{E}_{q_\sigma}[B_{T_\gamma} f_{T_\gamma}] = \mathbb{E}_{\mathbb{P}_\pi}[B_{T_\gamma} f_{T_\gamma}],$$

and, for every $t \geq 0$,

$$\mathbb{E}_{q_\sigma}[A_t^\sigma f_t] = \mathbb{E}_{\mathbb{P}_\pi}[A_t^\sigma f_t].$$

Proof. For the endpoint term, condition on $T_\gamma = t$. On the event $B_t = 1$, all continuation variables $C_0, \ldots, C_t$ equal one. Hence every decision needed to generate the prefix through a_t is active under AS, and all such decisions are active by definition under VT. The semantic law therefore uses π on the entire contributing prefix, exactly as $\mathbb{P}_\pi$ does. On $B_t = 0$, the contribution vanishes. Summing over t proves the first identity.

For the second identity, on $A_t^\sigma = 1$ every decision needed to generate the prefix through a_t is active under the corresponding semantics, so the contributing prefix law again agrees with $\mathbb{P}_\pi$. On $A_t^\sigma = 0$, the contribution is zero. □

Lemma G.8 (Endpoint reduction and active-decision weights)**.** *For either semantics $\sigma \in \{\mathrm{AS}, \mathrm{VT}\}$,*

$$\mathbb{E}_{q_\sigma}[\widehat{R}_\gamma] = \mu_\alpha^\pi(r) = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0} \gamma^t \Big(\prod_{k=0}^{t} \alpha_k\Big) r_t\Big].$$

Moreover, under $\mathbb{P}_\pi$, conditioned on the state-action trajectory τ,

$$\mathbb{E}_{\mathbb{P}_\pi}[A_t^{\mathrm{AS}} \mid \tau] = \gamma^t \prod_{k=0}^{t-1} \alpha_k, \qquad \mathbb{E}_{\mathbb{P}_\pi}[A_t^{\mathrm{VT}} \mid \tau] = \gamma^t.$$

Proof. By Lem. G.7 and the geometric-horizon identity in Theorem 2.2,

$$\mathbb{E}_{q_\sigma}[\widehat{R}_\gamma] = \frac{1}{1-\gamma} \mathbb{E}_{\mathbb{P}_\pi}[B_{T_\gamma} r_{T_\gamma}] = \mu_\alpha^\pi(r),$$

which is the displayed discounted expansion by the definition of μ_α^π.

For the decision gates, under $\mathbb{P}_\pi$ the geometric horizon T_γ is independent of the trajectory and continuation variables, with $\Pr(T_\gamma \geq t) = \gamma^t$. Conditional on the state-action trajectory τ, the continuation variables are independent and satisfy

$$\mathbb{E}_{\mathbb{P}_\pi}[C_k \mid \tau] = \alpha_k.$$

Therefore

$$\mathbb{E}_{\mathbb{P}_\pi}[A_t^{\mathrm{AS}} \mid \tau] = \gamma^t \prod_{k=0}^{t-1} \alpha_k, \qquad \mathbb{E}_{\mathbb{P}_\pi}[A_t^{\mathrm{VT}} \mid \tau] = \gamma^t.$$

□

G.6 AS/VT CLASSIFICATION THEOREM

Informally, AS and VT have the same survival-weighted task reward after marginalizing the geometric endpoint and continuation process. They differ only in the KL term, because they make different commitments about which post-continuation-failure actions remain semantic policy decisions.

Theorem G.9 (Exact AS and VT objectives)**.** *The semantic endpoint SDH ELBOs reduce to*

$$\mathcal{J}_{\mathrm{AS}}(\pi,\pi_0)=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t}\alpha_k\Big)r_t-\kappa\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t-1}\alpha_k\Big)\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)}\Big],$$

and

$$\mathcal{J}_{\mathrm{VT}}(\pi,\pi_0)=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t}\alpha_k\Big)r_t-\kappa\sum_{t\geq 0}\gamma^t\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)}\Big].$$

