ARRIVE AND SURVIVE: SCALING SAFE GOAL-CONDITIONED POLICY LEARNING USING ONE-BIT FAILURE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive reinforcement learning (CRL) scales effectively in goal-conditioned tasks through self-supervision. However, under failure termination, established CRL considers pre-failure future goals only when constructing positive samples, discarding the occupancy mass absorbed by failure. This omission induces a systematic overestimation bias in the implied goal-reaching values, giving short surviving futures disproportionately strong supervision. Such distortion can reinforce failure-prone behaviour through catastrophic failure bootstrapping or favour unsustainable goal reaching. To address this problem, we introduce two simple yet theoretically rigorous corrections: mass-weighted InfoNCE corrects the overweighting of short surviving futures in critic learning, and a log-survival-mass score restores the missing mass in policy optimization under the population formulation. The resulting method, Safe Contrastive Reinforcement Learning (Safe-CRL), uses the one-bit failure termination signal as its only safety supervision. Across twelve failure-prone robot navigation and locomotion tasks, Safe-CRL consistently improves mean survival time and achieves higher mean time at goal than Scaling-CRL, with substantial gains in nine tasks. It also preserves depth scalability, with deep policies exhibiting complex failure-avoidance behaviours. These results establish survival-mass correction as an effective completion of CRL for failure-prone goal-conditioned control.

1 INTRODUCTION

Scaling model capacity has become an important route toward more capable reinforcement learning (RL) policies (Schwarzer et al., 2023; Lee et al., 2025). In goal-conditioned RL, contrastive reinforcement learning (CRL) is particularly promising. It turns sparse goal-reaching reward into dense self-supervision through goal relabelling (Eysenbach et al., 2022; Andrychowicz et al., 2017), and learns to reach commanded goals via a contrastive objective. Recent work on Scaling-CRL has shown that this approach can effectively exploit deeper actor and critic networks to improve performance, establishing CRL as a scalable policy learning framework (Wang et al., 2025).

In practice, many goal-conditioned tasks involve *failure termination* – unsafe events like falls or collisions end the episode. For example, robot locomotion requires reaching the goal without falls (Berseth et al., 2020); navigation and autonomous driving require chasing moving goals while avoiding collisions (Belkhouche et al., 2006; Bansal et al., 2019). Therefore, a safe goal-conditioned policy must learn to *arrive and survive* while preserving model scalability, ideally using failure termination as its only additional safety supervision.

However, established CRL relabels pre-failure future goals for self-supervision, without considering the state occupancy absorbed by the failure (Eysenbach et al., 2022; Wang et al., 2025). Under such trajectory-wise relabelling, each trajectory's valid future-goal distribution is normalized to unit mass, discarding its total discounted pre-failure occupancy, which we term *survival mass*. This omission induces a systematic bias in the implied goal-reaching values in CRL and can have severe consequences. One is *catastrophic failure bootstrapping*: a self-reinforcing loop in which short, failure-prone trajectories dominate early replay and are repeatedly reinforced through self-supervision. Another one is *unsustainable goal reaching*: the agent reaches a goal but collides, overshoots, or

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.

falls shortly afterwards, as illustrated in Figure 1. Both behaviours can degrade performance in failure-prone tasks.

The discussion above raises the question: *How should CRL account for failure termination to realize scalable and safe goal-conditioned policy learning, using only a one-bit failure signal?*

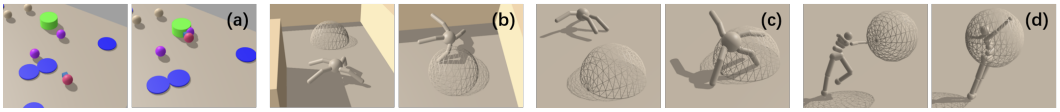


Figure 1: Examples of unsustainable goal reaching under Scaling-CRL. (a) The Point agent reaches the goal region, then collides with a moving obstacle; (b) the Ant reaches the goal but overshoots into the wall; (c–d) agents reach the goal but fall immediately afterwards.

In this paper, we first identify and formalize the systematic overestimation bias in established CRL under failure termination. To address this problem, we introduce *mass-weighted InfoNCE* and a *log-survival-mass score*, yielding *Safe-CRL* as a minimal, theoretically motivated correction of Scaling-CRL (Wang et al., 2025). Safe-CRL substantially improves survival and sustained goal occupancy in failure-prone tasks while preserving Scaling-CRL's depth scalability.

2 Problem Identification: The Missing Survival Mass in CRL

2.1 Preliminaries on CRL

We consider a goal-conditioned MDP with failure termination,

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \rho_0, \gamma, \mathcal{G}, \phi, F)\,, \tag{1}$$

where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces; $P(s' \mid s, a)$ is the state transition kernel, and ρ_0 is the initial-state distribution. $\gamma \in (0, 1)$ is the discount factor. The goal space is $\mathcal{G}$, and $\phi : \mathcal{S} \to \mathcal{G}$ maps a state to the achieved goal. A goal-conditioned policy $\pi(a \mid s, g)$ acts toward a commanded goal $g \in \mathcal{G}$. $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \{0, 1\}$ indicates whether a transition leads to a failure termination. We denote T_f as the random variable representing the failure horizon measured from (s_t, a_t).

CRL uses a contrastive critic and a goal-conditioned actor. It operates on *anchor-goal pairs*. The critic learns a contrastive score $f_\theta(x, g)$ between a state-action anchor $x = (s, a)$ and a goal g. Positive goals are obtained by hindsight future-goal relabelling (Eysenbach et al., 2022; Andrychowicz et al., 2017) that samples a horizon H from a discounted geometric distribution $q_\gamma(h) = (1-\gamma)\gamma^{h-1}$. The achieved goal $g_{t+H} = \phi(S_{t+H})$ is the positive goal for this anchor. Under failure termination, we analyse discounted future sampling normalized over each trajectory's valid pre-failure states. Goals associated with other anchors in a mini-batch are considered negatives. The InfoNCE loss trains the critic to distinguish positive and negative goals (van den Oord et al., 2018), which is interpreted as maximizing the likelihood of reachability in CRL (Wang et al., 2025).

At the population optimum, the contrastive score represents a positive-to-negative density ratio. Let $p^+(g \mid x)$ denote the positive goal distribution from anchor x and consider an anchor-independent negative proposal $p^-(g)$. The contrastive score estimates (Eysenbach et al., 2022)

$$f_\theta(x, g) \to \log \frac{p^+(g \mid x)}{p^-(g)} + c(x), \tag{2}$$

where $c(x)$ is an anchor-dependent offset that may depend on the action. Let $\bar{f}_\theta$ denote the calibrated score defined in Appendix A.3. At the population optimum, the action-independent negative proposal yields:

$$\arg\max_a \bar{f}_\theta(x, g) = \arg\max_a \log p^+(g \mid x). \tag{3}$$

2.2 The Overestimation Bias in CRL

Next, we analyse the bias induced by trajectory-normalized pre-failure relabelling. Throughout this analysis, π denotes the replay-induced continuation distribution, marginalized over behaviour goals when applicable. Occupancy and value are defined under this same distribution. The results concern the uncensored process under the conditions in Appendix A.5.

For an MDP with infinite horizon (no failure termination), *discounted state occupancy* is defined by

$$\mu_{\infty}^{\pi}(g \mid x) = \sum_{h=1}^{\infty} q_{\gamma}(h) p^{\pi}\left(\phi(s_{t+h}) = g \mid x_t = x\right), \tag{4}$$

Eysenbach et al. (2022) established the equivalence between the goal-reaching action-value function (Q-function) and this occupancy. Proposition 1 in Eysenbach et al. (2022) says

$$Q_g^{\pi}(s,a) = \mu_{\infty}^{\pi}(g \mid x). \tag{5}$$

Equation 4 forms a valid positive goal distribution with unit probability mass for equation 3. So, maximizing the contrastive score is equivalent to maximizing the goal-reaching Q-function.

Under failure termination, however, post-failure states are no longer valid future goals. In such settings, we propose:

Proposition 1 (Q-function is equivalent to failure-aware discounted future-goal occupancy) Let T_f denote the random failure horizon; we define

$$\mu_f^{\pi}(g \mid x) = \sum_{h=1}^{\infty} q_{\gamma}(h) p^{\pi}\left(\phi(s_{t+h}) = g,\ T_f > h \mid x_t = x\right), \tag{6}$$

For the normalized goal-reaching reward defined in Appendix A.1, the following equivalence holds:

$$Q_g^{\pi}(s,a) = \mu_f^{\pi}(g \mid x). \quad \square \tag{7}$$

The proof is provided in A.1. Unlike equation 4, μ_f^{π} is generally a sub-probability measure. Failure absorbs part of the discounted future occupancy. Its total mass is

$$Z^{\pi}(x) \triangleq \int_{\mathcal{G}} \mu_f^{\pi}(g \mid x)\, dg = \sum_{h=1}^{\infty} q_{\gamma}(h) \overset{\pi}{\Pr}\left(T_f > h \mid x\right) = \mathbb{E}_{H \sim q_{\gamma}}\left[\overset{\pi}{\Pr}\left(T_f > H \mid x\right)\right] \leq 1. \tag{8}$$

We call $Z^{\pi}(x)$ **survival mass**. Intuitively, $Z^{\pi}(x)$ is the probability that a future horizon sampled from q_{γ} still lies before failure.

