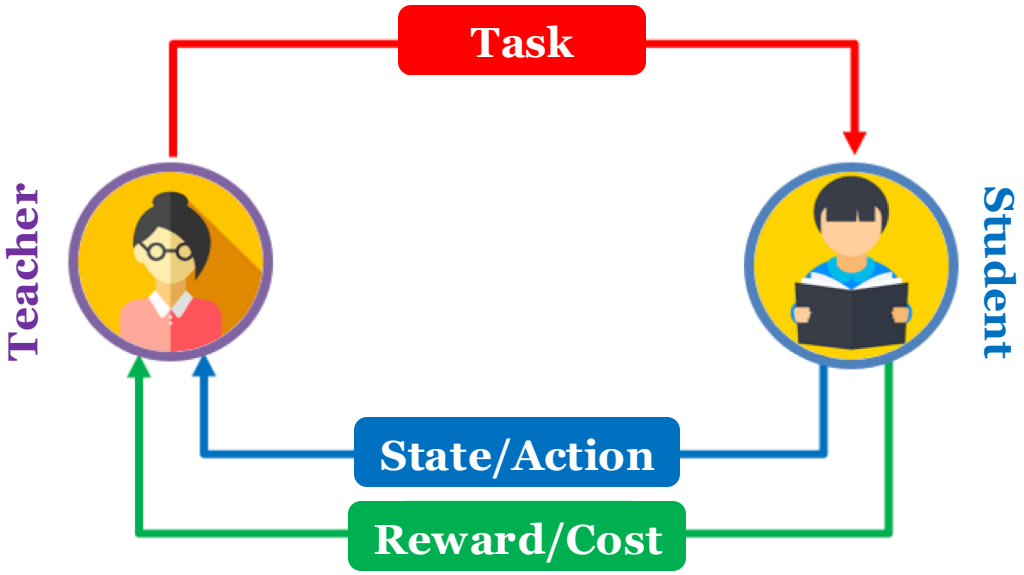
Safety-Prioritizing Curricula for Constrained Reinforcement Learning

Cevahir Koprulu[1], Thiago D. Simão[2], Nils Jansen[3], Ufuk Topcu[1]

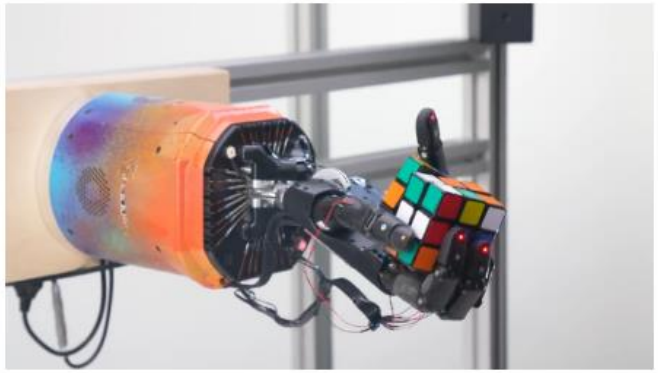
[1] The University of Texas at Austin, [2]Eindhoven University of Technology, [3]Ruhr University-Bochum

TLDR: Curriculum learning to improve safety during training and accelerate learning

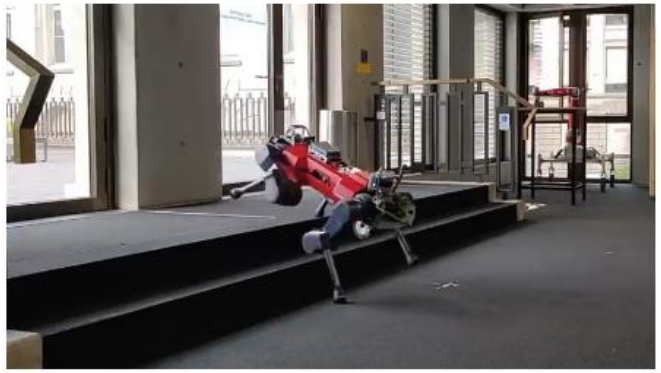
Curriculum Learning for RL



A sequence of tasks that gradually increase in difficulty to accelerate learning [1]



OpenAI's Rubik's cube [2]



ANYmal quadruped [3]

[1] Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *JMLR*.
[2] Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., ... & Zhang, L. (2019). Solving rubik's cube with a robot hand. arXiv:1910.07113.
[3] Rudin, N., Hoeller, D., Reist, P., & Hutter, M. (2022). Learning to walk in minutes using massively parallel deep reinforcement learning. In CoRL..

