



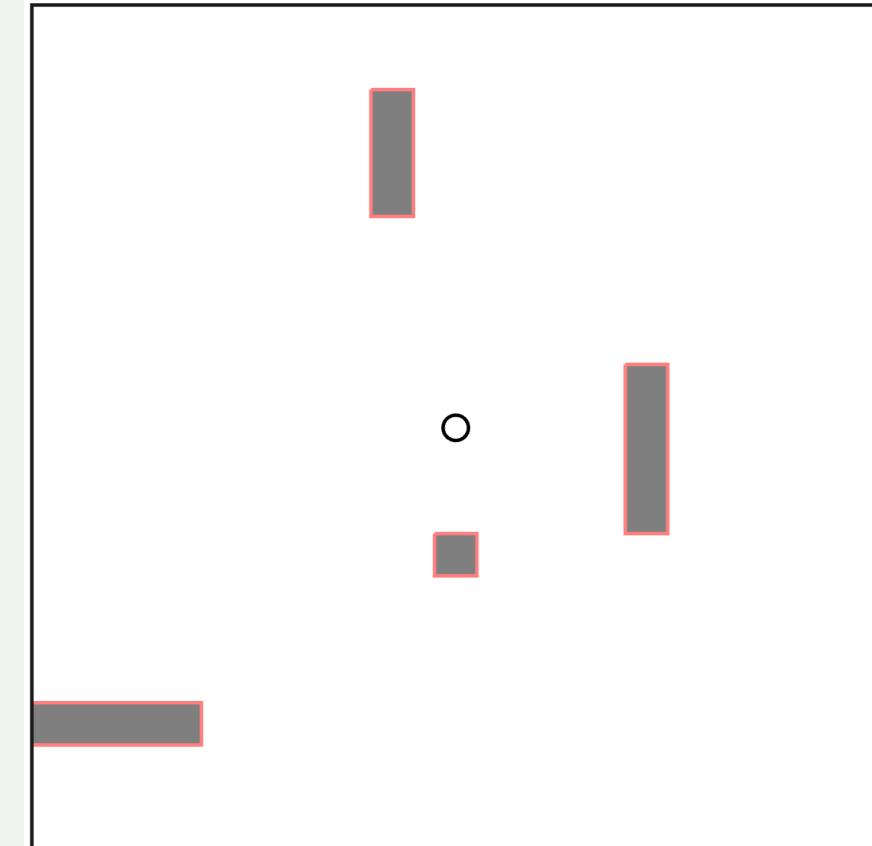
OCTOBER 1 - 5, 2023

IEEE/RSJ International Conference
on Intelligent Robots and Systems

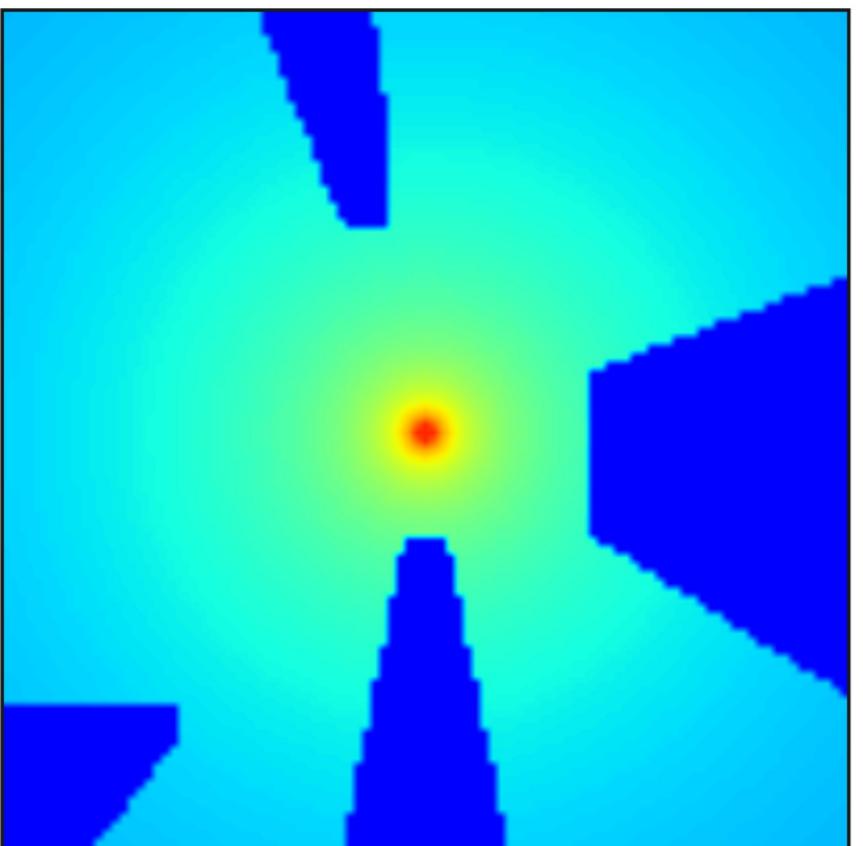
A Multiplicative Value Function for Safe and Efficient Reinforcement Learning

