Safe-CDT: Adaptive Target Scheduling for Safe Cross-Domain Deployment of Constrained Decision Transformers

Jiajun Shen [1] **Bijan SayyarRodsari** [1]

Abstract

Deploying Constrained Decision Transformers (CDT) across environments with different dynamics and safety budgets is challenging: fixed return-to-go and cost-to-go targets become brittle under distribution shift, leading to constraint violations. We propose Safe-CDT, a deployment framework combining adaptive target scheduling driven by Wilson-score achievability estimates, cost-aware multiplicative reweighting, and lightweight LoRA finetuning. We derive finite-episode probabilistic safety bounds, a critical-budget threshold, and a sufficient condition for cross-environment safety transfer. On the main DSRL cross-domain pair (CG1→CG2) under a consistent runtime-budget protocol, Safe-CDT achieves mean cost well below budget (8.56 at B=30) with the lowest observed violation rate among the baseline methods considered on that pair; additional pairs show environment-dependent safety behavior. We additionally evaluate CTG target responsiveness under environment shift via a 150-run target sweep, supporting the use of CTG conditioning as a deployment-time control variable for cost–reward tradeoff management.

1. Introduction

Safety constraints are central to sequential decision-making problems in robotics, autonomous systems, and industrial control, where an agent must maximize task performance while keeping cumulative safety cost below a specified budget. In practice, the data that are easiest to collect are often conservative logs from simplified task variants or low-risk conditions, while deployment environments are harder and potentially differ in dynamics. This mismatch creates a practical offline-to-online (O2O) challenge: leveraging offline data for sample efficiency while adapting online *without* sacrificing safety. A representative application is intralogistics with autonomous mobile robots, where logs from an earlier facility layout must drive deployment after reconfiguration, denser traffic, or a revised cumulative safety budget—the robot interface stays fixed even though the risk geometry becomes harder, so what is actually needed is online *recalibration*, not retraining.

Decision Transformer (DT) casts RL as conditional sequence modeling, generating actions from a trajectory context and a user-specified return-to-go (RTG) token (Chen et al., 2021). Constrained Decision Transformer (CDT) extends this to constrained MDPs by additionally conditioning on a cost-to-go (CTG) token (Liu et al., 2023). This conditioning interface is attractive because it exposes a direct control knob at inference time. However, two coupled difficulties arise in deployment. First, the correct target pair is unknown and may drift under distribution shift. Second, when targets fall outside the support of offline data, target mismatch can directly cause unsafe behavior—an issue well-known in unconstrained DT adaptation (Zheng et al., 2022; Yan et al., 2024), but more severe in safe RL where mismatch translates directly into constraint violations.

A natural response is online finetuning, yet O2O RL research has shown that successful adaptation depends critically on calibration and careful bridging of offline and online distributions, not merely additional gradient updates (Lee et al., 2021; Nakamoto et al., 2023). These observations motivate a more targeted deployment loop: instead of spending online budget on redundant rollouts, use online feedback primarily to *calibrate conditioning targets* and acquire deployment data for lightweight adaptation.

We study this problem in a *cross-domain deployment* protocol where offline data are collected in an easier source environment and deployment occurs in a harder target environment with shifted dynamics. We first characterize why CDT fails in this regime, then propose Safe-CDT, a deployment procedure that calibrates RTG/CTG conditioning online and finetunes the model with a small online buffer.

Contributions.

[1]AI Center of Excellence, Rockwell Automation, Austin, TX 78759, USA. Correspondence to: Jiajun Shen <sjjvic@gmail.com>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

1. A **cost-prioritized adaptive target scheduler** using conservative Wilson-score lower bounds on cost achievability to tighten or relax CTG online, while gating reward-seeking behavior through a cost-prioritized controller.
2. A **lightweight O2O deployment loop** combining adaptive target scheduling, cost-aware multiplicative reweighting, and LoRA finetuning designed for limited online interaction.
3. A **compact deployment-time safety theory**: single- and multi-episode probabilistic safety bounds, a critical-budget threshold, and a cross-environment safety-transfer condition.
4. **Empirical evaluation under a consistent runtime-budget protocol**: in the cross-method baseline comparison, Safe-CDT achieves the lowest observed violation rate on the main cross-domain pair (14.4%) with mean cost well below budget. Baseline characterization identifies three failure modes in offline safe RL: budget saturation, reward collapse, and seed instability.
5. **CTG target-responsiveness evaluation**: a 150-run experiment combining static target sweeps and runtime step-change interventions shows that the CTG conditioning interface is a directionally responsive deployment-time control variable.

2. Related Work

Offline safe RL. Safe RL is broadly surveyed in (García & Fernández, 2015; Gu et al., 2024). Within the offline regime, representative methods include CPQ, COptiDICE, and COPO (Xu et al., 2022; Lee et al., 2022; Polosky et al., 2022), while more recent work such as OASIS, FISOR, and CAPS improve learning via data shaping, feasibility filtering, or adaptive decision rules (Yao et al., 2024; Zheng et al., 2024; Chemingui et al., 2025; Liu et al., 2024). These focus on offline learning quality under a fixed deployment setting.