However, trajectory-normalized pre-failure relabelling discards this mass. Let τ denote a future trajectory from anchor $x_t = x$, and let L_{τ} be its number of valid future states before failure. The realized failure-aware occupancy and survival mass are

$$\hat{\mu}_{\tau}(g \mid x) = \sum_{h=1}^{L_{\tau}} q_{\gamma}(h) \mathbf{1}\{\phi(s_{t+h}) = g\}, \quad \hat{Z}_{\tau} = \sum_{h=1}^{L_{\tau}} q_{\gamma}(h) = 1 - \gamma^{L_{\tau}}. \tag{9}$$

Therefore, the trajectory-level positive distribution is $\hat{\mu}_{\tau}(g \mid x)/\hat{Z}_{\tau}$. By construction, on average,

$$\mathbb{E}_{\tau|x}[\hat{\mu}_{\tau}(g \mid x)] = \mu_f^{\pi}(g \mid x), \quad \mathbb{E}_{\tau|x}[\hat{Z}_{\tau}] = Z^{\pi}(x). \tag{10}$$

The population positive-goal distribution is

$$p_{\text{CRL}}^{+}(g \mid x) = \mathbb{E}_{\tau|x}\left[\frac{\hat{\mu}_{\tau}(g \mid x)}{\hat{Z}_{\tau}}\right]. \tag{11}$$

Because $\hat{Z}_{\tau} \in (0,1]$, if we denote by $\hat{Q}_{\text{CRL},g}^{\pi}(s,a)$ the goal-reaching value implicitly learned by CRL, we have:

$$\boxed{\hat{Q}_{\text{CRL},g}^{\pi}(s,a) = p_{\text{CRL}}^{+}(g \mid x) = \mathbb{E}_{\tau|x}\left[\frac{\hat{\mu}_{\tau}(g \mid x)}{\hat{Z}_{\tau}}\right] \geq \mathbb{E}_{\tau|x}[\hat{\mu}_{\tau}(g \mid x)] = \mu_f^{\pi}(g \mid x) = Q_g^{\pi}(s,a).} \tag{12}$$

Equation 12 establishes a non-negative occupancy overestimation under trajectory-normalized pre-failure relabelling. When this distortion in relabelling changes action preferences, it can favour failure-prone behaviour.

The omission of survival mass impacts both the critic and actor learning of CRL. Figure 2 conceptualizes the two consequences. For critic learning, ignoring survival mass normalizes each trajectory,

giving short trajectories too much relative weight. Therefore, the learned positive distribution is distorted. Correcting this distortion is not sufficient. Contrastive learning can recover only the normalized average conditional occupancy $\bar{\mu}_f^\pi$, whereas the true goal-reaching value factorizes as

$$Q_g^\pi(s,a) = \mu_f^\pi(g \mid x) = Z^\pi(x)\bar{\mu}_f^\pi(g \mid x). \tag{13}$$

Omitting $Z^\pi(x)$ can change the action ranking induced by the failure-aware occupancy. We next derive complementary corrections for the trajectory-wise normalization and the missing total mass.

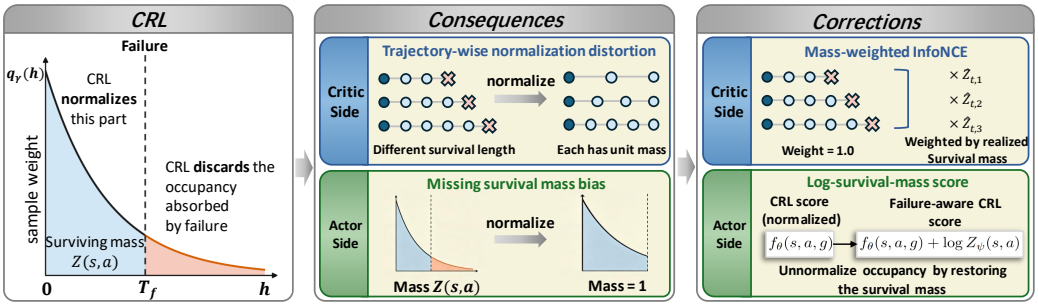


Figure 2: Failure termination removes part of the discounted future-goal occupancy. Established CRL normalizes the remaining pre-failure futures, whereas Safe-CRL corrects the trajectory-wise normalization distortion in critic learning and restores the missing total mass in actor scoring.

3 SAFE-CRL

3.1 CORRECTING OVERESTIMATION BIAS

We address the two effects of missing survival mass through complementary critic and actor corrections, as illustrated in Figure 2. For critic learning, we restore the relative contribution by weighting the InfoNCE row by the realized survival mass $\hat{Z}_i$. This gives the *mass-weighted InfoNCE* (MW-InfoNCE):

$$\mathcal{L}_{\text{MW-NCE}} = \frac{1}{B}\sum_{i=1}^{B} \hat{Z}_i \ell_i^{\text{NCE}}, \tag{14}$$

where B is the batch size and ℓ_i^{NCE} is the row-wise InfoNCE loss (van den Oord et al., 2018; Bortkiewicz et al., 2025) for each anchor. This gives the following proposition:

Proposition 2 (Mass-weighted InfoNCE corrects trajectory-wise normalization distortion) Equation 14 induces, at the population level, the conditional distribution associated with the failure-aware occupancy,

$$p_{\text{W}}^+(g \mid x) = \frac{\mu_f^\pi(g \mid x)}{Z^\pi(x)} = \bar{\mu}_f^\pi(g \mid x). \quad \square \tag{15}$$

The proof is provided in A.2. It says that multiplying each row by $\hat{Z}_i$ cancels the trajectory-wise normalization in expectation. On the actor side, the correction is direct:

Corollary 1 (Restoring survival mass in actor scoring) Under the calibrated CRL score convention in equation 3, the optimal mass-weighted contrastive score satisfies

$$\bar{f}_W^*(x,g) = \log \frac{\bar{\mu}_f^\pi(g \mid x)}{p^-(g)} = \log \mu_f^\pi(g \mid x) - \log Z^\pi(x) - \log p^-(g) \tag{16}$$

Because $\log p^-(g)$ is action-independent,

$$\arg\max_a \left[\bar{f}_W^*(x,g) + \log Z^\pi(x)\right] = \arg\max_a \mu_f^\pi(g \mid x) = \arg\max_a Q_g^\pi(s,a). \quad \square \tag{17}$$

The derivation is provided in Appendix A.3. At the population optimum, the calibrated mass-weighted score combined with the exact survival mass recovers the action ranking of the failure-aware occupancy, making the implicit goal-reaching value *unbiased*.

3.2 IMPLEMENTATION

We implement a lightweight Z-encoder $Z_\psi(x) \in (0, 1)$ to learn the survival mass. For an anchor x_t, its realized survival mass $\hat{Z}_t$ weights the InfoNCE row as in equation 14. In parallel, we sample $H_t \sim q_\gamma$ and train the Z-encoder with binary cross-entropy (BCE), assigning a target of 0 if failure is observed at or before H_t and 1 otherwise. Appendix A.5 describes truncation handling. For actor updates, the $\log Z$ term predicted by the Z-encoder is added to the objective:

$$f_{\text{corrected}}(s, a, g) = f_\theta(s, a, g) + \log Z_\psi(s, a). \tag{18}$$

We retain Scaling-CRL's treatment of the contrastive-score offset and introduce no additional calibration step. All other components, including the ResNet-like network architecture, remain unchanged (Wang et al., 2025). We call the resulting method **Safe Contrastive Reinforcement Learning (Safe-CRL)**.

Why not model failure as an explicit future outcome? In principle, formulating failure as an absorbing outcome would preserve the missing mass. However, this changes the established CRL outcome space: failure is shared across many anchors and therefore requires special treatment in contrastive negative sampling, while it is not a valid commanded goal for actor learning. Safe-CRL instead retains the original achieved-goal space and restores the missing survival mass directly.

4 EXPERIMENTS

Environments and Baselines We evaluate Safe-CRL on 12 failure-prone robot navigation and locomotion tasks built in Brax (Freeman et al., 2021), as illustrated in Figure 3. We use the MJX backend (Todorov et al., 2012; Zakka et al., 2025). For Humanoid, the actuator gear is weakened to make the locomotion task harder. Following the lidar-style observations used in Safety-Gymnasium (Ji et al., 2023), a 2D lidar is added to the observation to detect walls and obstacles. Appendix B provides detailed environment and failure specifications.

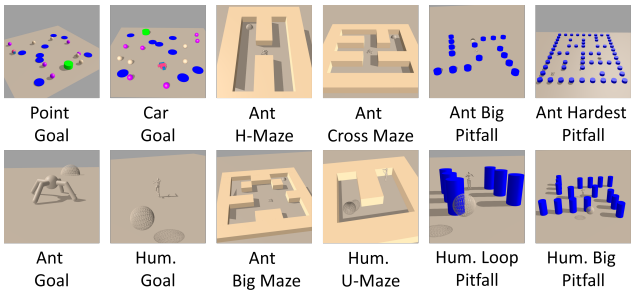


Figure 3: Illustrations of the experimental environments. The goal is represented by a bright green cylinder (for Point and Car Goal) or a meshed sphere (for Ant and Humanoid locomotion tasks). Entering task-specific failure regions (hazards, obstacles, or gremlins) or a fall triggers failure termination.

We use Scaling-CRL (Wang et al., 2025) as the primary baseline because Safe-CRL is an extension of it. An additional hindsight-relabelling baseline, SAC-HER (Haarnoja et al., 2018; Andrychowicz et al., 2017), is reported in Appendix C. For the main benchmark, both methods use 64-layer actor and critic networks, except on Point Goal and Car Goal, where both use 8 layers. Safe-CRL additionally uses a 4-layer Z-encoder, while all other training settings are identical to Scaling-CRL. We set $\gamma = 0.99$ and the episode time limit to 1000 steps in all environments.