Nick Bührer¹, Zhejun Zhang¹, Alexander Liniger¹, Fisher Yu¹, Luc Van Gool^{1,2}.¹ Computer Vision Lab, ETH Zurich, Switzerland. ² PSI, KU Leuven, Belgium.

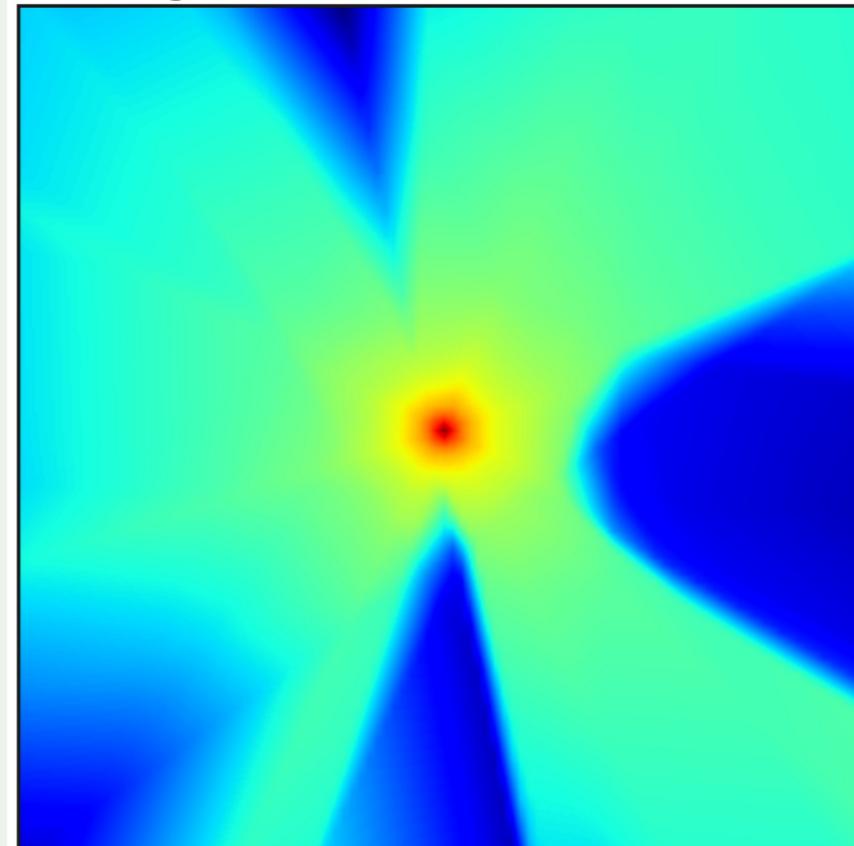
Motivation



Environment with Constraints

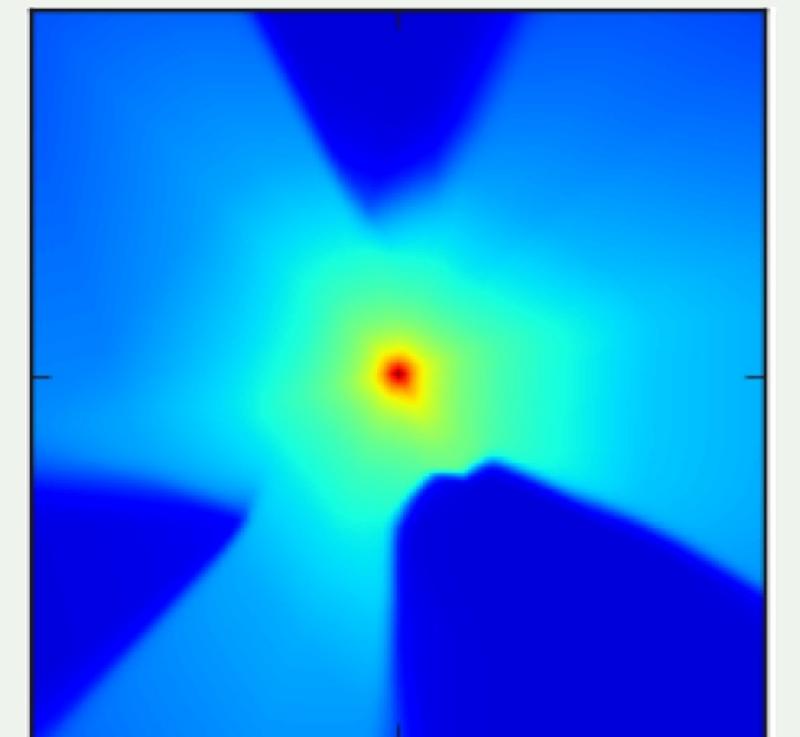


Ground-Truth Value

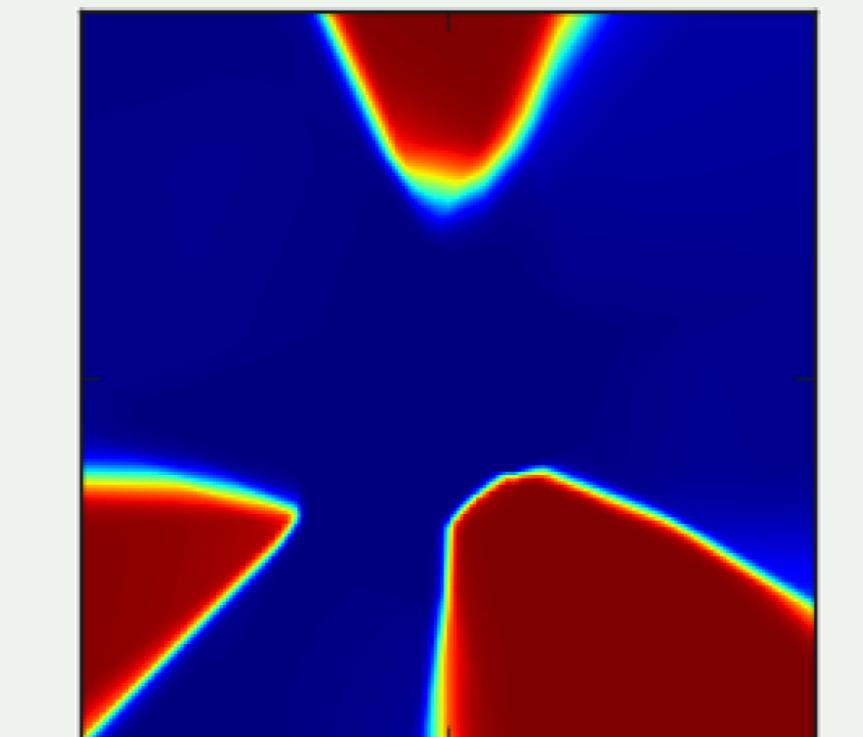


Regular Value Function

- Sharp discontinuities are hard to learn.
- Poor training stability and sample efficiency.



Multiplicative Value Function



Probabilistic Safety Critic



Constraint-Free Reward Critic

Method

Multiplicative Value/Q Function:

$$V^{\pi}_{\text{mult}}(s) = (\bar{V}^{\pi}(s) - \bar{v}_{\min}) \cdot (1 - \Phi^{\pi}(s)) + \bar{v}_{\min}$$

$$Q^{\pi}_{\text{mult}}(s, a) = (\bar{Q}^{\pi}(s, a) - \bar{q}_{\min}) \cdot (1 - \Psi^{\pi}(s, a)) + \bar{q}_{\min},$$

$$\bar{v}_{\min} := \min_s \bar{V}^{\pi}(s), \bar{q}_{\min} := \min_{s,a} \bar{Q}^{\pi}(s, a)$$

Multiplicative Advantage:

(V1) Bootstrap Q:

$$A^{\pi}_{\text{mult}}(s_t, a_t) = [\bar{r}_t + \gamma V^{\pi}_{\text{mult}}(s_{t+1})] - V^{\pi}_{\text{mult}}(s_t)$$

(V2) W/O bootstrap: $A^{\pi}_{\text{mult}}(s_t, a_t) = Q^{\pi}_{\text{mult}}(s_t, a_t) - V^{\pi}_{\text{mult}}(s_t)$ (V3) Bootstrap the safety critic inside $Q^{\pi}_{\text{mult}}(s_t, a_t)$:

$$(\bar{Q}^{\pi}(s_t, a_t) - \bar{q}_{\min}) \cdot (1 - (r_{c,t} + \gamma_c \Phi^{\pi}(s_{t+1}))) + \bar{q}_{\min}$$

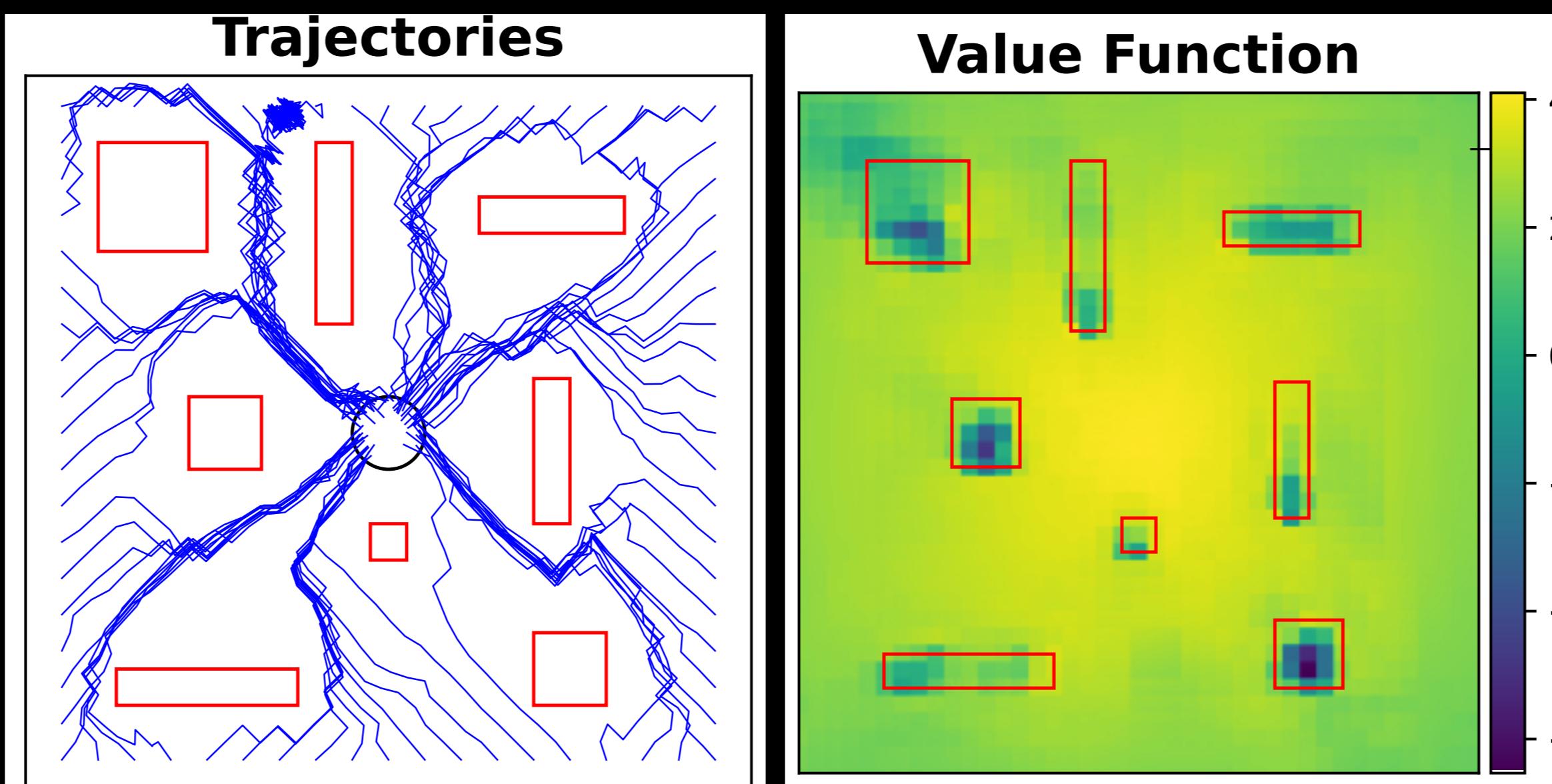