Decision transformers for safe RL. DT reframes offline RL as autoregressive sequence modeling conditioned on desired return (Chen et al., 2021). CDT extends this to CMDPs with cost conditioning (Liu et al., 2023), and more recent work conditions on temporal logic specifications (Guo et al., 2024). These demonstrate that transformer-based policies are appealing for offline safe control but do not address safe deployment under environment shift.

Offline-to-online adaptation. Representative O2O RL approaches include balanced replay, policy expansion, actor–critic alignment, and Cal-QL (Lee et al., 2021; Zhang et al., 2023; Yu et al., 2023; Nakamoto et al., 2023). For sequence-modeling policies, ODT and RL Gradients as Vitamin extend DT to online finetuning (Zheng et al., 2022; Yan et al., 2024). In safe RL, GOLD and FOSP study O2O adaptation with actor–critic or world-model machinery (Li et al., 2024; Cao et al., 2025). Our focus is narrower: deployment of a target-conditioned transformer under environment shift and changing budgets, where the main control variable is the conditioning target itself. To our knowledge, prior DT-based safe RL work has not studied this source-to-target cross-domain protocol with deployment-time target calibration as the primary mechanism.

Uncertainty quantification for offline decision-making. Distribution-free uncertainty quantification, in particular conformal prediction, is increasingly used for calibrated risk estimates from limited interaction data (Angelopoulos & Bates, 2023). We use the classical Wilson-score interval (Wilson, 1927) as a small-sample, conservative lower bound on cost achievability, providing a lightweight UQ signal directly inside the deployment loop.

3. Method

3.1. Problem Setup

We consider a CMDP $(\mathcal{S}, \mathcal{A}, P, R, C, \gamma, d)$, where the goal is to learn a policy π maximizing expected return subject to a cost constraint:

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\Big[\sum_t \gamma^t R_t\Big] \text{ s.t. } \mathbb{E}_{\tau\sim\pi}\Big[\sum_t \gamma^t C_t\Big] \le d, \quad (1)$$

where $R_t = R(s_t, a_t)$ and $C_t = C(s_t, a_t)$. In our cross-domain deployment protocol, a fixed offline dataset is collected in a source environment $\mathcal{M}_s$, while deployment and online adaptation occur in a harder target environment $\mathcal{M}_t$ sharing the same task family and state–action dimensionality but with shifted dynamics.

3.2. Safe-CDT Overview

Safe-CDT builds on CDT as a target-conditioned sequence model. We retain the pretrained CDT backbone and adapt it online using a small LoRA module (Hu et al., 2022): only $\sim$65k parameters are updated ($\approx$1.9% of the 3.5M-parameter model). During deployment, each episode triggers three coupled operations: target scheduling (every 5 episodes), reweighting multiplier update (after a warmup period), and lightweight LoRA finetuning (after each episode). Figure 1 contrasts this closed-loop deployment with the standard fixed-target baseline.

3.3. Adaptive Target Scheduler

The scheduler controls a normalized cost ratio $CR_k \in [0, 1]$; the actual CTG target at stage k is $CTG_k = CR_k \cdot B$. The scheduler uses the Wilson-score lower bound A_C^{lower} as a conservative estimate of cost achievability and updates CR