Evaluation metrics We report three complementary metrics: (1) *time at goal*, the number of episode steps spent within the commanded goal region, jointly reflecting goal attainment and persistence near the goal (Wang et al., 2025; Bortkiewicz et al., 2025); (2) *survival time*, the number of steps before failure termination or the episode time limit; and (3) *goal coverage*, the percentage of episodes in which the commanded goal is reached by the robot at least once. Together, time at goal and goal coverage distinguish sustained goal occupancy from first-hit goal attainment; our theoretical objective concerns discounted occupancy under failure termination. Each metric is averaged over 128 parallel evaluation environments. We report the mean and standard deviation over five independent random seeds for the main benchmark and depth-scaling experiments, and over three seeds for the remaining experiments and ablations, following Scaling-CRL (Wang et al., 2025).

5 Results and Discussion

Main benchmark Figure 4 shows the training curve of time at goal, and Table 1 further compares the three metrics of the final policy. Figure D.1 presents the training curves of survival time and goal coverage. Overall, Safe-CRL achieves higher mean time at goal than Scaling-CRL in all 12 tasks. The gains are modest on Point Goal, Car Goal, and Ant H-Maze, but pronounced on the remaining nine tasks. For example, average time at goal increases from 276.8 to 684.0 steps on Ant Goal, from 22.2 to 266.4 on Humanoid Goal, and from 8.2 to 474.1 on Humanoid Big Pitfall. Safe-CRL also achieves consistently higher mean survival time in all 12 tasks, indicating a better safety performance.

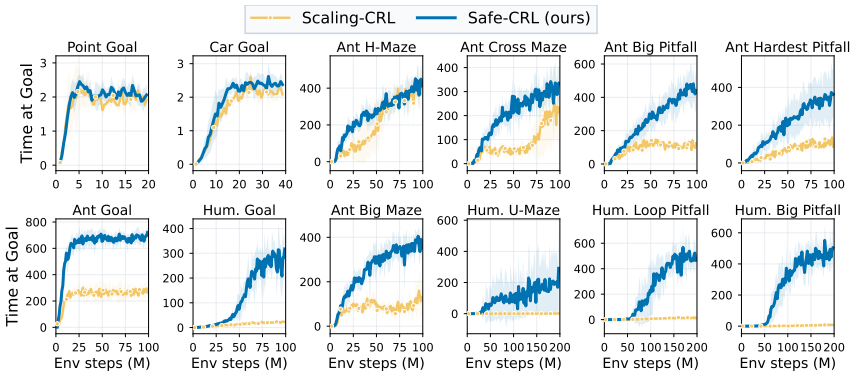


Figure 4: Comparison of time at goal between Scaling-CRL and the proposed Safe-CRL method in failure-prone navigation and locomotion tasks (mean ± standard deviation over five random seeds). Both methods use 64-layer actor and critic networks (8-layer for Point Goal and Car Goal). Safe-CRL achieves higher or on-par mean time at goal across all tasks, with pronounced gains in nine.

Table 1: Final performance on the main benchmark, mean ± standard deviation over 5 random seeds.

	Time at goal		Survival time		Goal coverage (%)	
Environment	Scaling-CRL	Safe-CRL	Scaling-CRL	Safe-CRL	Scaling-CRL	Safe-CRL
Point Goal	1.94 ± 0.11	$\mathbf{1.98 \pm 0.09}$	555.1 ± 12.3	$\mathbf{631.5 \pm 21.0}$	81.6 ± 3.4	$\mathbf{83.0 \pm 1.1}$
Car Goal	2.17 ± 0.04	$\mathbf{2.39 \pm 0.06}$	464.9 ± 6.5	$\mathbf{550.1 \pm 15.9}$	74.2 ± 1.9	$\mathbf{79.7 \pm 0.7}$
Ant Goal	276.83 ± 12.81	$\mathbf{684.00 \pm 14.57}$	422.1 ± 16.2	$\mathbf{910.8 \pm 8.0}$	$\mathbf{87.2 \pm 0.9}$	81.0 ± 1.6
Humanoid Goal	22.19 ± 2.64	$\mathbf{266.43 \pm 20.70}$	145.7 ± 12.9	$\mathbf{386.3 \pm 19.9}$	60.9 ± 0.7	$\mathbf{78.5 \pm 0.9}$
Ant H-Maze	390.60 ± 52.80	$\mathbf{412.96 \pm 17.92}$	586.6 ± 66.6	$\mathbf{795.6 \pm 19.6}$	$\mathbf{58.3 \pm 4.6}$	54.0 ± 2.5
Ant Cross Maze	215.55 ± 72.07	$\mathbf{314.46 \pm 27.89}$	425.6 ± 92.7	$\mathbf{785.2 \pm 5.6}$	$\mathbf{56.1 \pm 1.2}$	46.0 ± 3.4
Ant Big Maze	122.12 ± 36.23	$\mathbf{370.51 \pm 14.73}$	300.2 ± 29.0	$\mathbf{768.4 \pm 11.3}$	$\mathbf{62.4 \pm 2.5}$	51.8 ± 1.7
Humanoid U-Maze	0.56 ± 0.56	$\mathbf{200.29 \pm 82.31}$	130.4 ± 42.4	$\mathbf{373.9 \pm 102.0}$	5.6 ± 5.6	$\mathbf{31.9 \pm 10.3}$
Ant Big Pitfall	116.84 ± 5.23	$\mathbf{452.03 \pm 42.61}$	294.9 ± 9.4	$\mathbf{821.1 \pm 16.1}$	$\mathbf{72.9 \pm 5.1}$	62.7 ± 4.6
Ant Hardest Pitfall	108.54 ± 11.02	$\mathbf{348.07 \pm 56.75}$	491.0 ± 6.7	$\mathbf{846.9 \pm 15.7}$	41.8 ± 2.4	$\mathbf{51.3 \pm 8.8}$
Humanoid Loop Pitfall	14.18 ± 2.43	$\mathbf{492.95 \pm 19.91}$	120.5 ± 22.3	$\mathbf{637.9 \pm 18.7}$	53.7 ± 19.3	$\mathbf{68.0 \pm 1.7}$
Humanoid Big Pitfall	8.21 ± 3.69	$\mathbf{474.12 \pm 18.91}$	111.8 ± 23.0	$\mathbf{646.5 \pm 17.4}$	30.8 ± 14.1	$\mathbf{66.7 \pm 1.9}$

Comparing goal coverage provides further evidence for the two failure modes of CRL discussed in Section 1. On five of the six Ant tasks, Safe-CRL improves mean time at goal and survival time but reduces goal coverage moderately or slightly, indicating a trade-off between sustained goal occupancy and reaching a broader set of commanded goals. For the more failure-prone Humanoid environments, Scaling-CRL performs worse than Safe-CRL on all three metrics. Particularly, the survival time is significantly worse. This observation indicates that the policy is stuck in falling-prone gestures, which is *catastrophic failure bootstrapping*. Safe-CRL mitigates the two failure modes effectively: it improves sustained goal occupancy for the stabler Ant robot and alleviates persistent failure-prone behaviour for the unstable Humanoid robot.

Safe-CRL preserves depth scalability A key finding of Scaling-CRL is that goal-reaching performance improves with deeper networks (Wang et al., 2025). To examine whether Safe-CRL preserves this depth scalability, we fix the 4-layer Z-encoder and scale the actor and critic jointly from 4 to 64 layers, further extending to 256 layers for the more challenging Humanoid U-Maze and Humanoid Big Pitfall tasks. Figure 5 shows time at goal and survival time over five random seeds, with 64-layer Scaling-CRL included as a reference.

Figure 5 shows that time at goal and survival time generally improve as network depth increases. Their performance often saturates around 8–16 layers in the easier Ant tasks. Harder Humanoid tasks benefit from greater depth: the improvement on Humanoid Big Pitfall is pronounced up to approximately 64 layers, and Humanoid U-Maze continues to improve up to 256 layers. Notably, 8-layer Safe-CRL is competitive with or outperforms 64-layer Scaling-CRL across the evaluated tasks. These results show that Safe-CRL retains benefits from increased depth while improving performance at the same environment-interaction budget.

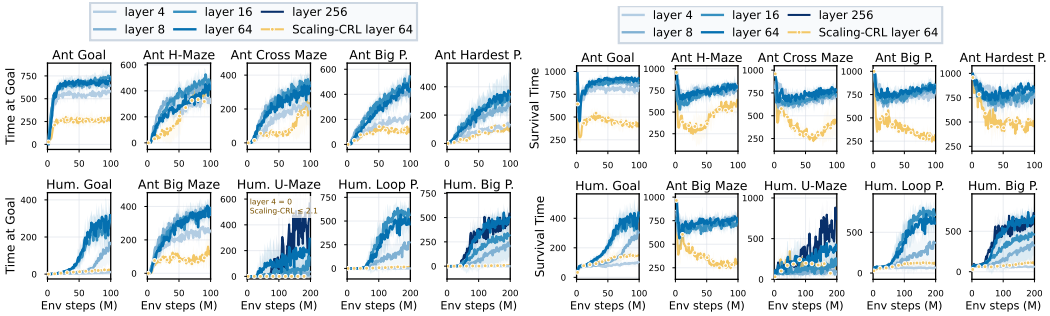


Figure 5: Time at goal and survival time for Safe-CRL using different actor and critic depths. The Z-encoder depth is 4. Point Goal and Car Goal are omitted from the depth-scaling study because they are easier than locomotion.