Can CL generate a sequence of tasks
to improve safety during training
and
speed up learning for constrained RL?

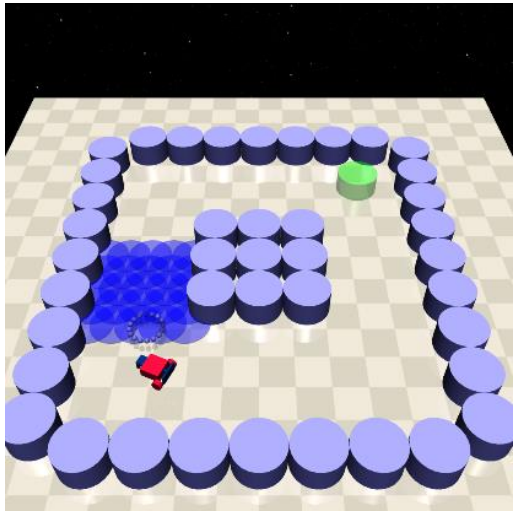
Contextual Constrained RL

Contextual Constrained MDP

$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{X}, \mathrm{M}, D, \gamma \rangle$

Context Space — Safety threshold

From contexts to constrained MDPs



Safety-goal

Optimal policy

Given a target context distribution φ

$$\pi^* \in \underset{\pi}{\operatorname{argmax}} \ \mathbb{E}_{\varphi}[V_r^{\pi}(\mathbf{x})]$$
$$\text{s.t.} \ \ \mathbb{E}_{\varphi}[V_c^{\pi}(\mathbf{x})] \leq D.$$

Our objective

Generate a sequence of contexts distributions $\{\varrho_l\}_{l=0}^{L,}$ that

1) accelerate learning an optimal policy, and
2) improve safety via reducing constraint violation regret

$$\mathrm{Reg}^{\mathrm{tr}}\left(\{\varrho_l\}_{l=0}^{L}, D\right) = \sum_{l=0}^{L} \left[\mathbb{E}_{\varrho_l}\left[V_c^{\pi_l}(\mathbf{x})\right] - D\right]_+$$

where $[y]_+ = \max\{y, 0\}$.

Failure of CL methods

They overlook the cost constraint!

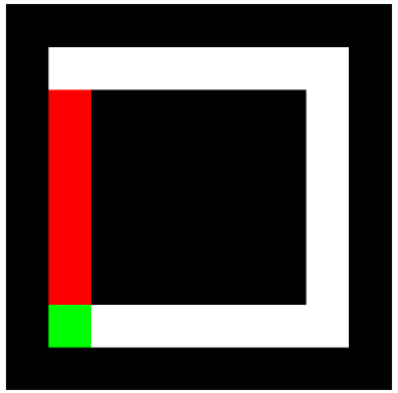
They prioritize contexts with high rewards, but also high costs

causing

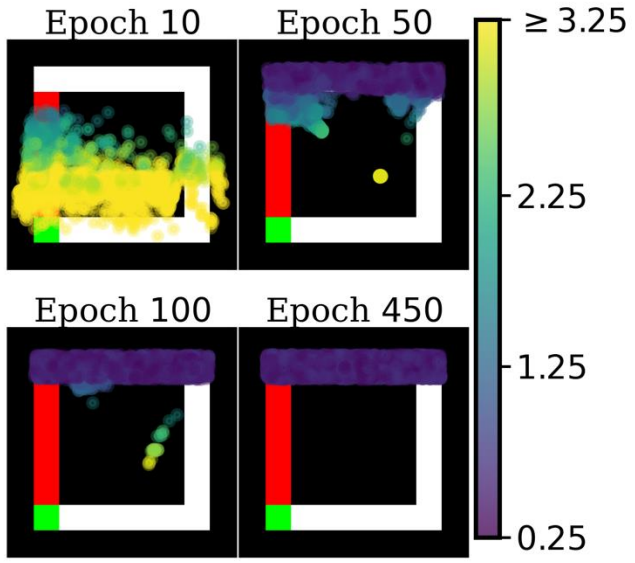
1) Constraint violations during training