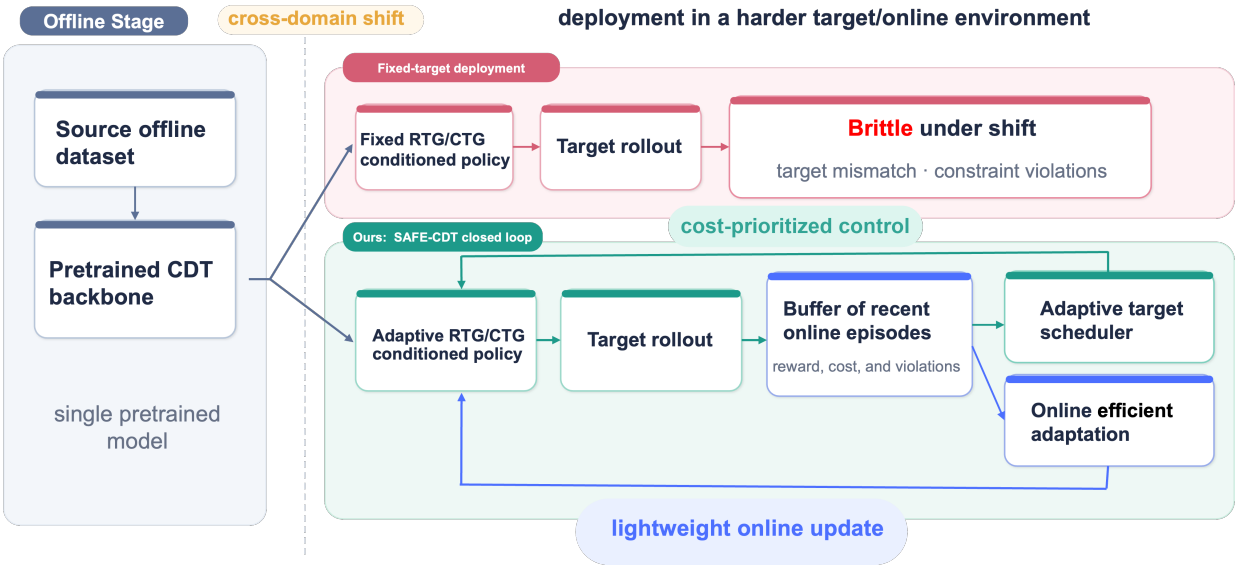

Figure 1. **Cross-domain deployment of a pretrained CDT: fixed-target baseline vs. Safe-CDT.** *Left:* a single CDT backbone is pretrained once on source offline data. *Top right:* a fixed RTG/CTG conditioned policy is brittle under cross-domain shift—target mismatch and constraint violations. *Bottom right:* Safe-CDT closes the deployment loop with cost-prioritized control—an adaptive target scheduler calibrates RTG/CTG from recent online rollouts, while lightweight LoRA adapters are finetuned with cost-aware replay reweighting.

according to:

$$\Delta CR_k = \begin{cases} -\Delta^C_{\text{step}}\, d_t, & \text{if } A^{\text{lower}}_C < 0.5 \ \text{(tighten)}, \\ +\Delta^C_{\text{step}}\, d_t, & \text{if } A^{\text{lower}}_C > 0.9 \ \text{(relax)}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where d_t is a momentum factor. The RTG target is updated by a separate achievability estimate but is frozen whenever the safety condition is not comfortably satisfied. Two design choices are key: the scheduler is *feedback-driven* (reacting to observed deployment behavior) and *asymmetric in objective* (cost calibration is prioritized over reward calibration).

3.4. Cost-Aware Reweighting and LoRA Finetuning

During online finetuning, Safe-CDT uses cost-aware multiplicative reweighting rather than an additive constraint penalty. For each replay sequence i in a minibatch, let $\tilde{C}_i$ be an unscaled CTG-based cost proxy and define its budget excess

$$e_i = \max(0, \tilde{C}_i - B). \tag{3}$$

The LoRA adapter is trained with a reweighted behavioral cloning loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} w_i \|\hat{a}_i - a_i\|^2, \tag{4}$$

where

$$\tilde{w}_i = \text{clip}(\exp(-\lambda e_i / B), 0.1, \infty)\,, \qquad w_i = \frac{\tilde{w}_i}{N^{-1}\sum_{j=1}^{N}\tilde{w}_j}. \tag{5}$$

The multiplier λ controls the strength of replay reweighting: it increases quickly when recent violations are frequent and relaxes slowly otherwise. Thus, high-cost sequences receive smaller training weight while the overall behavioral-cloning loss scale remains stable.

4. Deployment-Time Safety Analysis

We formalize three aspects of deployment-time behavior: finite-episode safety, threshold behavior in the budget, and transfer under environment shift. These results provide *probabilistic* safety bounds under the stated assumptions; empirical violation rates depend on how well the assumptions hold in practice. All proofs are deferred to Appendix A.

Notation. Let $c \in [c_{\min}, c_{\max}]$ denote the CTG condition, $\hat{C}$ the realized episode cost, $\phi(c) := \mathbb{E}[\hat{C} \mid \text{CTG} = c]$ the conditional mean cost, and $\sigma(c)$ a conditional sub-Gaussian scale parameter satisfying $\mathbb{E}[\exp(\lambda(\hat{C}-\phi(c))) \mid \text{CTG}{=}c] \le \exp(\lambda^2\sigma(c)^2/2)$ for all $\lambda \in \mathbb{R}$. The safety margin is $\Delta(c, B) := B - \phi(c)$.

Assumption 4.1 (Controllability)**.** *ϕ is nondecreasing and L_ϕ-Lipschitz on $[c_{\min}, c_{\max}]$.*

Assumption 4.2 (Concentration)**.** *$\hat{C} - \phi(c)$ is conditionally $\sigma(c)$-sub-Gaussian for every $c \in [c_{\min}, c_{\max}]$.*

4.1. Finite-Episode Safety

Theorem 4.1 (Single-episode safety)**.** *Under Assumption 4.2, if $\Delta(c^\star, B) = \delta > 0$, then $\mathbb{P}(\hat{C} \le B \mid \text{CTG}{=}c^\star) \ge 1 - \exp(-\delta^2/(2\sigma(c^\star)^2))$.*

Theorem 4.2 (Multi-episode safety). *For N conditionally independent episodes with the same $c^\star$ and $\Delta(c^\star, B) = \delta > 0$: $\mathbb{P}(\hat{C}_i \le B, \forall i) \ge (1 - \exp(-\delta^2/(2\sigma(c^\star)^2)))^N$. A sufficient condition for $\mathbb{P}(\hat{C}_i \le B, \forall i) \ge 1-\alpha$ is $\delta \ge \sigma(c^\star)\sqrt{2\log(N/\alpha)}$.*