Proof. By Thm. G.6,

$$\mathcal{J}_\sigma(\pi,\pi_0)=\mathbb{E}_{q_\sigma}[\widehat{R}_\gamma]-\kappa\mathbb{E}_{q_\sigma}\Big[\sum_{t\geq 0}A_t^\sigma\,\ell_t^\pi\Big],\qquad \ell_t^\pi:=\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)}.$$

Lemma G.8 gives the common reward term. For the policy-KL term, Lemma G.7, the stated absolute integrability, and the tower property give

$$\mathbb{E}_{q_\sigma}\Big[\sum_{t\geq 0}A_t^\sigma\,\ell_t^\pi\Big]=\mathbb{E}_{\mathbb{P}_\pi}\Big[\sum_{t\geq 0}A_t^\sigma\,\ell_t^\pi\Big]=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\mathbb{E}_{\mathbb{P}_\pi}[A_t^\sigma\mid\tau]\,\ell_t^\pi\Big].$$

Substituting the two conditional weights from Lemma G.8 yields the AS and VT expressions. □

Remark G.10 (Pathwise inference model). The endpoint potential is not a pathwise rewriting of a stopped-prefix CaI potential: exponentiating $B_{T_\gamma}r_{T_\gamma}/(1-\gamma)$ generally gives a different Gibbs posterior from exponentiating a cumulative prefix reward. The exact invariant needed here is the reduced variational policy objective. The endpoint construction yields the same survival-weighted reward functional $\mu_\alpha^\pi(r)$ together with the same semantic trajectory KLs, and all downstream Bellman and algorithmic results depend only on this reduced objective.

G.7 Consequences for dynamic programming

The classification has a direct algorithmic consequence. AS preserves the standard soft Bellman structure under a variable discount. VT separates the survival-return critic from an ordinary discounted policy-divergence term.

For AS, define

$$u_t:=\gamma^t\prod_{k=0}^{t-1}\alpha_k.$$

Then Thm. G.9 gives

$$\mathcal{J}_{\mathrm{AS}}(\pi,\pi_0)=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}u_t\Big(\alpha_t r_t-\kappa\ell_t^\pi\Big)\Big].$$

Thus AS admits the soft Bellman recursion

$$Q_{\mathrm{AS}}^\pi(s,a)=\alpha(s,a)r(s,a)+\gamma\alpha(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V_{\mathrm{AS}}^\pi(s')\right],$$

with

$$V_{\mathrm{AS}}^\pi(s)=\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[Q_{\mathrm{AS}}^\pi(s,a)-\kappa\log\frac{\pi(a\mid s)}{\pi_0(a\mid s)}\Big].$$

This preserves the usual timing convention: the reward appears in the Q-backup, while the policy regularizer appears in the V-backup. The current model-defined continuation probability $\alpha(s,a)$ gates the current reward and future value, but it does not gate the policy KL cost of choosing the current action.

VT has a different structure. Its objective decomposes as

$$\mathcal{J}_{\mathrm{VT}}(\pi,\pi_0)=R_{\mathrm{surv}}^\pi-\kappa K^\pi,$$

where

$$R_{\mathrm{surv}}^\pi:=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t}\alpha_k\Big)r_t\Big]\quad\text{and}\quad K^\pi:=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\gamma^t\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)}\Big].$$

The survival reward component satisfies the variable-discount recursion

$$Q^{\pi}_{\mathrm{surv}}(s,a) = \alpha(s,a)r(s,a) + \gamma\alpha(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^{\pi}_{\mathrm{surv}}(s')\right],$$

with

$$V^{\pi}_{\mathrm{surv}}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}_{\mathrm{surv}}(s,a)\right].$$

The KL component instead follows the ordinary discounted recursion

$$K^{\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\log\frac{\pi(a\mid s)}{\pi_0(a\mid s)} + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}K^{\pi}(s')\Big].$$

Therefore VT is not a single standard soft Bellman recursion on the original MDP with one Q-function and one V-function. Its reward and KL components live under different continuation structures. This is precisely the algorithmic distinction used in the main text: AS yields a SAC-style variable-discount soft Bellman operator, whereas VT naturally separates a survival-return critic from an ordinary discounted policy-divergence constraint, matching the MPO-style variational picture.