2) Suboptimal behavior at the end

An example: Safety-maze

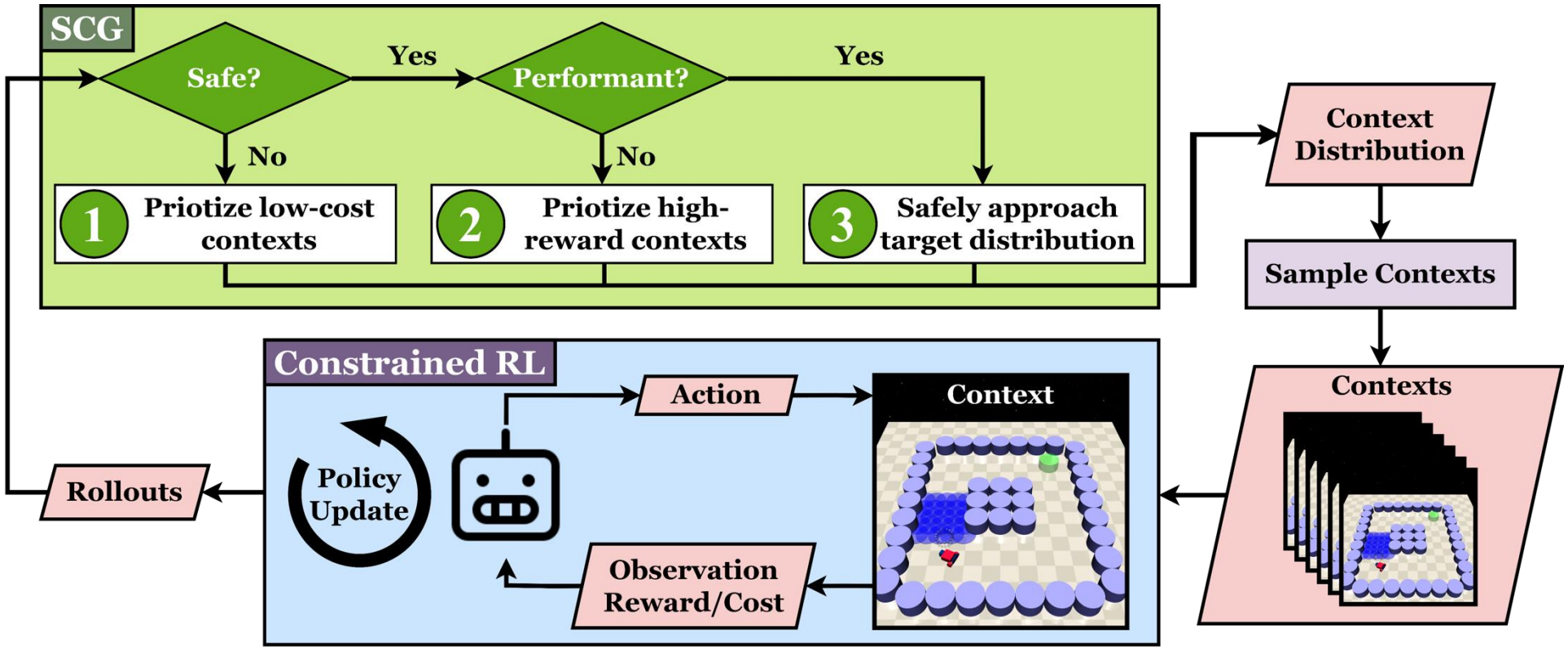


Curriculum progression of CURROT [4]