Z-encoder predictions track survival trends with little overhead Figure 6 compares predicted survival mass with evaluated survival time during training on the four Humanoid tasks. The quantities generally increase together, suggesting that Z-encoder predictions track changes in survival performance. This provides a qualitative consistency check rather than a direct assessment of survival-mass calibration.

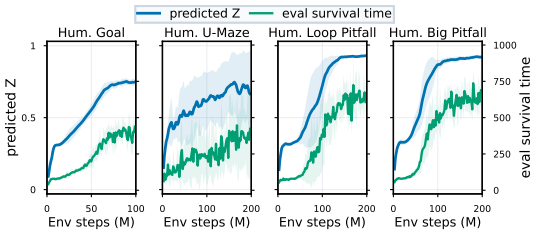


Figure 6: Predicted survival mass versus survival time in 128 evaluation runs. Actor and critic depths are 64.

The Z-encoder is computationally lightweight. Using the default 4-layer architecture increases the total number of model parameters by only approximately 4.2% (64-layer actor and critic networks) and the average wall-clock training time by 2.8% (Table D.1). Thus, the corrections introduce only modest computational overhead.

Deep Safe-CRL policies exhibit complex failure-avoidance behaviours We finally visualize representative rollouts of the learned policies (8 layers for Point Goal and Car Goal, 64 layers for the others) in Figure 7. Despite receiving only a one-bit failure signal, Safe-CRL learns diverse failure-avoidance behaviours while reaching the commanded goal. On Point Goal, the robot brakes and replans an S-shaped trajectory to pass between moving gremlins after the goal respawns. On Humanoid Goal, the robot walks backward toward a goal behind it and recovers from near-falls and overshooting after reaching the goal. On Humanoid Loop Pitfall, the humanoid adjusts its body orientation to avoid nearby hazard regions, while on the more difficult pitfall tasks, the humanoid and ant exhibit longer-horizon behaviours such as taking detours or slowing down to pass safely between hazards. These rollouts illustrate non-trivial failure-avoidance and recovery behaviours learned from a sparse one-bit failure signal.

Additional experimental results are provided in Appendix D.

6 Ablations

In this section, we mainly ablate the two corrections in Safe-CRL and examine the sensitivity to the coefficient of the $\log Z$ term. Additional ablations on Z-encoder depth, TD-based survival-mass estimation, and the effect of goal respawning are provided in Appendix D.

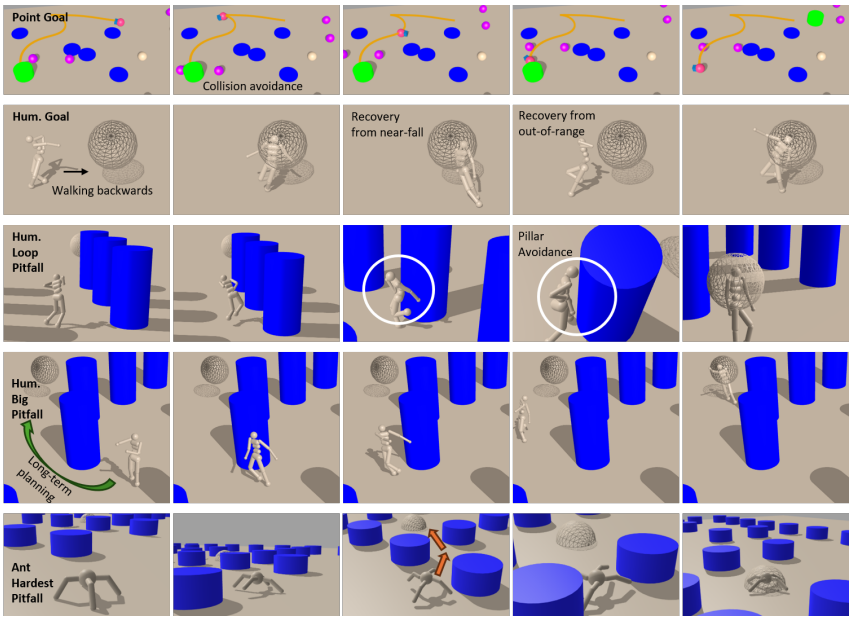


Figure 7: Visualization of the policy learned by Safe-CRL. Each row contains 5 snapshots captured from one evaluation episode, showing the key failure-avoidance behaviours of the agent.

Core component ablations Table 2 isolates the contributions of mass-weighted InfoNCE and the log-survival-mass correction. MW-InfoNCE alone produces only modest changes relative to Scaling-CRL, whereas the log-Z correction accounts for the dominant improvement across the four tasks. For example, on Humanoid Goal, the log-Z-only variant increases time at goal from 13.8 to 288.7 steps and survival time from 113.9 to 397.0 steps, while mass-weighted InfoNCE alone achieves 28.3 and 146.1 steps, respectively. A similar pattern is observed on Ant Big Maze and Ant Hardest Pitfall. MW-InfoNCE provides smaller additional improvements in the reported means, consistent with its complementary role in correcting trajectory-wise normalization rather than restoring the missing mass in actor scoring.

Table 2: Core component ablation of Safe-CRL. MW denotes MW-InfoNCE. Results are averaged over the final 10% of training and reported as mean $\pm$ standard deviation over three random seeds. The actor and critic depths are 16 (8 for Car Goal), and the Z-encoder depth is 4.

			Car Goal		Humanoid Goal		Ant Big Maze		Ant Hardest Pitfall	
Method	MW	log Z	Time at goal	Survival time	Time at goal	Survival time	Time at goal	Survival time	Time at goal	Survival time
Scaling-CRL	✗	✗	2.2±0.1	464.9±11.2	13.8±1.0	113.9±5.3	89.7±34.2	261.1±33.1	98.5±17.1	381.1±11.6
MW only	✓	✗	2.2±0.1	471.4±13.3	28.3±0.7	146.1±10.2	91.4±31.1	291.6±51.6	113.2±19.3	411.0±27.1
log Z only	✗	✓	**2.4**±0.1	537.1±61.3	288.7±21.6	397.0±21.8	337.0±28.9	693.5±33.7	271.2±28.3	722.9±32.6
Safe-CRL	✓	✓	**2.4**±0.2	**555.5**±49.1	**297.0**±19.4	**413.4**±23.9	**352.9**±11.1	**725.7**±31.7	**313.0**±50.7	**787.9**±51.1

Ablations on $\log Z$'s weight The population factorization of failure-aware occupancy gives a coefficient of 1.0 for $\log Z$ under the calibrated-score convention. To examine deviations from this value, we introduce an analysis coefficient β and compare 0.2, 1.0, and 5.0 in Figure 8. Increasing β generally increases survival time, whereas its effect on goal-reaching performance is environment-dependent. Safe-CRL fixes $\beta = 1$ because it is theory-derived rather than tuned as a safety–goal trade-off parameter; the experiment does not assume that this value maximizes every empirical metric in every environment.

7 Conclusion

This paper studied contrastive reinforcement learning under failure termination and identified a systematic occupancy overestimation bias induced by trajectory-normalized pre-failure relabelling. The analysis of this bias led to two minimal corrections, mass-weighted InfoNCE for critic learning and a log-survival-mass correction for policy optimization, which together form Safe-CRL. Across failure-prone navigation and locomotion tasks, Safe-CRL improves survival and sustained goal

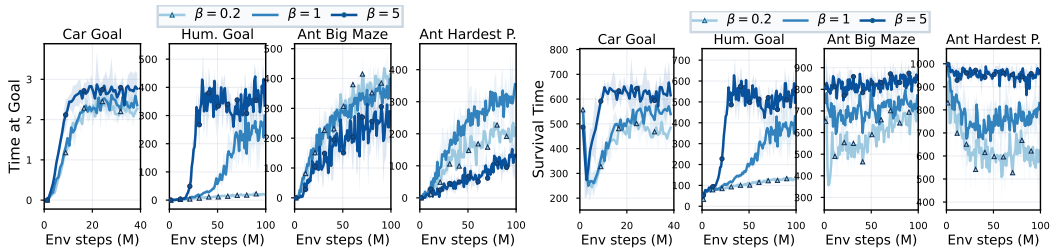


Figure 8: Ablations on the coefficient β of $\log Z$. The left four panels report time at goal, and the right four panels report survival time. We set the actor and critic depths to 16 layers (8 layers for Car Goal) and vary β.

occupancy while preserving CRL's depth scalability, using the one-bit failure signal as its only additional safety supervision. More broadly, our results suggest that under failure termination, scalable self-supervised policy learning should preserve not only the relative structure of observed futures, but also the total occupancy mass removed by termination.

Limitations We point out three limitations here: First, Safe-CRL is designed for failure-terminated goal-conditioned control and does not directly address non-terminating safety costs or constrained-RL objectives. Second, truncation introduces censoring approximations: contrastive weighting uses observed future support, whereas Z-encoder targets treat horizons without observed failure as survived. Third, Safe-CRL retains the exploration limitations of CRL, which can result in uneven goal-space coverage in difficult tasks (see an analysis in Figure D.5, Appendix D). Addressing these settings is also left for future work.

8 Related Work

Finally, we briefly review the related work. Goal-conditioned RL commonly uses achieved future states for hindsight relabelling, converting sparse goal-reaching experience into dense supervision (Andrychowicz et al., 2017). C-learning formulates goal achievement by classifying future versus independently sampled states (Eysenbach et al., 2021), while contrastive reinforcement learning (CRL) connects contrastive density-ratio estimation to discounted future-state occupancy and goal-conditioned value learning (Eysenbach et al., 2022). More recently, Scaling-CRL showed that this self-supervised objective benefits substantially from deeper networks, establishing CRL as a scalable goal-conditioned policy learning framework (Wang et al., 2025).