G.8 Stability of the variable-discount critic

Both semantics use the same survival-return critic. The relevant evaluation operator has reward $\bar{r}(s,a) = \alpha(s,a)r(s,a)$ and discount $\bar{\gamma}(s,a) = \gamma\alpha(s,a)$. Since $\alpha(s,a) \in [0,1]$, the effective discount satisfies $0 \le \bar{\gamma}(s,a) \le \gamma < 1$.

Lemma G.11 (Variable-discount contraction)**.** *Let $\bar{r}$ be bounded and let $\bar{\gamma} : \mathcal{S}\times\mathcal{A} \to [0,\gamma]$. For a fixed policy π, define*

$$(\mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}V)(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\bar{r}(s,a) + \bar{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}V(s')\right].$$

Then $\mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}$ is a γ-contraction in the sup norm:

$$\left\|\mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}V - \mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}W\right\|_{\infty} \le \gamma\|V-W\|_{\infty}.$$

Consequently, it has a unique fixed point and value iteration converges to it from any bounded initialization.

Proof. For any bounded V, W and any s,

$$\begin{aligned}\left|(\mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}V)(s) - (\mathcal{T}^{\pi}_{\bar{r},\bar{\gamma}}W)(s)\right| &= \left|\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\bar{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\big(V(s')-W(s')\big)\right]\right| \\ &\le \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\bar{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}|V(s')-W(s')|\right] \\ &\le \gamma\|V-W\|_{\infty}.\end{aligned}$$

Taking the supremum over s gives the contraction bound. Existence, uniqueness, and convergence follow from the Banach fixed point theorem. □

Thus SDH shortens credit assignment locally without breaking the contraction structure of discounted policy evaluation.

G.9 Summary: ELBO-valid SDH regularization

The survival-constrained formulation identifies the random-horizon payoff optimized by SDH. This section applies the Gibbs variational construction directly to that same geometric endpoint, scaled by $(1-\gamma)^{-1}$ to match the Bellman-return normalization. Because the endpoint payoff is uniformly bounded for bounded rewards, the corresponding evidence is finite for every $\kappa > 0$.

The reward side and policy-KL side play different roles. The common endpoint potential uses $B_{T_\gamma} r_{T_\gamma}/(1-\gamma)$. The active decision gates

$$A^{\mathrm{AS}}_t = H_t\prod_{k=0}^{t-1}C_k, \qquad A^{\mathrm{VT}}_t = H_t$$

determine which policy likelihood factors appear in the semantic trajectory laws. The resulting weighted log-ratio terms are therefore genuine trajectory KLs, not pseudo-KL penalties.

After marginalization, AS and VT share exactly the same survival-weighted task objective and differ only in the policy-KL horizon induced by their post-continuation semantics. AS yields the survival-weighted KL-regularized Bellman recursion used by AS-SAC; VT pairs the same survival-weighted task critic with ordinary discounted, KL-constrained policy improvement in VT-MPO. The endpoint Gibbs posterior itself need not coincide with a stopped-prefix Gibbs posterior; the exact object preserved, and the only object used downstream, is the reduced regularized policy objective.

H ABSORBING STATE: UNIFORM PRIOR AND AS-SAC

This section derives the algorithmic consequences of the absorbing-state (AS) semantics when the CaI prior is a normalized uniform policy. The main point is simple but important: under AS, constants in the policy-ratio term are no longer globally constant, because they are accumulated only while the stochastic decision horizon survives. This yields a survival-weighted living cost and a SAC-style Bellman recursion under survival shaping, while temperature adaptation remains standard SAC-style policy-improvement tuning.