Apply to SAC and PPO

$$\text{SAC: } \max_{\theta} \mathbb{E}_{a_{\theta} \sim \pi_{\theta}} [Q^{\pi_{\theta}}_{\text{mult}}(s, a_{\theta}) - \alpha \log \pi_{\theta}(s_{\theta} | x)]$$

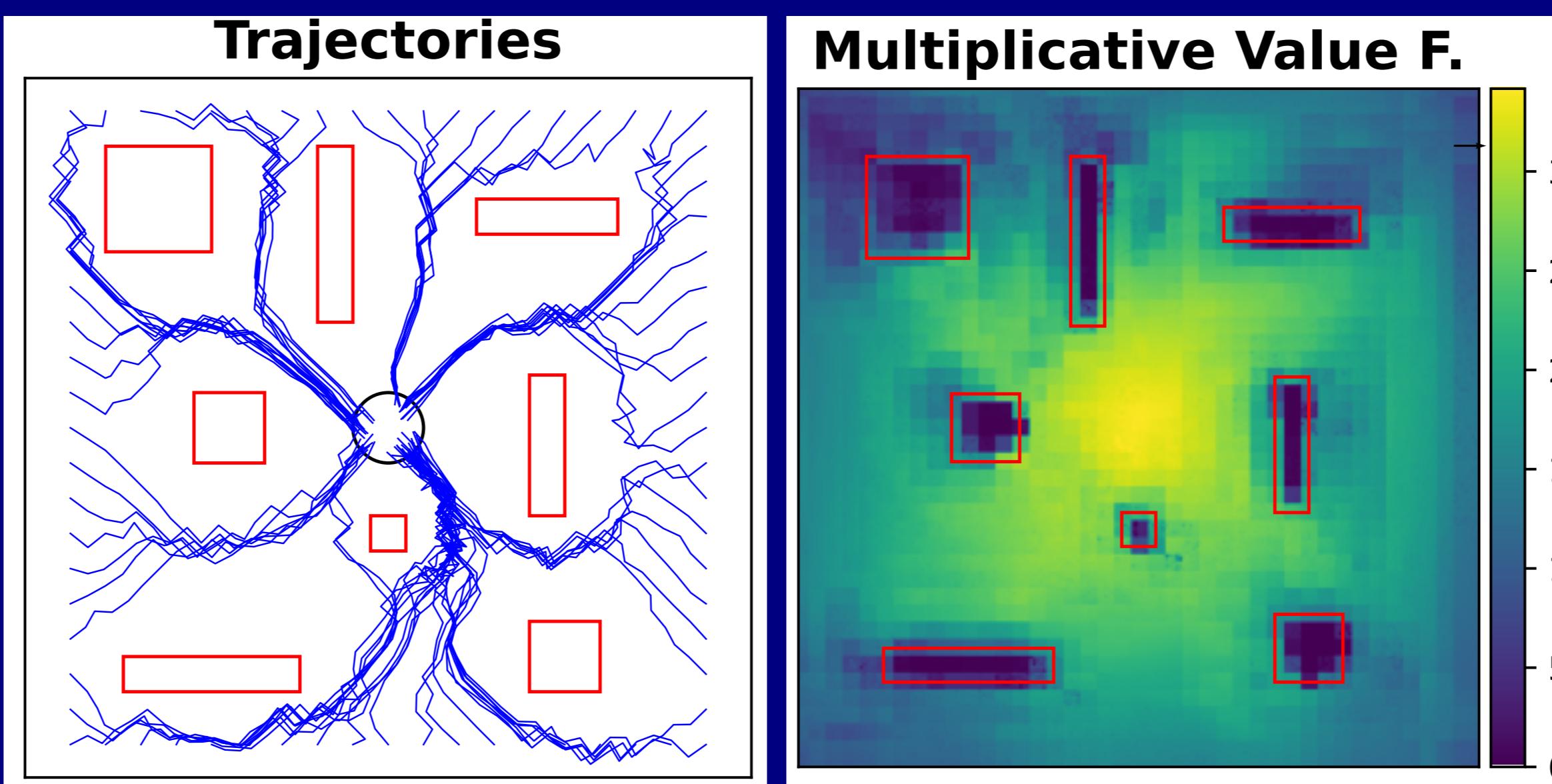
$$\text{PPO: } \max_{\theta} \mathbb{E}_{a \sim \pi_{\theta}} \left[\min \left\{ \frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}_{\text{mult}}(s, a), g(\epsilon, A^{\pi_{\theta_k}}_{\text{mult}}(s, a)) \right\} \right]$$