Several studies have examined biases or missing information in goal-conditioned learning from observed futures. Distributional Distance Classifiers separate horizon-dependent goal-reaching quantities from reachability probability (Akella et al., 2023), while USHER corrects hindsight-selection bias under stochastic dynamics (Schramm et al., 2023). Survival Value Learning and Survival Reinforcement Learning instead formulate goal-conditioned learning through time-to-event distributions and survival analysis (Tiofack et al., 2026; Nguimatsia-Tiofack et al., 2026). Recent work has also studied safety-aware exploration arising from CRL representations (Bastankhah et al., 2026). These works address horizon dependence, stochastic hindsight bias, temporal goal-reaching structure, or exploration. Safe-CRL studies a distinct failure-termination effect: trajectory-wise normalization of pre-failure futures discards survival mass and distorts the implied failure-aware occupancy.

A separate line of conventional safe RL formulates safety through constrained Markov decision processes (CMDPs), where a policy maximizes task return subject to explicit cumulative-cost constraints (Altman, 1999). Representative approaches include Constrained Policy Optimization (Achiam et al., 2017), Lyapunov-based safe RL (Chow et al., 2018), and Lagrangian-PID methods (Stooke et al., 2020). These methods assume explicit cost signals and do not require unsafe events to terminate the episode.

In summary, existing work has studied scalable CRL, stochastic and horizon-dependent goal reachability, and survival-based goal-conditioned learning. To our knowledge, prior work has not identified the systematic overestimation bias that arises under trajectory-normalized pre-failure relabelling in CRL, nor the accompanying trajectory-wise normalization distortion in critic learning.

REPRODUCIBILITY STATEMENT

Detailed theoretical assumptions, proofs, environment and failure specifications, baseline implementation details, additional experimental results, and ablations are provided in the appendices. We will publicly release the source code, environment implementations, training configurations, and evaluation scripts to ensure the reproducibility of the reported results.

AI USE STATEMENT

Generative AI tools were used for language editing, literature search, and feedback on the presentation of theoretical assumptions and interpretation of experimental results. The authors developed the research ideas, problem formulation, methodology, theoretical analysis, experiments, and conclusions. Particularly, the identification of the overestimation bias originated with the authors. All AI-assisted text and suggested references were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

Ravi Tej Akella, Benjamin Eysenbach, Jeff Schneider, and Ruslan Salakhutdinov. Distributional distance classifiers for goal-conditioned reinforcement learning. In *ICML 2023 Workshops: Frontiers4LCD*, 2023. URL https://openreview.net/forum?id=yuLAGjRwk9.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Boca Raton, FL, 1999. ISBN 9780849303821.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/453fadbd8a1a3af50a9df4df899537b5-Abstract.html.

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *Robotics: Science and Systems*, 2019.

Mahsa Bastankhah, Grace Liu, Dilip Arumugam, Thomas L. Griffiths, and Benjamin Eysenbach. Demystifying the mechanisms behind emergent exploration in goal-conditioned rl. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mwgYORsqtv.

Fethi Belkhouche, Boumediene Belkhouche, and Parviz Rastgoufard. Line of sight robot navigation toward a moving goal. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(2):255–267, 2006.

Glen Berseth, Daniel Geng, Coline Devin, Dinesh Jayaraman, Chelsea Finn, and Sergey Levine. SMiRL: Surprise minimizing RL in entropic environments. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lDbaVYvH.

Michał Bortkiewicz, Władysław Pałucki, Vivek Myers, Tadeusz Dziarmaga, Tomasz Arczewski, Łukasz Kuciński, and Benjamin Eysenbach. Accelerating goal-conditioned reinforcement learning algorithms and research. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2858e880333b3cd64f8192f13ddcca2f-Abstract-Conference.html.

Yinlam Chow, Ofir Nachum, Edgar A. Duéñez-Guzmán, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8103–8112, 2018.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ovpk6kH1scl.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Ruslan Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2580. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e7663e974c4ee7a2b475a4775201ce1f-Abstract-Conference.html.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax: A differentiable physics engine for large scale rigid body simulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d1f491a404d6854880943e5c3cd9ca25-Abstract-round1.html.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

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*, volume 36, pp. 18964–18993, 2023. doi: 10.52202/075280-0831. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3c557a3d6a48cc99444f85e924c66753-Abstract-Datasets_and_Benchmarks.html.

Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R. Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jXLiDKsuDo.

Franki Nguimatsia-Tiofack, Fabian Schramm, Théotime Le Hellard, and Justin Carpentier. Survival reinforcement learning: Toward scalable self-supervised RL. *arXiv preprint arXiv:2605.31273*, 2026. URL https://arxiv.org/abs/2605.31273.

Liam Schramm, Yunfu Deng, Edgar Granados, and Abdeslam Boularias. USHER: Unbiased sampling for hindsight experience replay. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2073–2082. PMLR, 2023. URL https://proceedings.mlr.press/v205/schramm23a.html.

Max Schwarzer, Johan Samir Obando Ceron, Aaron Courville, Marc G Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level atari with human-level efficiency. In *International Conference on Machine Learning*, pp. 30365–30380. PMLR, 2023.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 2020.

Franki Nguimatsia Tiofack, Fabian Schramm, Théotime Le Hellard, and Justin Carpentier. SVL: Goal-conditioned reinforcement learning as survival learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

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*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzciński, and Benjamin Eysenbach. 1000 layer networks for self-supervised RL: Scaling depth can enable new goal-reaching capabilities. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5268. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e74ee34cc0f2d0780f34ee77d8fba25b-Abstract-Conference.html.

Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A. Kahrs, Carmelo Sferrazza, Yuval Tassa, and Pieter Abbeel. MuJoCo Playground. *arXiv preprint arXiv:2502.08844*, 2025. URL https://arxiv.org/abs/2502.08844.

A Proofs and Technical Details

This appendix provides the detailed derivations for the theoretical results in Sections 2 and 3. We first show that the failure-aware discounted future-goal occupancy is the correct goal-conditioned value, then prove that survival-mass weighting in the InfoNCE loss corrects the trajectory-wise normalization distortion induced by future relabelling. Finally, we derive the log-survival-mass correction for actor optimization.

A.1 Proof of Proposition 1

We prove Proposition 1 using measurable goal sets, which avoids point-mass notation for continuous goal spaces. For an anchor $x = (s, a)$ and a measurable goal set $B \subseteq \mathcal{G}$, define the failure horizon as

$$T_f \triangleq \inf \{h \geq 1 : F(s_{t+h-1}, a_{t+h-1}, s_{t+h}) = 1\}, \tag{A.1}$$

with $T_f = \infty$ if no failure occurs. Therefore, $T_f > h$ means that S_{t+h} lies on the valid pre-failure portion of the trajectory. The corresponding failure-aware discounted future-goal occupancy is

$$\mu_f^\pi(B \mid x) = \sum_{h=1}^{\infty} q_\gamma(h) \overset{\pi}{\Pr}(\phi(S_{t+h}) \in B, T_f > h \mid x_t = x), \quad q_\gamma(h) = (1-\gamma)\gamma^{h-1}. \tag{A.2}$$

Consider the normalized goal-reaching transition reward

$$r_B(s, a, s') = (1-\gamma)\mathbf{1}\{F(s, a, s') = 0, \phi(s') \in B\}. \tag{A.3}$$

The corresponding discounted action value is

$$Q_B^\pi(s, a) = \mathbb{E}^\pi \left[\sum_{k=0}^{\infty} \gamma^k r_B(s_{t+k}, a_{t+k}, s_{t+k+1}) \mid s_t = s, a_t = a \right]. \tag{A.4}$$

Because failure immediately terminates the episode, a valid goal-reaching reward at horizon $k+1$ is obtained only when $T_f > k+1$. Substituting the reward definition gives

$$Q_B^\pi(s, a) = \sum_{k=0}^{\infty} (1-\gamma)\gamma^k \overset{\pi}{\Pr}(\phi(s_{t+k+1}) \in B, T_f > k+1 \mid x_t = x). \tag{A.5}$$

Letting $h = k+1$ directly yields

$$Q_B^\pi(s, a) = \sum_{h=1}^{\infty} q_\gamma(h) \overset{\pi}{\Pr}(\phi(S_{t+h}) \in B, T_f > h \mid x_t = x) = \mu_f^\pi(B \mid x). \tag{A.6}$$

This proves Proposition 1. For discrete goals, or equivalently using density notation for continuous goal spaces, this gives the pointwise relation used in the main text,

$$Q_g^\pi(s,a) = \mu_f^\pi(g \mid x). \tag{A.7}$$

Taking $B = \mathcal{G}$ further gives

$$Z^\pi(x) = \mu_f^\pi(\mathcal{G} \mid x) = \sum_{h=1}^{\infty} q_\gamma(h) \overset{\pi}{\Pr}(T_f > h \mid x), \tag{A.8}$$

which is exactly the defined survival mass.