H.1 AS OBJECTIVE UNDER A NORMALIZED UNIFORM PRIOR

Recall the AS objective from Thm. G.9:

$$\mathcal{J}_{\mathrm{AS}}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t}\alpha_k\Big)r_t - \kappa\sum_{t\geq 0}\gamma^t\Big(\prod_{k=0}^{t-1}\alpha_k\Big)\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)}\right], \tag{53}$$

where $\alpha_t = \alpha(s_t, a_t)$ and $r_t = r(s_t, a_t)$. Assume throughout this section that the reference policy is normalized uniform:

$$\log\pi_0(a\mid s) = -\ell_c \qquad \forall(s,a). \tag{54}$$

For a finite action space, $\ell_c = \log|\mathcal{A}|$. In continuous control, ℓ_c should be interpreted relative to a chosen reference measure, and is implemented through a target-entropy constant.

Define the AS survival-to-decision weight

$$w_t(\tau) := \gamma^t\prod_{k=0}^{t-1}\alpha(s_k,a_k), \qquad w_0 = 1, \tag{55}$$

and the survival-shaped reward and discount

$$\tilde{r}(s,a) := \alpha(s,a)r(s,a), \qquad \tilde{\gamma}(s,a) := \gamma\alpha(s,a). \tag{56}$$

Corollary H.1 (CaI-AS with a normalized uniform prior)**.** *Under* (54)*, the AS objective can be written as*

$$\mathcal{J}_{\mathrm{AS}}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t\geq 0} w_t(\tau)\Big(\tilde{r}(s_t,a_t) + \kappa\,\mathcal{H}(\pi(\cdot\mid s_t)) - \kappa\,\ell_c\Big)\right]. \tag{57}$$

Equivalently, with

$$Z(\pi) := \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0} w_t(\tau)\Big], \tag{58}$$

we have the decomposition

$$\mathcal{J}_{\mathrm{AS}}(\pi) = \underbrace{\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0} w_t(\tau)\big(\tilde{r}_t + \kappa\mathcal{H}(\pi(\cdot\mid s_t))\big)\Big]}_{\textit{survival-shaped maximum-entropy return}} - \underbrace{\kappa\ell_c Z(\pi)}_{\textit{feasibility-weighted living cost}}. \tag{59}$$

Proof. Substituting $\log\pi_0(a\mid s) = -\ell_c$ into (53) gives

$$-\kappa\log\frac{\pi(a_t\mid s_t)}{\pi_0(a_t\mid s_t)} = -\kappa\log\pi(a_t\mid s_t) - \kappa\ell_c.$$

Moreover,

$$\gamma^t\prod_{k=0}^{t}\alpha_k r_t = w_t(\tau)\alpha_t r_t = w_t(\tau)\tilde{r}_t.$$

Taking the conditional expectation over $a_t\sim\pi(\cdot\mid s_t)$ turns $-\mathbb{E}[\log\pi(a_t\mid s_t)\mid s_t]$ into $\mathcal{H}(\pi(\cdot\mid s_t))$, yielding (57). The decomposition (59) follows by collecting the constant per-decision term into $Z(\pi)$. □

The corollary identifies the main difference from standard maximum-entropy RL. With ordinary geometric discounting, a constant per step contributes a fixed offset proportional to $(1-\gamma)^{-1}$. Under AS, the corresponding mass $Z(\pi)$ depends on the policy through $\alpha(s,a)$. Hence the living-cost term cannot generally be discarded without changing the optimization problem.

Remark H.2 (Finite action spaces). If $|\mathcal{A}| < \infty$ and $\pi_0(a \mid s) = 1/|\mathcal{A}|$, then $\ell_c = \log|\mathcal{A}|$ and

$$\kappa \ell_c Z(\pi) = \kappa \log(|\mathcal{A}|) \cdot \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t\geq 0}\gamma^t \prod_{k=0}^{t-1}\alpha(s_k,a_k)\Big].$$

Thus the normalized uniform prior induces a regularization on the expected feasibility-weighted decision horizon.