Results

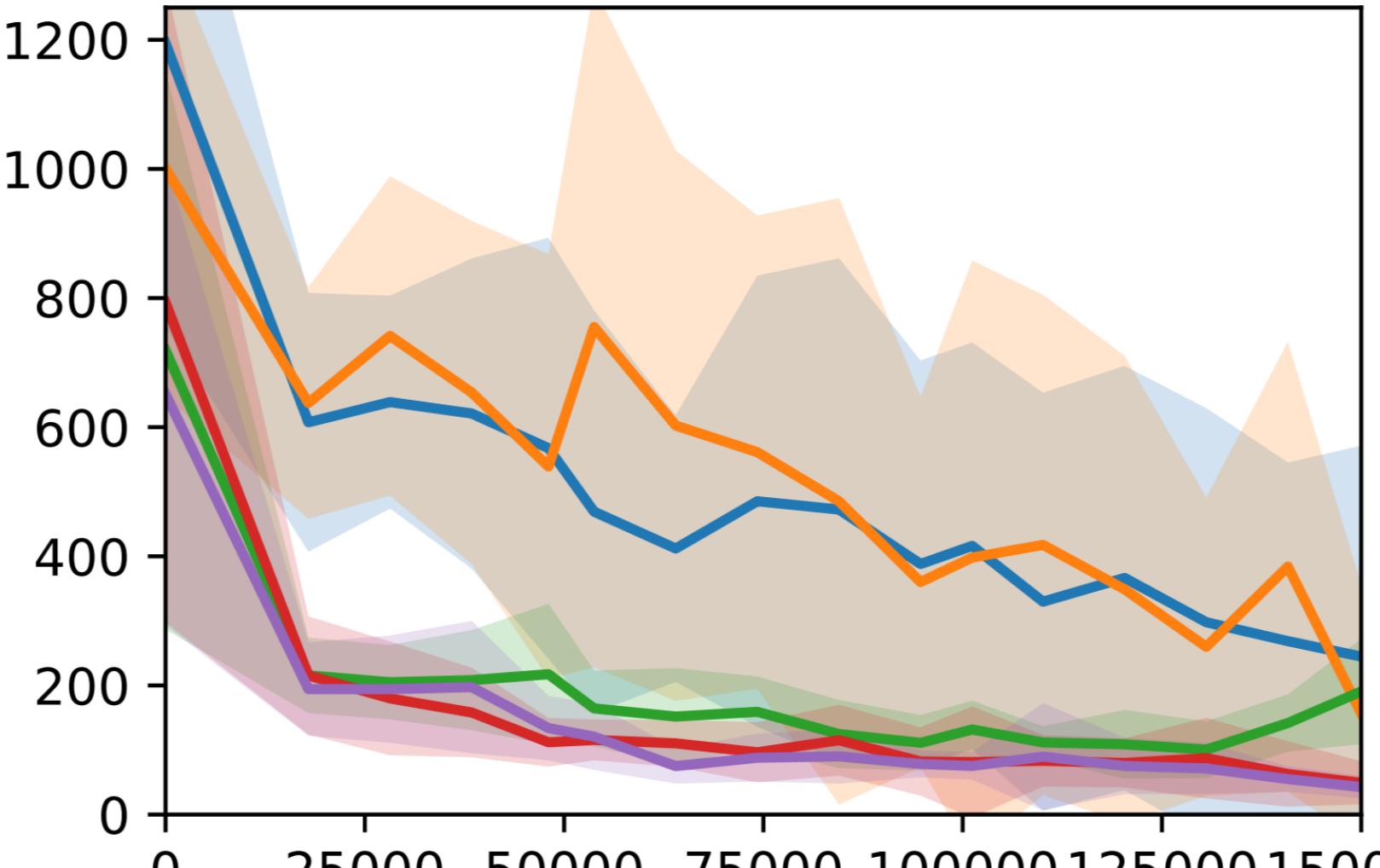
SAC Lagrange



SAC Mult Clipped



Value Loss for Lunar Lander