A.2 Proof of Proposition 2

Consider an anchor $x_t = x = (s,a)$ and a realized trajectory τ from this anchor. Let L_τ denote the number of valid future states before failure. For a measurable goal set $B \subseteq \mathcal{G}$, define the realized failure-aware discounted future-goal measure as

$$\hat{\mu}_\tau(B \mid x) = \sum_{h=1}^{L_\tau} q_\gamma(h) \mathbf{1}\{\phi(S_{t+h}) \in B\}. \tag{A.9}$$

Its total mass is

$$\hat{Z}_\tau = \hat{\mu}_\tau(\mathcal{G} \mid x) = \sum_{h=1}^{L_\tau} q_\gamma(h) = 1 - \gamma^{L_\tau}. \tag{A.10}$$

Conditioned on the realized trajectory τ, established CRL samples positive goals only from these valid pre-failure futures. Therefore, for $\hat{Z}_\tau > 0$, the corresponding normalized positive-goal distribution is

$$r_\tau(B \mid x) = \frac{\hat{\mu}_\tau(B \mid x)}{\hat{Z}_\tau}. \tag{A.11}$$

Without mass weighting, averaging this distribution over trajectories gives the established CRL positive distribution,

$$p^+_{\mathrm{CRL}}(B \mid x) = \mathbb{E}_{\tau|x}\left[r_\tau(B \mid x)\right] = \mathbb{E}_{\tau|x}\left[\frac{\hat{\mu}_\tau(B \mid x)}{\hat{Z}_\tau}\right]. \tag{A.12}$$

In general,

$$\mathbb{E}_{\tau|x}\left[\frac{\hat{\mu}_\tau(B \mid x)}{\hat{Z}_\tau}\right] \neq \frac{\mathbb{E}_{\tau|x}\left[\hat{\mu}_\tau(B \mid x)\right]}{\mathbb{E}_{\tau|x}\left[\hat{Z}_\tau\right]}. \tag{A.13}$$

Thus, trajectory-wise normalization does not generally recover the normalized population failure-aware occupancy. We now consider the mass-weighted InfoNCE objective. Multiplying the complete contrastive row associated with trajectory τ by $\hat{Z}_\tau$ induces the weighted positive measure

$$\nu_W(B \mid x) \triangleq \mathbb{E}_{\tau|x}\left[\hat{Z}_\tau r_\tau(B \mid x)\right]. \tag{A.14}$$

Using the definition of r_τ, the trajectory-wise normalization cancels exactly:

$$\hat{Z}_\tau r_\tau(B \mid x) = \hat{\mu}_\tau(B \mid x). \tag{A.15}$$

Therefore,

$$\nu_W(B \mid x) = \mathbb{E}_{\tau|x}\left[\hat{\mu}_\tau(B \mid x)\right] = \mu_f^\pi(B \mid x). \tag{A.16}$$

Taking $B = \mathcal{G}$ gives the total weighted mass

$$\nu_W(\mathcal{G} \mid x) = \mathbb{E}_{\tau|x}\left[\hat{Z}_\tau\right] = Z^\pi(x). \tag{A.17}$$

So, the normalized positive distribution induced by mass-weighted InfoNCE is

$$p^+_W(B \mid x) = \frac{\nu_W(B \mid x)}{\nu_W(\mathcal{G} \mid x)} = \frac{\mu_f^\pi(B \mid x)}{Z^\pi(x)}. \tag{A.18}$$

Equivalently, using the density notation adopted in the main text,

$$p_W^+(g \mid x) = \frac{\mu_f^\pi(g \mid x)}{Z^\pi(x)} = \bar{\mu}_f^\pi(g \mid x). \tag{A.19}$$

This proves Proposition 2. Mass weighting therefore corrects the trajectory-wise normalization distortion and recovers the correct conditional shape of the failure-aware occupancy. However,

$$\int_{\mathcal{G}} p_W^+(g \mid x)\, dg = 1, \tag{A.20}$$

so the contrastive distribution remains normalized, and the total survival mass $Z^\pi(x)$ is still absent from the critic score. This remaining factor is restored in actor optimization by the log-survival-mass correction derived in Appendix A.3.

For anchors with $Z^\pi(x) = 0$, the normalized distribution above is undefined, but such anchors receive zero weight in the mass-weighted population objective; this case is also discussed separately in Appendix A.5.

Finite-minibatch estimator For a mini-batch of B independently sampled anchor-trajectory pairs, the implemented critic loss is

$$\widehat{\mathcal{L}}_{\text{MW-NCE}} = \frac{1}{B} \sum_{i=1}^{B} \hat{Z}_i \ell_i^{\text{NCE}}, \tag{A.21}$$

where ℓ_i^{NCE} denotes the complete row-wise InfoNCE loss for anchor i. Its expectation satisfies

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{\text{MW-NCE}}\right] = \mathbb{E}\left[\hat{Z}\ell^{\text{NCE}}\right]. \tag{A.22}$$

Thus, the unnormalized mini-batch average is a direct Monte Carlo estimator of the population mass-weighted contrastive risk. No normalization of the weights within each mini-batch is required.

In particular, the self-normalized alternative

$$\frac{\sum_{i=1}^{B} \hat{Z}_i \ell_i^{\text{NCE}}}{\sum_{i=1}^{B} \hat{Z}_i} \tag{A.23}$$

contains a random denominator and is not the same finite-minibatch estimator as the objective used in Safe-CRL.

The negative goals appearing in each InfoNCE row remain sampled from the ordinary mini-batch goal marginal, as in established CRL. Row weighting changes the contribution of the anchor-positive pair to the population risk but does not require the negative proposal to be re-weighted. The resulting negative proposal therefore remains anchor-independent, as required by the density-ratio interpretation used in Appendix A.3.

A.3 Proof of Corollary 1

Let $p^-(g)$ denote the anchor-independent negative proposal. Under the support condition stated in Appendix A.5, the population optimum of InfoNCE identifies the positive-to-negative density ratio up to an anchor-dependent additive offset:

$$f_W^*(x, g) = \log \frac{p_W^+(g \mid x)}{p^-(g)} + c(x). \tag{A.24}$$

Because $x = (s, a)$, the offset $c(x)$ may in general depend on the action. Therefore, the row-wise InfoNCE objective alone does not make scores directly comparable across different actions. To remove this arbitrary offset, define the calibrated contrastive score

$$\bar{f}_W(x, g) = f_W(x, g) - \log \mathbb{E}_{g' \sim p^-} \left[\exp f_W(x, g')\right]. \tag{A.25}$$

This normalization defines the calibrated score used in the analysis; it is not an additional implementation step in Safe-CRL, which retains Scaling-CRL's offset treatment. At the population optimum,

$$\mathbb{E}_{g' \sim p^-} \left[\exp f_W^*(x, g')\right] = \mathbb{E}_{g' \sim p^-} \left[\exp\left(\log \frac{p_W^+(g' \mid x)}{p^-(g')} + c(x)\right)\right]. \tag{A.26}$$

Therefore,

$$\mathbb{E}_{g'\sim p^-}\left[\exp f^*_W(x,g')\right] = e^{c(x)}\int_{\mathcal{G}} p^-(g')\frac{p^+_W(g'\mid x)}{p^-(g')}\,dg'. \tag{A.27}$$

Because $p^+_W(\cdot\mid x)$ is a normalized distribution,

$$\int_{\mathcal{G}} p^+_W(g'\mid x)\,dg' = 1, \tag{A.28}$$

and therefore

$$\mathbb{E}_{g'\sim p^-}\left[\exp f^*_W(x,g')\right] = e^{c(x)}. \tag{A.29}$$

Substituting this result into the calibrated score gives

$$\bar{f}^*_W(x,g) = \log\frac{p^+_W(g\mid x)}{p^-(g)}. \tag{A.30}$$

Using Proposition 2, we have:

$$p^+_W(g\mid x) = \frac{\mu^\pi_f(g\mid x)}{Z^\pi(x)}, \tag{A.31}$$

so that

$$\bar{f}^*_W(x,g) = \log\frac{\mu^\pi_f(g\mid x)}{Z^\pi(x)p^-(g)}. \tag{A.32}$$

Equivalently,

$$\bar{f}^*_W(x,g) = \log\mu^\pi_f(g\mid x) - \log Z^\pi(x) - \log p^-(g). \tag{A.33}$$

Adding the log survival mass therefore yields

$$\bar{f}^*_W(x,g) + \log Z^\pi(x) = \log\mu^\pi_f(g\mid x) - \log p^-(g). \tag{A.34}$$

For a fixed state s and commanded goal g, the negative proposal $p^-(g)$ is independent of the candidate action a. So, we have:

$$\arg\max_a\left[\bar{f}^*_W(s,a,g) + \log Z^\pi(x)\right] = \arg\max_a \log\mu^\pi_f(g\mid x). \tag{A.35}$$

Because the logarithm is strictly increasing,

$$\arg\max_a\left[\bar{f}^*_W(s,a,g) + \log Z^\pi(x)\right] = \arg\max_a \mu^\pi_f(g\mid x). \tag{A.36}$$

Finally, Proposition 1 gives

$$\mu^\pi_f(g\mid x) = Q^\pi_g(s,a), \tag{A.37}$$

and

$$\arg\max_a\left[\bar{f}^*_W(s,a,g) + \log Z^\pi(x)\right] = \arg\max_a Q^\pi_g(s,a). \tag{A.38}$$

This proves Corollary 1. The coefficient of the log-Z correction is exactly 1.0 because the failure-aware occupancy factorizes multiplicatively as

$$\mu^\pi_f(g\mid x) = Z^\pi(x)p^+_W(g\mid x). \tag{A.39}$$

Thus, the unit coefficient 1.0 follows from this population factorization; its empirical effect is examined in the coefficient ablation.