Counter-Example H.3 (Dropping the living cost can change the optimizer)**.** Consider the one-state MDP with $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a_{\text{cont}}, a_{\text{stop}}\}$. The transition is $P(s \mid s,a) = 1$ for both actions,

$$\alpha(s, a_{\text{cont}}) = 1, \qquad \alpha(s, a_{\text{stop}}) = 0,$$

and

$$r(s, a_{\text{cont}}) = r > 0, \qquad r(s, a_{\text{stop}}) = 0.$$

Let $\pi_p(a_{\text{cont}} \mid s) = p$ and $\pi_p(a_{\text{stop}} \mid s) = 1-p$, with $p \in (0,1)$. For a normalized uniform prior over the two actions, $\ell_c = \log 2$. Writing $h(p) := -p\log p - (1-p)\log(1-p)$, the exact AS objective is

$$\mathcal{J}_{\text{AS}}(\pi_p) = \frac{pr + \kappa h(p) - \kappa\log 2}{1-\gamma p}. \tag{60}$$

If the living-cost term is removed, the resulting objective becomes

$$\mathcal{J}_{\text{AS-N}}(\pi_p) = \frac{pr + \kappa h(p)}{1-\gamma p}. \tag{61}$$

These objectives can have different strictly stochastic maximizers. For example, with $(\gamma, \kappa, r) = (0.9, 1, 0.4)$,

$$\arg\max_{p\in(0,1)} \mathcal{J}_{\text{AS}}(\pi_p) \approx 0.707, \qquad \arg\max_{p\in(0,1)} \mathcal{J}_{\text{AS-N}}(\pi_p) \approx 0.984.$$

H.2 AS admits a SAC-style Bellman recursion

The preceding result explains why AS is not identical to standard SAC. Nevertheless, AS retains the essential algorithmic structure of SAC: policy evaluation is a soft Bellman recursion on the survival-shaped MDP.

Theorem H.4 (AS soft policy evaluation)**.** *Fix a policy π and a normalized uniform prior. Under AS semantics, the corresponding soft action-value function satisfies*

$$Q^\pi(s,a) := \tilde{r}(s,a) + \tilde{\gamma}(s,a)\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V^\pi(s')\big], \tag{62}$$

$$V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\big[Q^\pi(s,a) - \kappa(\log\pi(a \mid s) + \ell_c)\big]. \tag{63}$$

The associated Bellman operator is a γ-contraction whenever rewards are bounded.

Proof. Starting from (57), condition on an initial state-action pair (s,a). The immediate reward contribution is $\tilde{r}(s,a)$, the immediate prior-normalization contribution is $-\kappa\ell_c$, and future terms are discounted by $\tilde{\gamma}(s,a) = \gamma\alpha(s,a)$. The entropy term appears in the next state value through the standard soft value definition. Because $\tilde{\gamma}(s,a) \leq \gamma < 1$, contraction follows from Lem. G.11. □

Thus AS-SAC differs from SAC only through the shaped reward, the state-action-dependent discount, and the living-cost baseline.

For a replay transition $(s,a,r,s',d) \sim \mathcal{D}$, define

$$\alpha = \alpha(s,a), \qquad \tilde{r} = \alpha r, \qquad \tilde{\gamma} = (1-d)\gamma\alpha.$$

Sampling $a' \sim \pi_\phi(\cdot \mid s')$, the single-critic AS-SAC target is

$$\hat{V}(s') := \min_{i\in\{1,2\}} \bar{Q}_{\bar{\theta}_i}(s',a') - \kappa\big(\log\pi_\phi(a' \mid s') + \ell_c\big), \qquad y := \tilde{r} + \tilde{\gamma}\hat{V}(s'). \tag{64}$$

In continuous-action experiments, ℓ_c is implemented by a target-entropy baseline, so the soft value subtracts the current KL-to-prior surrogate rather than only the log-policy term. Setting this baseline to zero gives the ablation that drops the AS living cost.

H.2.1 ADAPTIVE TEMPERATURE AS A POLICY-IMPROVEMENT DUAL

The CaI derivation determines the AS objective for any fixed temperature κ; it does not require κ itself to be adapted through a trajectory-level survival-weighted information constraint. We instead adapt κ through a local policy-improvement constraint.