4.2. Critical Budget and Threshold Behavior

Definition 4.1 (Critical budget). *$B^\star := \inf_{c \in [c_{\min}, c_{\max}]} \big(\phi(c) + \kappa\sigma(c)\big)$, where $\kappa > 0$ is the target confidence coefficient.*

Theorem 4.3 (Budget threshold). *Let $\sigma_{\max} := \sup_c \sigma(c)$ and assume the infimum defining $B^\star$ is attained at $c^\dagger$. **(a)** If $B < B^\star$, no CTG satisfies $\Delta(c, B) \ge \kappa\sigma(c)$, so the concentration framework cannot certify exponentially small violation probability. **(b)** If $B > B^\star$, then $\mathbb{P}(\hat{C} > B \mid c^\dagger) \le \exp\big(-(B-B^\star)^2/(2\sigma^2_{\max})\big)$. A union bound over N episodes gives $\mathbb{P}(\exists i : \hat{C}_i > B) \le N \exp\big(-(B-B^\star)^2/(2\sigma^2_{\max})\big)$. **(c)** $\mathbb{P}(\exists i : \hat{C}_i > B) \le \alpha$ holds when $B \ge B^\star + \sigma_{\max}\sqrt{2\log(N/\alpha)}$.*

4.3. Cross-Environment Transfer

Definition 4.2 (Environment shift). $\Delta_{\mathcal{M}} := \sup_c |\phi_1(c) - \phi_2(c)|$.

Theorem 4.4 (Zero-step safety transfer). *If $c^\star$ achieves margin $\delta_1 := B - \phi_1(c^\star)$ in source $\mathcal{M}_1$ and $\bar{\sigma}_2 := \sup_c \sigma_2(c)$, then the target margin satisfies $\delta_2 \ge \delta_1 - \Delta_{\mathcal{M}}$. If $\Delta_{\mathcal{M}} < \delta_1 - \kappa\bar{\sigma}_2$, then $\mathbb{P}(\hat{C} > B \mid \mathrm{CTG}{=}c^\star, \mathcal{M}_2) \le e^{-\kappa^2/2}$.*

Theorem 4.5 (Unified threshold). *For M environments with $B^\star_{\max} := \max_m \inf_c(\phi_m(c) + \kappa\sigma_m(c))$ and $\sigma_{\max} := \max_m \sup_c \sigma_m(c)$, if $B > B^\star_{\max}$:*

$$\mathbb{P}\big(\exists m, i : \hat{C}_{m,i} > B\big) \le MN \exp\Big(-\frac{(B - B^\star_{\max})^2}{2\sigma^2_{\max}}\Big).$$

5. Experiments

5.1. Setup

We evaluate on the DSRL benchmark (Liu et al., 2024) in a cross-domain deployment protocol: pretraining uses an easier Level-1 source, deployment occurs in the harder Level-2 target. We study three task families: CarGoal (CG), CarCircle (CC), and PointCircle (PC). All experiments use a **consistent runtime-budget protocol** where the wrapper budget exactly matches the stated evaluation budget B. Online deployment uses 50 episodes; CTG/RTG scheduling updates every 5 episodes; LoRA finetuning performs 10 gradient steps per episode; the reweighting multiplier updates after a 10-episode warmup. Metrics: violation rate (VR, percentage of episodes exceeding the budget), mean reward, and average cost.

Baselines include: (i) **Offline safe RL**: BCQL, BEARL, CPQ (Fujimoto et al., 2019; Kumar et al., 2019; Xu et al., 2022); (ii) **CDT-centered**: zero-shot CDT (no adaptation) and CDT with naive LoRA finetuning.

Table 1. Cross-domain deployment at budget $B{=}30$ under a consistent runtime-budget protocol. CG1→CG2 uses 15 seeds; CC and PC use 3 seeds.

Pair	Reward	Cost	VR (%)
CG1→CG2	$-15.93_{\pm 2.01}$	$8.56_{\pm 2.21}$	$\mathbf{14.4}_{\pm 6.3}$
CC1→CC2	$0.77_{\pm 0.29}$	$9.05_{\pm 1.60}$	$24.7_{\pm 5.0}$
PC1→PC2	$-12.32_{\pm 3.30}$	$19.21_{\pm 3.34}$	$56.0_{\pm 11.1}$

Table 2. Budget sensitivity for Safe-CDT on CG1→CG2 (3 seeds, 50 episodes each).