A.4 Population target of the Z-encoder

The Z-encoder is trained from a one-sample Monte Carlo survival target. For a trajectory τ, we sample $H\sim q_\gamma$ and define $Y = \mathbf{1}\{T_f > H\}$. For the uncensored process,

$$\mathbb{E}_H[Y\mid\tau] = \sum_{h=1}^{L_\tau} q_\gamma(h) = \hat{Z}_\tau, \qquad \mathbb{E}_{\tau,H}[Y\mid x] = \mathbb{E}_\tau[\hat{Z}_\tau\mid x] = Z^\pi(x). \tag{A.40}$$

For a fixed anchor x, the corresponding binary cross-entropy risk is

$$\mathcal{L}_x(z) = -\mathbb{E}\left[Y \log z + (1-Y)\log(1-z) \mid x\right], \qquad z \in (0,1). \tag{A.41}$$

For $0 < Z^\pi(x) < 1$, its population minimizer is

$$Z^*_\psi(x) = \mathbb{E}[Y \mid x] = Z^\pi(x). \tag{A.42}$$

For $Z^\pi(x) \in \{0,1\}$, the infimum is approached as z tends to the corresponding boundary.

Because binary cross-entropy is affine in its target label, sampling one Y provides an unbiased Monte Carlo estimator of the soft-target BCE based on $\hat{Z}_\tau$. Under time-limit truncation, horizons without an observed failure are treated as survived, yielding the finite-support approximation discussed in Appendix A.5.

A.5 Technical conditions

The theoretical results above rely on the following conditions.

(1) Replay-induced trajectory distribution The notation π denotes the replay-induced trajectory distribution represented by the data used for contrastive learning. When replay contains trajectories collected under different commanded goals, π is the corresponding goal-marginalized replay-induced trajectory distribution; the same replay-induced trajectory distribution must be used consistently in the definitions of μ^π_f and Z^π.

(2) Support of the negative proposal For the contrastive density ratio to be well-defined, the negative proposal satisfies

$$p^-(g) > 0 \qquad \text{whenever} \qquad p^+_{\mathrm{W}}(g|x) > 0. \tag{A.43}$$

This is the standard support condition for contrastive density-ratio estimation.

(3) Continuous goal spaces For continuous goal spaces, expressions such as $\mu^\pi_f(g|x)$ and $p^+_{\mathrm{W}}(g|x)$ denote densities with respect to a common dominating measure. Equivalently, all results can be stated directly for measurable goal sets without assuming the existence of densities.

(4) Zero-mass trajectories If a trajectory contains no valid future state for an anchor, then

$$L_\tau = 0, \qquad \hat{Z}_\tau = 0. \tag{A.44}$$

Such an anchor has zero weight in the mass-weighted contrastive objective and therefore does not require a valid positive goal in the population formulation. In implementation, rows with $L_\tau = 0$ are masked from the InfoNCE objective before positive-goal sampling.

(5) Failure termination versus right censoring The theoretical quantities μ^π_f and Z^π are defined for the uncensored failure-terminated process. In implementation, following Scaling-CRL (Wang et al., 2025; Bortkiewicz et al., 2025), futures beyond the available replay support or time limit are marked as invalid, and we do not apply an explicit right-censoring correction. When this support ends because of time-limit truncation rather than failure, the trajectory is right-censored in survival-analysis terminology (Tiofack et al., 2026). Consequently, $\hat{Z}_t$ should be interpreted operationally as the realized survival mass of the future support available to the sampler. Importantly, the sample-wise identity underlying Proposition 2 remains exact on this observed support,

$$\hat{Z}_\tau r_\tau(g \mid x) = \hat{\mu}_\tau(g \mid x). \tag{A.45}$$

Therefore, mass weighting still exactly cancels the trajectory-wise normalization introduced by the implemented future sampler. For Z-encoder training, we instead sample $H \sim q_\gamma$ and use $Y = \mathbf{1}\{T_f > H\}$. Horizons extending beyond the available trajectory without an observed failure are treated as valid survival samples with $Y = 1$; horizons crossing an observed failure remain valid with $Y = 0$. Without censoring, $\mathbb{E}[Y \mid x] = Z^\pi(x)$. Under time-limit truncation, treating horizons without an observed failure as survived yields a finite-support approximation that can overestimate the uncensored $Z^\pi(x)$. We do not apply an explicit censoring correction. Consequently, exact cancellation on the observed support does not imply exact recovery of the uncensored occupancy by the combined implemented score.

B ENVIRONMENTS

We evaluate Safe-CRL on 12 failure-prone goal-conditioned robot control tasks implemented in Brax (Freeman et al., 2021) with the MJX backend (Todorov et al., 2012; Zakka et al., 2025), providing richer friction and contact models. In all environments, failure produces a one-bit failure signal and immediately terminates the episode. Time-limit truncation is not treated as failure.

Point Goal and Car Goal are adopted from the level-2 Safety-Gymnasium navigation tasks (Ji et al., 2023), using the same environmental parameters. Entering an obstacle, hazard (pitfall), or moving-gremlin region causes failure, and the goal respawns after being reached. The remaining Ant and Humanoid tasks are adopted from Scaling-CRL (Wang et al., 2025). Falling is failure in Goal and Maze tasks, while Pitfall tasks additionally terminate when the robot enters a hazard region. Locomotion goals do not respawn after being reached.

Navigation tasks Point Goal and Car Goal are adopted from the level-2 Safety-Gymnasium navigation tasks (Ji et al., 2023). Goals respawn after being reached, and entering a designated obstacle, hazard, or moving-gremlin region triggers failure termination. The number of obstacles, hazards, gremlins, and their sizes, layout generation rules, playground sizes, 3-channel 16-bin 2D lidar observations, and ego-robot states are all identical to the original Safety-Gymnasium package.

Locomotion tasks The Ant and Humanoid tasks are adapted from Scaling-CRL (Wang et al., 2025) and JaxGCRL (Bortkiewicz et al., 2025). Goals remain fixed after being reached. Falling terminates all locomotion tasks, while Pitfall variants additionally terminate when the robot enters a designated hazard region. Maze-wall contact itself is not treated as failure. For locomotion tasks, we also equip the agents with a 16-bin, one-channel 2D lidar to detect surrounding walls and pitfalls.

Table B.1 summarizes the task-specific termination rules. To make the locomotion tasks sufficiently failure-prone, we reduce the ground sliding friction coefficient from 1.0 to 0.6, creating a slippery ground. In particular, for the Humanoid robot, actuator gears of the torso, lower limbs, and arms are weakened, as listed in Table B.2. These modified dynamics are shared by both the Scaling-CRL baseline and the proposed Safe-CRL.

Table B.1: Summary of the experimental environments and failure mechanisms.

Environment	Agent	Goal Respawn	Failure Termination
Point Goal	Point	Yes	Obstacle, hazard, or moving gremlin
Car Goal	Car	Yes	Obstacle, hazard, or moving gremlin
Ant Goal	Ant	No	Fall
Humanoid Goal	Humanoid	No	Fall
Ant H-Maze	Ant	No	Fall
Ant Cross Maze	Ant	No	Fall
Ant Big Maze	Ant	No	Fall
Humanoid U-Maze	Humanoid	No	Fall
Ant Big Pitfall	Ant	No	Fall or enter hazard region
Ant Hardest Pitfall	Ant	No	Fall or enter hazard region
Humanoid Loop Pitfall	Humanoid	No	Fall or enter hazard region
Humanoid Big Pitfall	Humanoid	No	Fall or enter hazard region

C SAC-HER BASELINE

We additionally compare Safe-CRL with SAC-HER, combining Soft Actor-Critic (Haarnoja et al., 2018) with Hindsight Experience Replay (Andrychowicz et al., 2017). The implementation is based on JaxGCRL (Bortkiewicz et al., 2025). Scaling-CRL has previously shown that contrastive self-supervision substantially outperforms conventional HER-based goal-conditioned RL in large-scale policy learning (Wang et al., 2025). Here, SAC-HER serves only as an additional reference for conventional hindsight-based goal-conditioned RL in the same failure-terminated setting. This comparison is not intended to test the CRL-specific missing-survival-mass mechanism identified in Section 2. As shown in Figure C.1, SAC-HER performs substantially worse than Safe-CRL on the four representative failure-prone tasks, particularly on the more challenging locomotion environments.

Table B.2: Humanoid actuator gears under the default setting and in our experiments.

Actuator gear	default	ours
abdomen_y	350	100
abdomen_z	350	100
abdomen_x	350	100
right_hip_x	350	100
right_hip_z	350	100
right_hip_y	350	300
right_knee	350	200
left_hip_x	350	100
left_hip_z	350	100
left_hip_y	350	300
left_knee	350	200
right_shoulder1	100	25
right_shoulder2	100	25
right_elbow	100	25
left_shoulder1	100	25
left_shoulder2	100	25
left_elbow	100	25

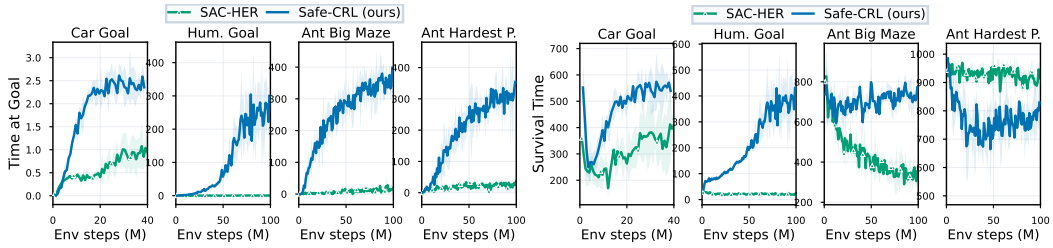


Figure C.1: Comparison with SAC-HER on four representative failure-prone tasks. Curves report time at goal and survival time. Safe-CRL uses 16 layers (8 layers for Car Goal) for the actor and critic networks and 4 layers for the Z-encoder.