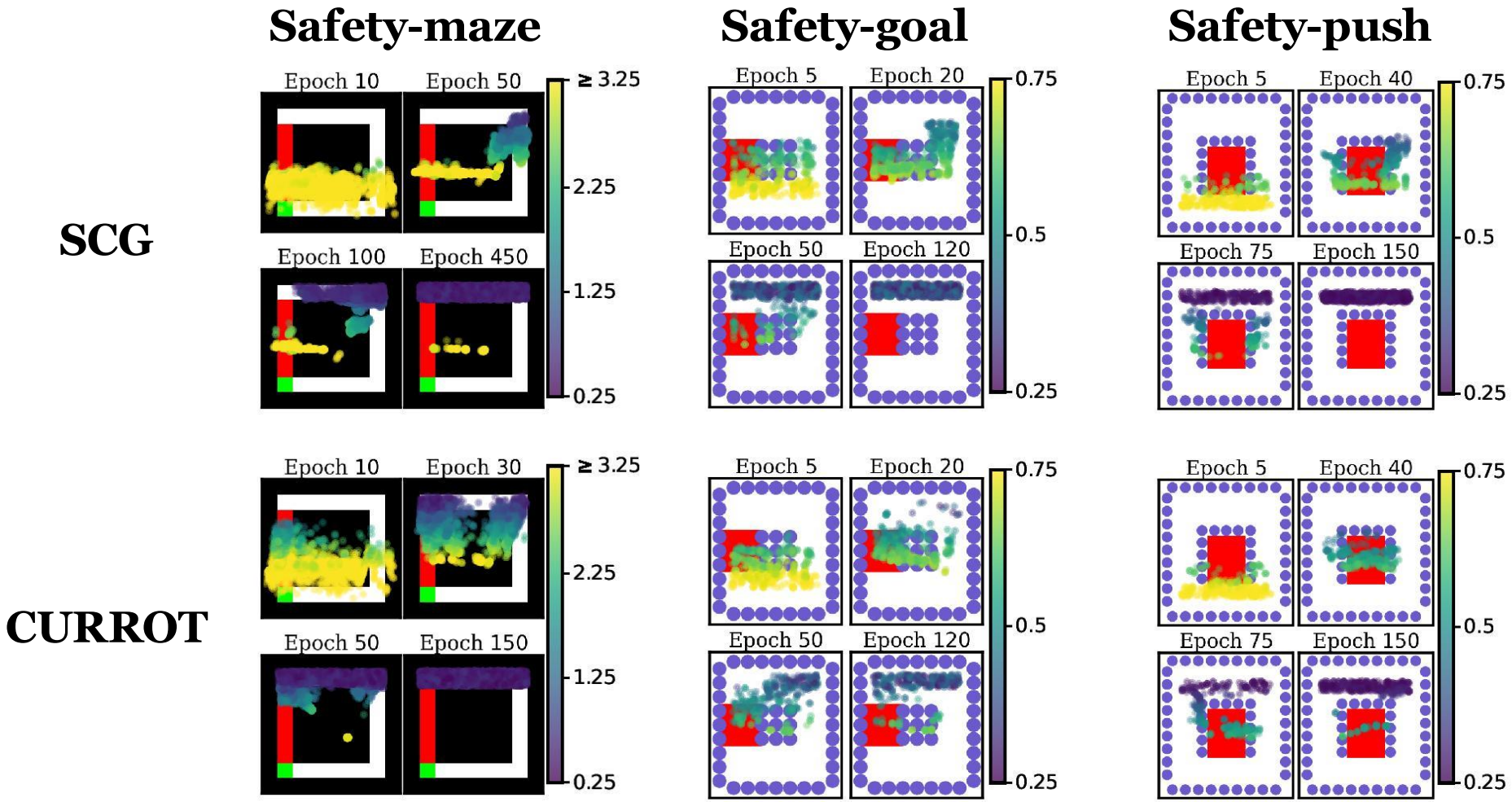
[4] Klink, P., Yang, H., D'Eramo, C., Peters, J., & Pajarinen, J. (2022, June). Curriculum reinforcement learning via constrained optimal transport. In International Conference on Machine Learning (pp. 11341-11358). PMLR.