Budget	Reward	Cost	VR (%)
$B{=}25$	$-13.91_{\pm 0.92}$	$8.17_{\pm 1.49}$	$24.7_{\pm 3.1}$
$B{=}30$	$-15.40_{\pm 2.99}$	$7.87_{\pm 0.67}$	$\mathbf{12.7}_{\pm 3.1}$
$B{=}35$	$-15.71_{\pm 2.87}$	$10.56_{\pm 1.87}$	$14.0_{\pm 5.3}$
$B{=}40$	$-17.33_{\pm 0.26}$	$9.18_{\pm 2.56}$	$9.3_{\pm 4.2}$
$B{=}50$	$-14.88_{\pm 1.23}$	$11.40_{\pm 1.14}$	$10.0_{\pm 3.5}$

5.2. Cross-Domain Deployment

Table 1 presents the main evaluation at budget $B{=}30$. On the main pair (CG1→CG2, 15 seeds), Safe-CDT achieves mean cost 8.56 (well below $B{=}30$) with VR 14.4%; in the cross-method comparison in Table 4, this is the lowest observed violation rate among the baseline methods considered on this pair. Safety behavior is environment-dependent: CC1→CC2 yields VR 24.7% and PC1→PC2 yields VR 56.0%, demonstrating that no uniform low-violation claim is warranted.

5.3. Budget Sensitivity

Table 2 reports VR for CG1→CG2 across five budget levels. Violation rates are lower for $B{\ge}30$ than for $B{=}25$, consistent with larger budgets providing more room for the adaptive scheduler. However, VR is non-monotone within $B{\ge}30$ (e.g., $B{=}35$ yields higher VR than $B{=}30$), indicating budget sensitivity rather than a discrete threshold effect.

5.4. Ablation and CDT Attribution

Table 3 compares method variants on CG1→CG2 at $B{=}30$. All three variants achieve similar violation rates (14–17%) and mean costs well below budget; differences are within the standard deviation across seeds. Notably, zero-shot CDT inference (no adaptation, no scheduler) already achieves VR 15.6% on this pair, within one standard deviation of the full method. This indicates that the pre-trained CDT backbone's conservatism is the primary driver of low VR on CG1→CG2; the scheduler's contribution is best understood as providing *runtime adjustability* rather than unconditional VR reduction.

Table 3. Ablation and CDT attribution on CG1→CG2 at B=30 (3 seeds, 50 episodes each).

Variant	VR (%)↓	Reward	Cost
Safe-CDT (full)	$16.7_{\pm 5.8}$	$-14.67_{\pm 0.97}$	$8.97_{\pm 1.49}$
No reweighting	$15.3_{\pm 2.3}$	$-16.15_{\pm 1.67}$	$8.56_{\pm 0.44}$
Scheduler only	$14.0_{\pm 7.2}$	$-14.97_{\pm 2.68}$	$8.41_{\pm 1.91}$
CDT zero-shot	$15.6_{\pm 7.0}$	$-14.74_{\pm 2.73}$	$8.09_{\pm 1.84}$
CDT + naive LoRA	$14.8_{\pm 6.7}$	$-16.47_{\pm 2.01}$	$8.38_{\pm 1.69}$

Table 4. Cross-method comparison on CG1→CG2, B=30. Safe-CDT uses online adaptation (50 episodes, 15 seeds); OSRL baselines use offline-only evaluation (20 episodes, 3 seeds).

Method	Reward	Cost	VR (%)	N
Safe-CDT	$-15.93_{\pm 2.01}$	$\mathbf{8.56}_{\pm 2.21}$	$\mathbf{14.4}_{\pm 6.3}$	15
BEARL	$7.89_{\pm 1.06}$	$30.75_{\pm 0.39}$	$95.0_{\pm 5.0}$	3
BCQL	$3.93_{\pm 1.28}$	$24.72_{\pm 1.08}$	$73.3_{\pm 7.6}$	3
CPQ	$-0.55_{\pm 0.34}$	$11.75_{\pm 3.49}$	$28.3_{\pm 15.3}$	3

5.5. Baseline Comparison

Table 4 places Safe-CDT alongside offline safe RL baselines on CG1→CG2. The comparison reveals three distinct constraint-behavior archetypes: *Budget-saturating* (BEARL, BCQL): high reward but near-total budget exhaustion, with costs pinned at or near B. *Over-conservative* (CPQ): low cost but near-zero or negative reward; on Point-Circle1, CPQ additionally exhibits seed instability (VR spans 0%–100% across seeds). *Conservative-adaptive* (Safe-CDT): lowest VR and lowest cost among these baseline methods on CG1→CG2, with negative reward reflecting conservative target selection.

5.6. CTG Target Controllability

The CDT attribution above motivates a direct question: does the cost-to-go conditioning interface actually respond to target changes under environment shift? If cost behavior is insensitive to the CTG target, the "runtime knob" interpretation has little content.

We sweep 5 CTG target ratios (CR $\in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) across all 3 pairs with 5 seeds each (50 episodes/seed), using zero-shot CDT inference (no adaptation, no scheduler, RTG ratio $= 0.5$, B=30). Table 5 reports the results.

All three pairs show positive Spearman $\rho > 0.5$ between CTG target ratio and mean cost, indicating a directionally correct cost response. CC1→CC2 achieves the strongest correlation (ρ=0.9, p=0.037). We additionally run a **runtime step-change intervention**: after 50 episodes at one CTG level, the target switches (low→high or high→low) for 50 more episodes. All three pairs satisfy a directional consistency criterion (cost moves in the expected direction in $\geq$3 of 5 seeds). Figure 2 visualizes both experiments.