D Supplementary results and ablations

This appendix provides additional quantitative results and analyses that complement the main experiments.

Additional training curves for the main benchmark Figure D.1 shows the training curves of survival time and goal coverage for the main benchmark.

Effect of goal respawning Point Goal and Car Goal respawn a new commanded goal after the previous goal is reached, so the policy must continue acting and can still encounter failure afterwards. We compare this default setting with a fixed-goal variant in Figure D.2. Safe-CRL retains an advantage in both settings, but the improvement in survival time is larger when goals respawn: $+13.8\%$ versus $+5.0\%$ for Point Goal and $+18.3\%$ versus $+2.2\%$ for Car Goal. This result is consistent with the survival-mass interpretation: when control continues after reaching a goal, failure removes additional future occupancy and therefore has a larger effect on subsequent goal-conditioned behaviour.

Z-encoder depth We compare 4-layer and 64-layer Z-encoders on four representative tasks while keeping the actor and critic depths fixed. Increasing the Z-encoder depth provides no consistent improvement in either time at goal or survival time, suggesting that a lightweight estimator is sufficient in these experiments.

TD-based survival-mass estimation Safe-CRL trains the Z-encoder from $Y_t = \mathbf{1}\{T_f > H_t\}$ with $H_t \sim q_\gamma$. We additionally examine whether $Z(s, a)$ can instead be learned through temporal-difference (TD) learning. We replace the Z-encoder objective with a clipped double-Q formulation (Fujimoto et al., 2018) that estimates a goal-independent survival-mass function.

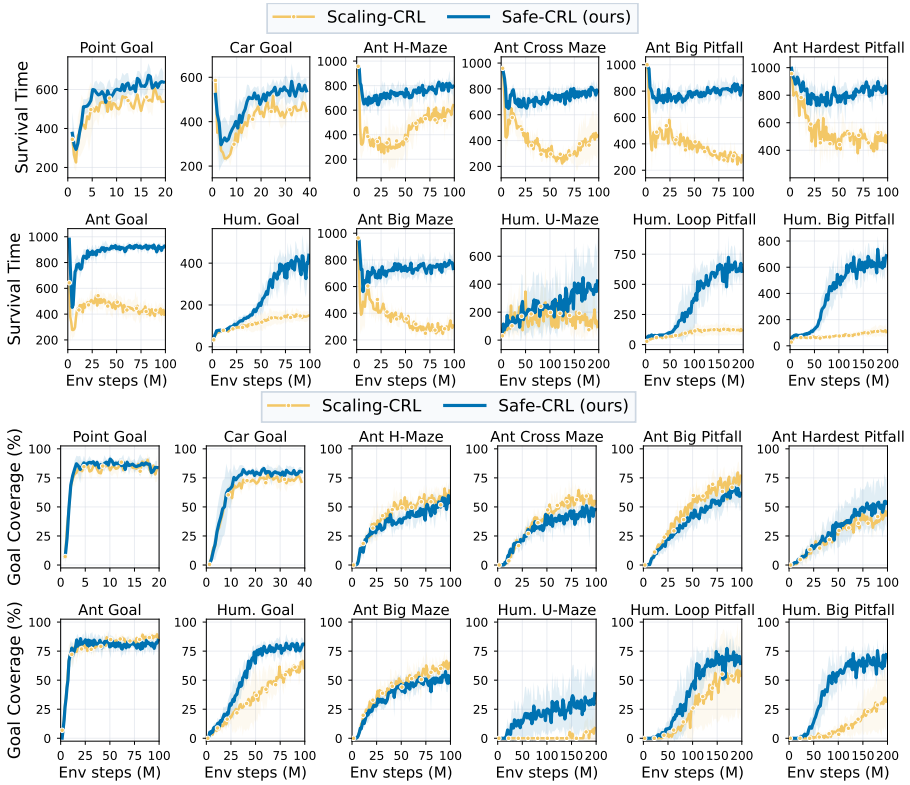


Figure D.1: Survival time (top) and goal coverage (bottom) on the 12-task benchmark (mean ± std).

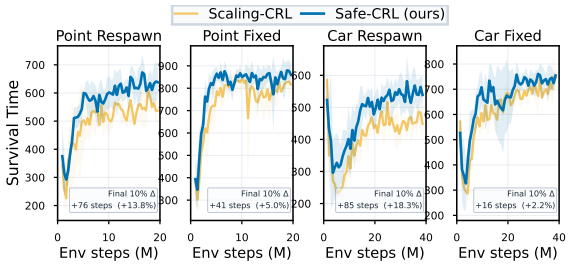


Figure D.2: Comparison of survival time with respawned and fixed goals in Point Goal and Car Goal. Percentages show the mean advantage of Safe-CRL over Scaling-CRL during the final 10% of training.

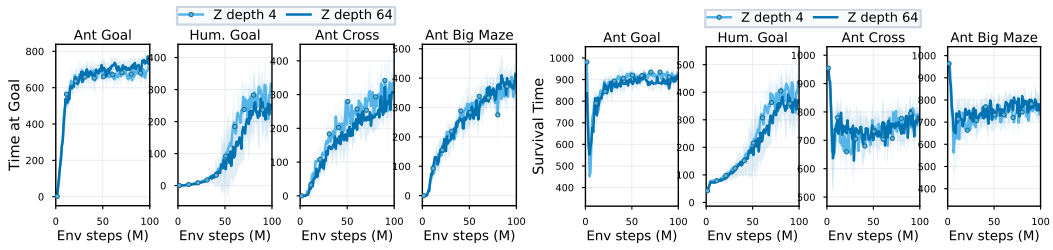


Figure D.3: Ablation on Z-encoder depth. Actor and critic depths are fixed at 16 layers.

As shown in Figure D.4, TD-based estimation produces mixed results. It substantially degrades goal-reaching and survival performance in Car Goal and Humanoid Goal, while remaining competitive on the selected Ant tasks. We therefore use direct Monte Carlo survival-mass estimation as the default because it is simpler and provides more consistent performance across the evaluated environments.

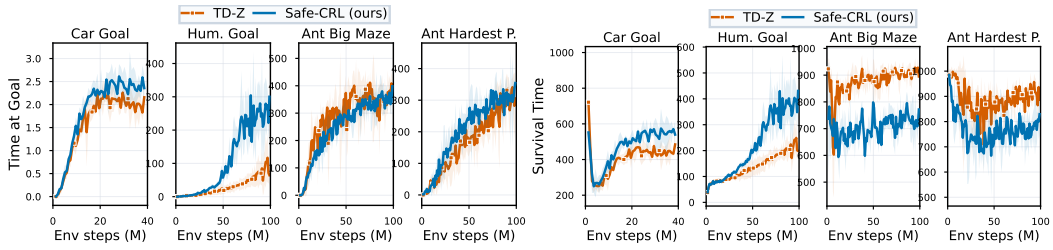


Figure D.4: Ablations on TD-learned Z. Actor and critic depths are set to 16 layers (8 layers for Car Goal).

Computational overhead Table D.1 reports the wall-clock training time. Humanoid locomotion tasks are tested on a modified NVIDIA RTX 4090 (48 GB) and the others are tested on a modified NVIDIA RTX 4080 (32 GB). For each environment, Scaling-CRL and Safe-CRL are measured on the same hardware configuration. Using the main benchmark architectures and a 4-layer Z-encoder, the average training time increases from 9.07 to 9.32 hours per 100M environment steps, corresponding to an overhead of 2.8%.

Table D.1: Average wall-clock training time per 100M environment steps.

Environment	Scaling-CRL	Safe-CRL (ours)
Point Goal	2.92 h	3.07 h
Car Goal	8.73 h	8.94 h
Ant Goal	6.65 h	6.71 h
Humanoid Goal	7.28 h	7.74 h
Ant H-Maze	10.20 h	10.58 h
Ant Cross Maze	10.48 h	10.64 h
Ant Big Maze	10.49 h	10.70 h
Humanoid U-Maze	11.09 h	11.13 h
Ant Big Pitfall	9.85 h	10.01 h
Ant Hardest Pitfall	10.42 h	10.66 h
Humanoid Loop Pitfall	10.55 h	11.01 h
Humanoid Big Pitfall	10.21 h	10.70 h
Mean	9.07 h	**9.32 h (↑ 2.8%)**

Exploration remains a limitation Safe-CRL corrects the occupancy information used for critic and policy learning, but does not modify the exploration mechanism of CRL. Figure D.5 visualizes goal-reaching success rates across different goal locations. The learned success distributions can remain spatially uneven, including asymmetric performance at geometrically similar goal locations. For example, for Hum. Loop Pitfall, the top-left corner and the bottom-left corner represent the left-back and right-back side of the humanoid robot, respectively. The right-back side has a higher goal-reaching success rate than the left-back side. This pattern shows spatially uneven goal-reaching performance and indicates that restoring survival mass does not by itself guarantee uniform goal-space coverage.

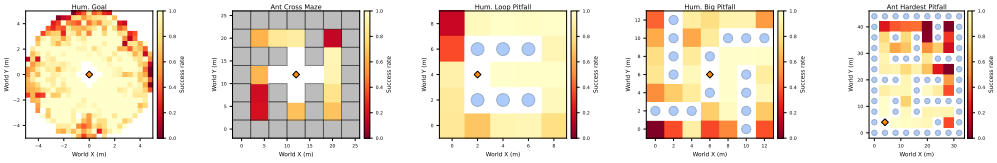


Figure D.5: Success rate for goals at different places in the playground. Darker colors indicate lower goal-reaching success rates. The robot initially faces the +x axis direction (→). The diamond marks the initial position of the robot.