At each actor update, hold fixed the current regularized AS critic estimate

$$\widehat{Q}^{\mathrm{AS}}(s,a) := \min_i Q_{\phi_i}(s,a),$$

and let π' denote the candidate next policy. For a fixed state distribution d, taken here as the replay-state distribution $\mathcal{D}$, consider

$$\max_{\pi'} \ \mathbb{E}_{s\sim d,\ a\sim\pi'(\cdot|s)}\Big[\widehat{Q}^{\mathrm{AS}}(s,a)\Big] \quad \text{subject to} \quad \mathbb{E}_{s\sim d} D_{\mathrm{KL}}\Big(\pi'(\cdot\mid s)\,\|\,\pi_0(\cdot\mid s)\Big) \le \varepsilon. \tag{65}$$

Its policy-improvement Lagrangian is

$$\mathcal{L}_{\mathrm{PI}}(\pi',\kappa) = \mathbb{E}_{s\sim d}\Big[\mathbb{E}_{a\sim\pi'(\cdot|s)}\widehat{Q}^{\mathrm{AS}}(s,a) - \kappa D_{\mathrm{KL}}\Big(\pi'(\cdot\mid s)\,\|\,\pi_0(\cdot\mid s)\Big)\Big] + \kappa\varepsilon. \tag{66}$$

For fixed $\kappa > 0$, the nonparametric optimizer is the usual soft policy-improvement operator

$$\pi'_\kappa(a\mid s) \propto \pi_0(a\mid s)\exp\Big\{\frac{\widehat{Q}^{\mathrm{AS}}(s,a)}{\kappa}\Big\}. \tag{67}$$

Let $g(\kappa) := \max_{\pi'} \mathcal{L}_{\mathrm{PI}}(\pi',\kappa)$ denote the dual function. Under exact inner maximization, the envelope theorem gives

$$g'(\kappa) = \varepsilon - \mathbb{E}_{s\sim d} D_{\mathrm{KL}}\Big(\pi'_\kappa(\cdot\mid s)\,\|\,\pi_0(\cdot\mid s)\Big).$$

Projected dual descent therefore yields

$$\kappa \leftarrow \Big[\kappa + \beta_\kappa\Big(\mathbb{E}_{s\sim d} D_{\mathrm{KL}}\Big(\pi'_\kappa(\cdot\mid s)\,\|\,\pi_0(\cdot\mid s)\Big) - \varepsilon\Big)\Big]_+. \tag{68}$$

For a normalized uniform prior, $\log\pi_0(a\mid s) = -\ell_c$, and therefore

$$D_{\mathrm{KL}}(\pi'\|\pi_0) = \ell_c - \mathcal{H}(\pi').$$

Defining $\bar{\mathcal{H}} := \ell_c - \varepsilon$ gives

$$\kappa \leftarrow \Big[\kappa + \beta_\kappa\Big(\bar{\mathcal{H}} - \mathbb{E}_{s\sim d}\mathcal{H}(\pi(\cdot\mid s))\Big)\Big]_+, \tag{69}$$

which is the standard SAC-style temperature controller (21). In our off-policy realization we take $d = \mathcal{D}$, so the entropy target is imposed on the replay-state distribution used for actor improvement. With a parametric actor, the SAC actor update approximately solves the inner policy-improvement problem within the policy class, while the temperature update provides a stochastic dual step evaluated at the current actor. Under exact inner maximization, this reduces to the dual gradient above.

This local policy-improvement interpretation does not remove the AS living cost. For a normalized uniform prior,

$$-\kappa\log\frac{\pi'(a\mid s)}{\pi_0(a\mid s)} = -\kappa\log\pi'(a\mid s) - \kappa\ell_c.$$

The current $-\kappa\ell_c$ term is independent of the candidate action distribution and therefore does not affect the local actor argmax. Future occurrences of this term under the currently evaluated policy are already represented in the regularized AS critic $\widehat{Q}^{\mathrm{AS}}$ through the AS Bellman recursion. Thus the actor uses the usual SAC objective, while AS policy evaluation retains the CaI-derived survival-weighted living cost.

Importantly, the temperature update above is the dual of the local replay-state policy-improvement constraint (65), not of a trajectory-level survival-weighted information constraint. Its entropy statistic is therefore averaged under $d = \mathcal{D}$ without an additional survival weight. Survival semantics remain in policy evaluation through the regularized AS critic and its state-action-dependent discount.