Table 5. CTG target controllability: mean cost as a function of CTG target ratio (zero-shot CDT, B=30, 5 seeds, 50 ep/seed).

	CTG target ratio				
Pair	0.2	0.4	0.6	0.8	1.0
CG1→CG2	7.74	8.17	7.42	9.45	9.08
CC1→CC2	15.16	16.03	16.66	16.28	16.95
PC1→PC2	27.17	28.04	27.78	27.45	28.74

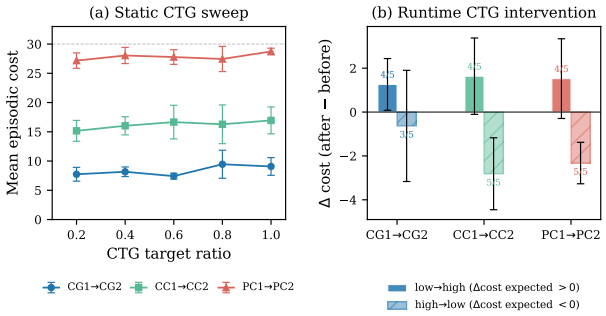

Figure 2. **CTG target responsiveness** across three deployment pairs (zero-shot CDT, B=30). *Left:* Static sweep—mean cost vs. CTG target ratio; all pairs show a positive trend. *Right:* Runtime intervention—cost delta after mid-deployment target switch; bars show expected direction with seed-level consistency annotations. The CTG interface is directionally responsive but not monotone.

These results indicate that CTG conditioning provides a *measurable* deployment-time tradeoff knob: higher CTG targets are associated with higher realized cost. The claim is deliberately narrow—cost response is not perfectly monotone, PC1→PC2 exhibits cost saturation near B, and no formal safety guarantee follows. The adaptive scheduler leverages this responsiveness for online calibration.

6. Conclusion

We presented Safe-CDT, an offline-to-online deployment framework for constrained decision transformers under environment shift. The key idea is that RTG/CTG conditioning should be treated not as a fixed offline hyperparameter but as a feedback-controlled deployment variable calibrated online with conservative feedback. Under a consistent runtime-budget protocol, Safe-CDT achieves the lowest observed violation rate in the cross-method baseline comparison on the main cross-domain pair (14.4%) with mean cost well below budget. A 150-run CTG target-sweep experiment supports that the conditioning interface is directionally responsive under environment shift, consistent with the runtime-adjustable control interpretation. Probabilistic safety bounds formalize the relationship between budget margin, cost concentration, and violation probability.

Limitations. Safety behavior is environment-dependent: VR varies from 14% to 56% across deployment pairs, and no uniform low-violation claim is made. The CDT backbone's intrinsic conservatism accounts for most of the ob-

served cost control on the main pair. Our experiments are confined to DSRL tasks with vector observations and shared task families. Extending to visual observations, larger morphology changes, and multiple simultaneous constraints is future work.

Impact Statement

This paper presents work whose goal is to advance safe deployment of reinforcement learning agents. The methods are designed to reduce deployment-time constraint risk during online adaptation, which is relevant to safety-critical applications. We do not foresee specific negative societal consequences beyond those common to the broader field of machine learning.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Cao, C., Xin, Y., Wu, S., He, L., Yan, Z., Tan, J., and Wang, X. Fosp: Fine-tuning offline safe policy through world models. In *International Conference on Learning Representations*, 2025.

Chemingui, Y., Deshwal, A., Wei, H., Fern, A., and Doppa, J. R. Constraint-adaptive policy switching for offline safe reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:15722–15730, 2025.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062, 2019.

García, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., and Knoll, A. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Guo, Z., Zhou, W., and Li, W. Temporal logic specification-conditioned decision transformer for offline safe reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17003–17019, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lee, J., Paduraru, C., Mankowitz, D. J., Heess, N., Precup, D., Kim, K.-E., and Guez, A. Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation. *arXiv preprint arXiv:2204.08957*, 2022.

Lee, S., Seo, Y., Lee, K., Abbeel, P., and Shin, J. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, Proceedings of Machine Learning Research, 2021.

Li, J., Liu, X., Zhu, B., Jiao, J., Tomizuka, M., Tang, C., and Zhan, W. Guided online distillation: Promoting safe reinforcement learning by offline demonstration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7447–7454, 2024.

Liu, Z., Guo, Z., Yao, Y., Cen, Z., Yu, W., Zhang, T., and Zhao, D. Constrained decision transformer for offline safe reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21611–21630, 2023.

Liu, Z., Guo, Z., Lin, H., Yao, Y., Zhu, J., Cen, Z., Hu, H., Yu, W., Zhang, T., Tan, J., and Zhao, D. Datasets and benchmarks for offline safe reinforcement learning. *Data-centric Machine Learning Research*, 1:1–29, 2024.

Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., and Levine, S. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Polosky, N., Da Silva, B. C., Fiterau, M., and Jagannath, J. Constrained offline policy optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Xu, H., Zhan, X., and Zhu, X. Constraints penalized q-learning for safe offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 8753–8760, 2022.

Yan, K., Schwing, A. G., and Wang, Y.-X. Reinforcement learning gradients as vitamin for online finetuning decision transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yao, Y., Cen, Z., Ding, W., Lin, H., Liu, S., Zhang, T., Yu, W., and Zhao, D. Oasis: Conditional distribution shaping for offline safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yu, Z., Zhang, X., Xu, W., and Lu, Z. Actor-critic alignment for offline-to-online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Zhang, H., Xu, W., and Yu, H. Policy expansion for bridging offline-to-online reinforcement learning. In *International Conference on Learning Representations*, 2023.

Zheng, Q., Zhang, A., and Grover, A. Online decision transformer. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27042–27059, 2022.

Zheng, Y., Li, J., Yu, D., Yang, Y., Li, S. E., Zhan, X., and Liu, J. Safe offline reinforcement learning with feasibility-guided diffusion model. In *International Conference on Learning Representations*, 2024.

A. Proofs of Main Theorems

A.1. Proof of Theorem 4.1

Set $X := \hat{C} - \phi(c^\star)$. By Assumption 4.2, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[e^{\lambda X} \mid \mathrm{CTG} = c^\star\big] \le \exp\left(\frac{\lambda^2 \sigma(c^\star)^2}{2}\right).$$

Since $\Delta(c^\star, B) = \delta$, we have $B = \phi(c^\star) + \delta$, so $\mathbb{P}(\hat{C} > B \mid \mathrm{CTG} = c^\star) = \mathbb{P}(X > \delta \mid \mathrm{CTG} = c^\star)$. By Markov's inequality with $\lambda > 0$:

$$\mathbb{P}(X > \delta \mid \mathrm{CTG} = c^\star) \le e^{-\lambda\delta}\mathbb{E}[e^{\lambda X} \mid \mathrm{CTG} = c^\star] \le \exp\left(-\lambda\delta + \frac{\lambda^2 \sigma(c^\star)^2}{2}\right).$$

Optimizing at $\lambda^\star = \delta/\sigma(c^\star)^2$ gives $\mathbb{P}(\hat{C} > B \mid \mathrm{CTG} = c^\star) \le \exp(-\delta^2/(2\sigma(c^\star)^2))$.

A.2. Proof of Theorem 4.2

By conditional independence, $\mathbb{P}(\bigcap_i A_i) = \prod_i \mathbb{P}(A_i)$ where $A_i := \{\hat{C}_i \le B\}$. Theorem 4.1 gives $\mathbb{P}(A_i) \ge 1 - \exp(-\delta^2/(2\sigma(c^\star)^2))$, proving the product bound. For the $1-\alpha$ guarantee, the union bound gives $\mathbb{P}(\bigcup_i A_i^c) \le N\exp(-\delta^2/(2\sigma(c^\star)^2))$, and setting this $\le \alpha$ yields $\delta \ge \sigma(c^\star)\sqrt{2\log(N/\alpha)}$.

A.3. Proof of Theorem 4.3

(a) If $B < B^\star$, then $\phi(c) + \kappa\sigma(c) \ge B^\star > B$ for all c, so $\Delta(c, B) < \kappa\sigma(c)$ everywhere.

(b) At the minimizer $c^\dagger$: $\phi(c^\dagger) + \kappa\sigma(c^\dagger) = B^\star$, so $\delta^\dagger = B - \phi(c^\dagger) = B - B^\star + \kappa\sigma(c^\dagger) \ge B - B^\star$. Theorem 4.1 at $c^\dagger$ with $\sigma(c^\dagger) \le \sigma_{\max}$ gives $\mathbb{P}(\hat{C} > B \mid c^\dagger) \le \exp(-(B - B^\star)^2/(2\sigma_{\max}^2))$. The union bound over N episodes yields the stated result.

(c) Setting $N\exp(-(B - B^\star)^2/(2\sigma_{\max}^2)) \le \alpha$ and solving gives $B \ge B^\star + \sigma_{\max}\sqrt{2\log(N/\alpha)}$.

A.4. Proof of Theorem 4.4

By Definition 4.2, $|\phi_1(c^\star) - \phi_2(c^\star)| \le \Delta_{\mathcal{M}}$, so $\phi_2(c^\star) \le \phi_1(c^\star) + \Delta_{\mathcal{M}}$. Subtracting from B: $\delta_2 = B - \phi_2(c^\star) \ge B - \phi_1(c^\star) - \Delta_{\mathcal{M}} = \delta_1 - \Delta_{\mathcal{M}}$. Under condition $\Delta_{\mathcal{M}} < \delta_1 - \kappa\bar{\sigma}_2$, we get $\delta_2 > \kappa\bar{\sigma}_2 \ge \kappa\sigma_2(c^\star)$, and Theorem 4.1 in $\mathcal{M}_2$ gives $\mathbb{P}(\hat{C} > B \mid c^\star, \mathcal{M}_2) \le e^{-\kappa^2/2}$.