Safe Curriculum Generation



Curricula Progression

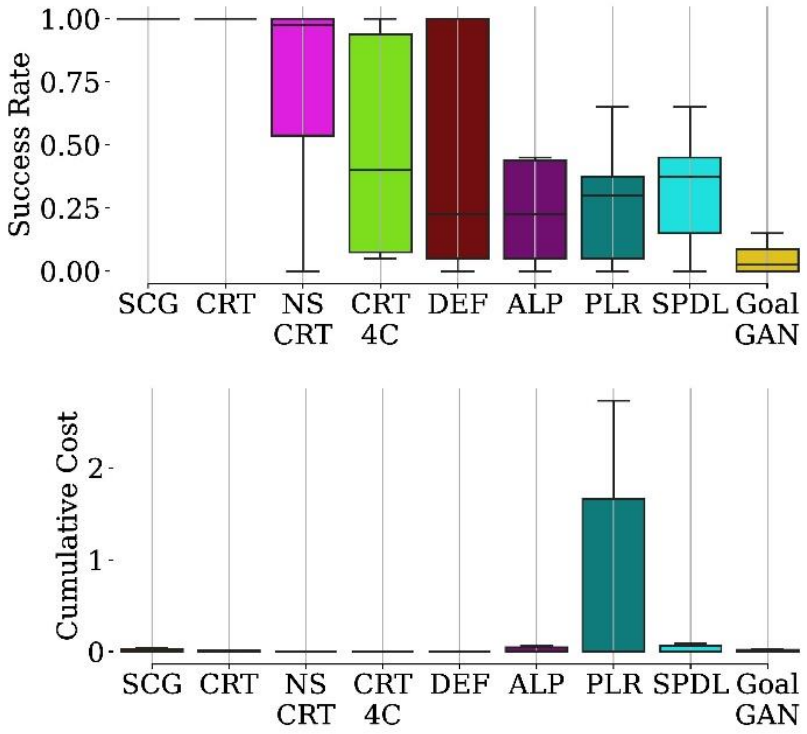
TLDR: SCG identifies safe contexts early on, whereas **CURROT** fails to avoid hazards, causing high CV regret.



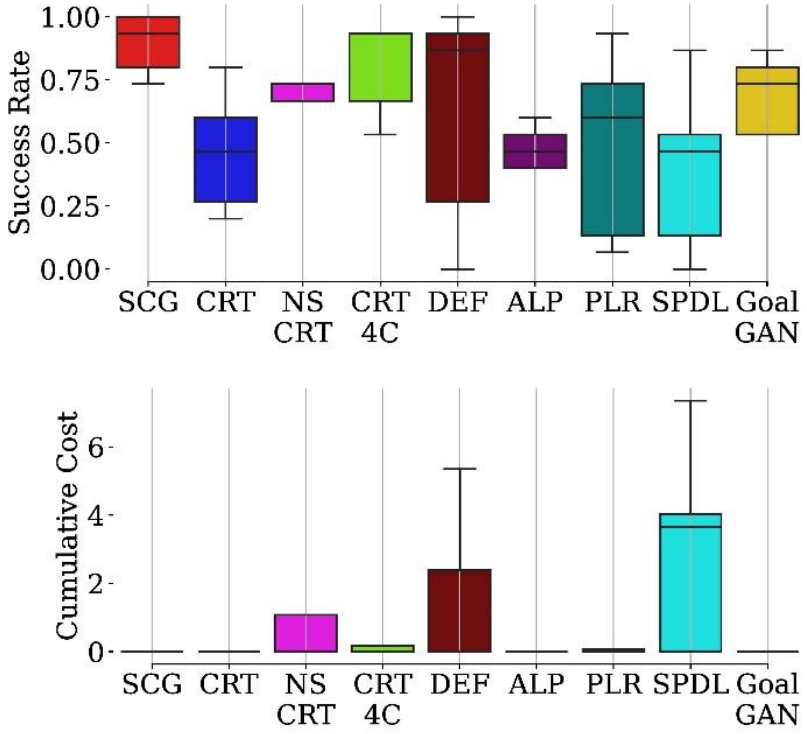
Learning Optimal Policies

TLDR: SCG learns policies that **achieve zero cost** in target contexts, which satisfies the cost constraint, and **highest success rates**.

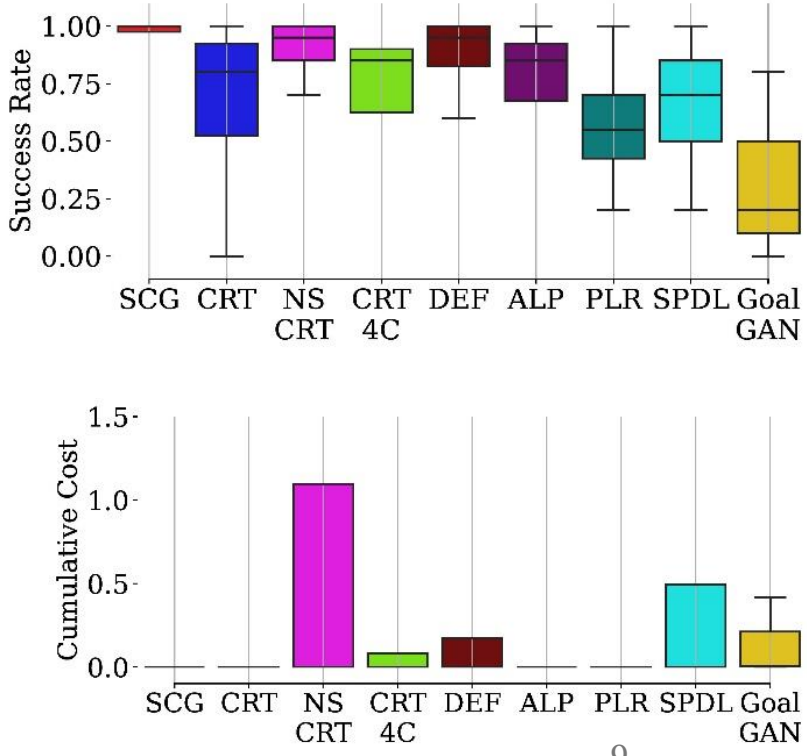
Safety-maze



Safety-goal



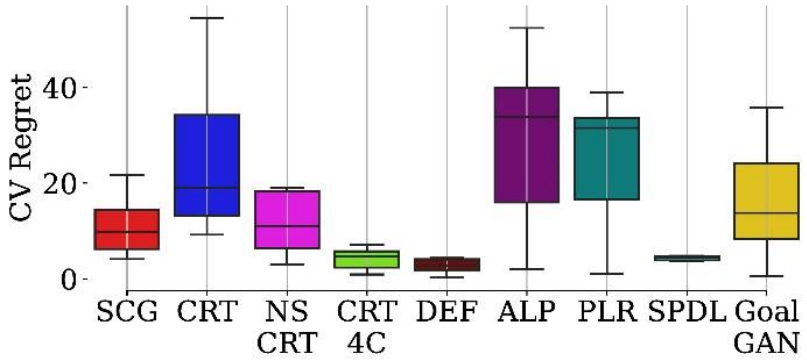
Safety-push



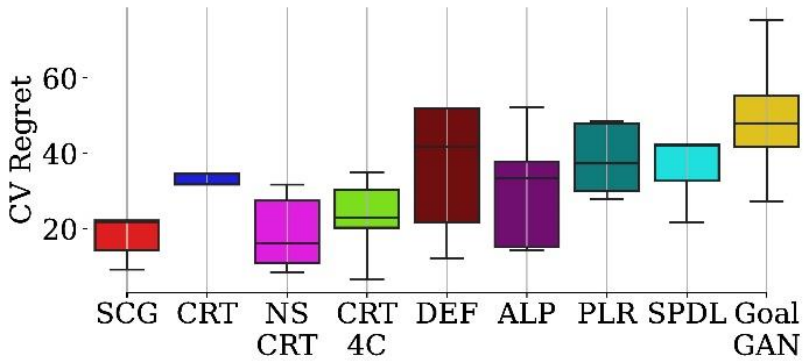
Reducing constraint violations

TLDR: Among the approaches that learn optimal policies, SCG achieves the lowest CV regret.

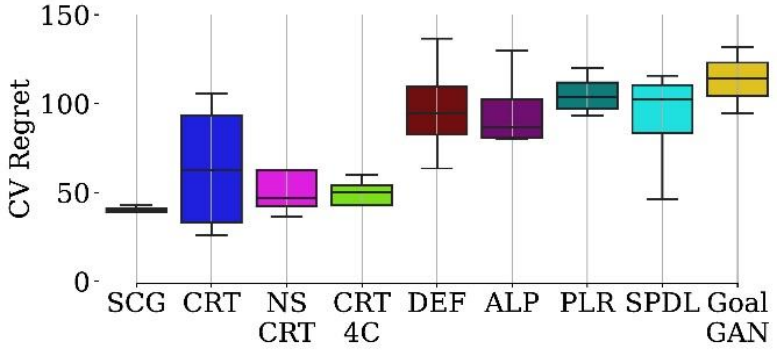
Safety-maze



Safety-goal



Safety-push



Thank you!

Cevahir Koprulu
Email: cevahir.koprulu@utexas.edu
Website: https://cevahir-koprulu.github.io/