A.5. Proof of Theorem 4.5

Since $B > B^\star_{\max} \ge B^\star_m$ for all m, Theorem 4.3(b) applies in each environment. Because $B - B^\star_m \ge B - B^\star_{\max}$, a union bound over M environments and N episodes per environment gives the stated bound.

B. Auxiliary Stochastic-Approximation Result

Theorem B.1 (Idealized scheduler convergence)**.** *Let $\{c_k\} \subseteq [c_{\min}, c_{\max}]$ satisfy the projected stochastic-approximation recursion $c_{k+1} = \Pi_{[c_{\min}, c_{\max}]}(c_k + \eta_k(h(c_k) + M_{k+1}))$, where Π is Euclidean projection, $\eta_k > 0$ with $\sum \eta_k = \infty$ and $\sum \eta_k^2 < \infty$, $\{M_{k+1}\}$ is a martingale-difference sequence with bounded variance σ_M^2, and h is continuous with strict drift $(c - c^\star)h(c) \le -m(c - c^\star)^2$. Then $c_k \to c^\star$ almost surely.*

Proof. Define $V_k := (c_k - c^\star)^2$. By nonexpansiveness of projection and the recursion: $V_{k+1} \le V_k + 2\eta_k(c_k - c^\star)(h(c_k) + M_{k+1}) + \eta_k^2(h(c_k) + M_{k+1})^2$. Taking conditional expectation and using $\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = 0$, boundedness $|h| \le H$, and strict drift: $\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \le V_k - 2m\eta_k V_k + C_0\eta_k^2$ where $C_0 = 2H^2 + 2\sigma_M^2$. Since $\sum \eta_k^2 < \infty$, Robbins–Siegmund's theorem gives $V_k \to 0$ a.s., hence $c_k \to c^\star$ a.s. □

Remark B.1. *This is a rigorous result for an abstract diminishing-step controller. It should not be read as a theorem about the exact fixed-step Wilson-score scheduler of Section 3.*

Table 6. Runtime CTG target intervention (zero-shot CDT, B=30, 5 seeds, 100 episodes/seed). Cost deltas compare episodes 50–99 against episodes 0–49.

Pair	Switch	Δ Cost	Δ VR (pp)	Direction
CG1→CG2	low→high	$+1.26 \pm 1.18$	$+5.6$	4/5
CG1→CG2	high→low	-0.63 ± 2.53	-2.8	3/5
CC1→CC2	low→high	$+1.64 \pm 1.74$	$+4.4$	4/5
CC1→CC2	high→low	-2.81 ± 1.64	-9.2	5/5
PC1→PC2	low→high	$+1.52 \pm 1.82$	$+0.0$	4/5
PC1→PC2	high→low	-2.32 ± 0.95	-5.2	5/5

Table 7. Per-environment baseline results (B=30, 3 seeds, 20 evaluation episodes/seed).

Pair	Method	Reward	Cost	VR (%)
CG1→CG2	BEARL	7.89 ± 1.06	30.75 ± 0.39	95.0
CG1→CG2	BCQL	3.93 ± 1.28	24.72 ± 1.08	73.3
CG1→CG2	CPQ	-0.55 ± 0.34	11.75 ± 3.49	28.3
CC1→CC2	BEARL	8.29 ± 0.67	31.00 ± 0.00	100.0
CC1→CC2	BCQL	5.35 ± 0.42	31.00 ± 0.00	100.0
CC1→CC2	CPQ	-0.07 ± 0.43	1.37 ± 0.75	0.0
PC1→PC2	BEARL	12.81 ± 2.05	26.97 ± 5.24	83.3
PC1→PC2	BCQL	16.89 ± 3.11	17.22 ± 7.93	51.7
PC1→PC2	CPQ	-1.05 ± 2.69	20.98 ± 17.49	71.7

C. Runtime CTG Intervention Details

Table 6 reports the full runtime intervention results. For each pair, we compare fixed controls (CTG ratios 0.2, 0.6, 1.0 held constant for 100 episodes) against step-change schedules that switch after 50 episodes. Direction consistency reports the number of seeds (out of 5) whose cost delta has the expected sign.

The clearest intervention signal appears on CC1→CC2, where high→low reduces mean cost by 2.81 with 5/5 seed consistency. CG1→CG2 passes but is borderline: the high→low effect is small (-0.63) and passes at the minimum threshold (3/5).

D. Per-Environment Baseline Breakdown

Table 7 reports the offline safe RL baseline results broken down by environment pair.

BEARL exhibits budget saturation on CC1→CC2 (cost $= 31$ in all seeds). BCQL similarly saturates on CC but shows lower cost on PC. CPQ is bimodal on PC1→PC2: cost spans 0 to 31 across seeds, with one seed fully violating and another fully safe.