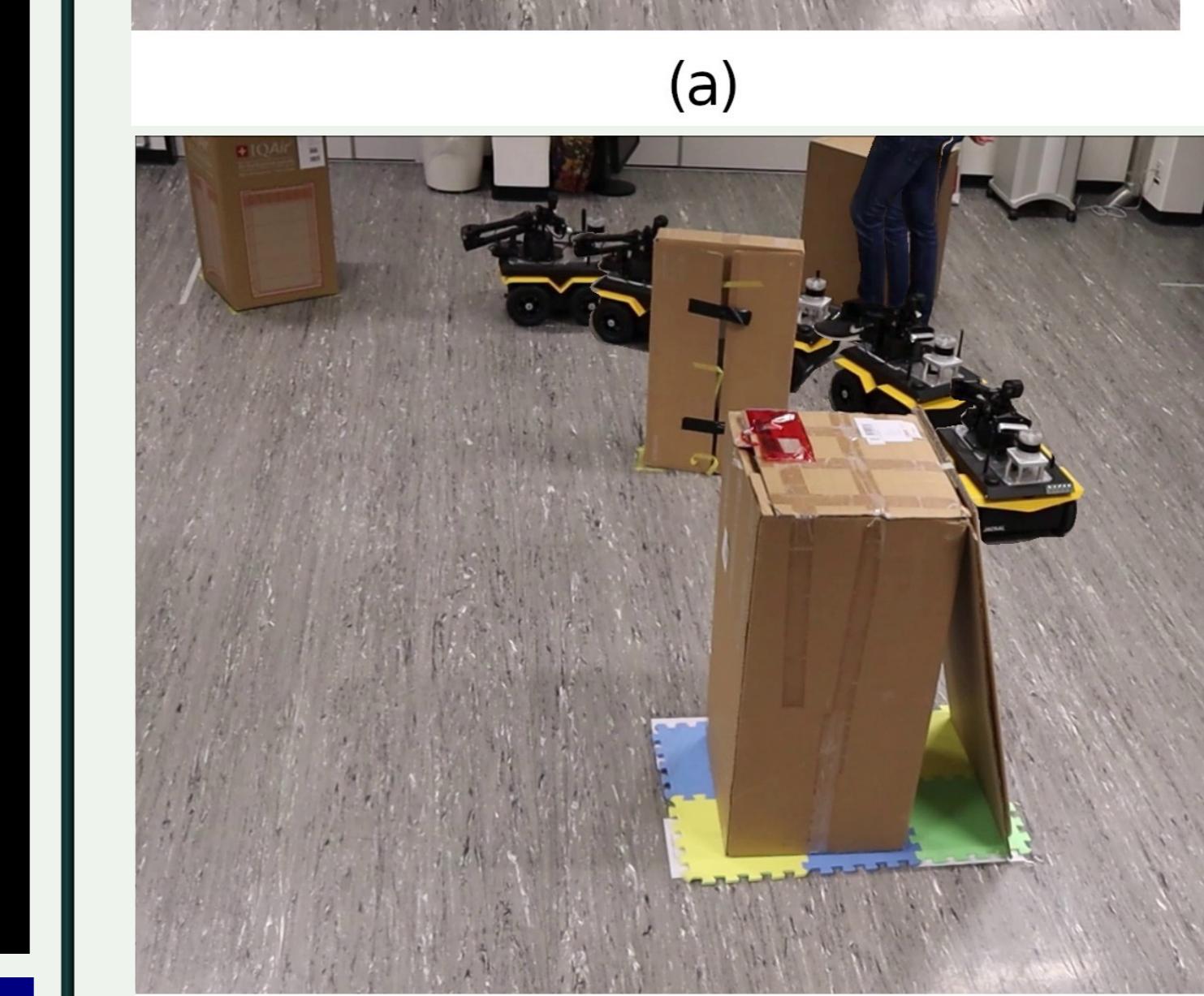
Improved

- Training stability
- Sample efficiency
- Value matching to the obstacles

Real-World Experiments



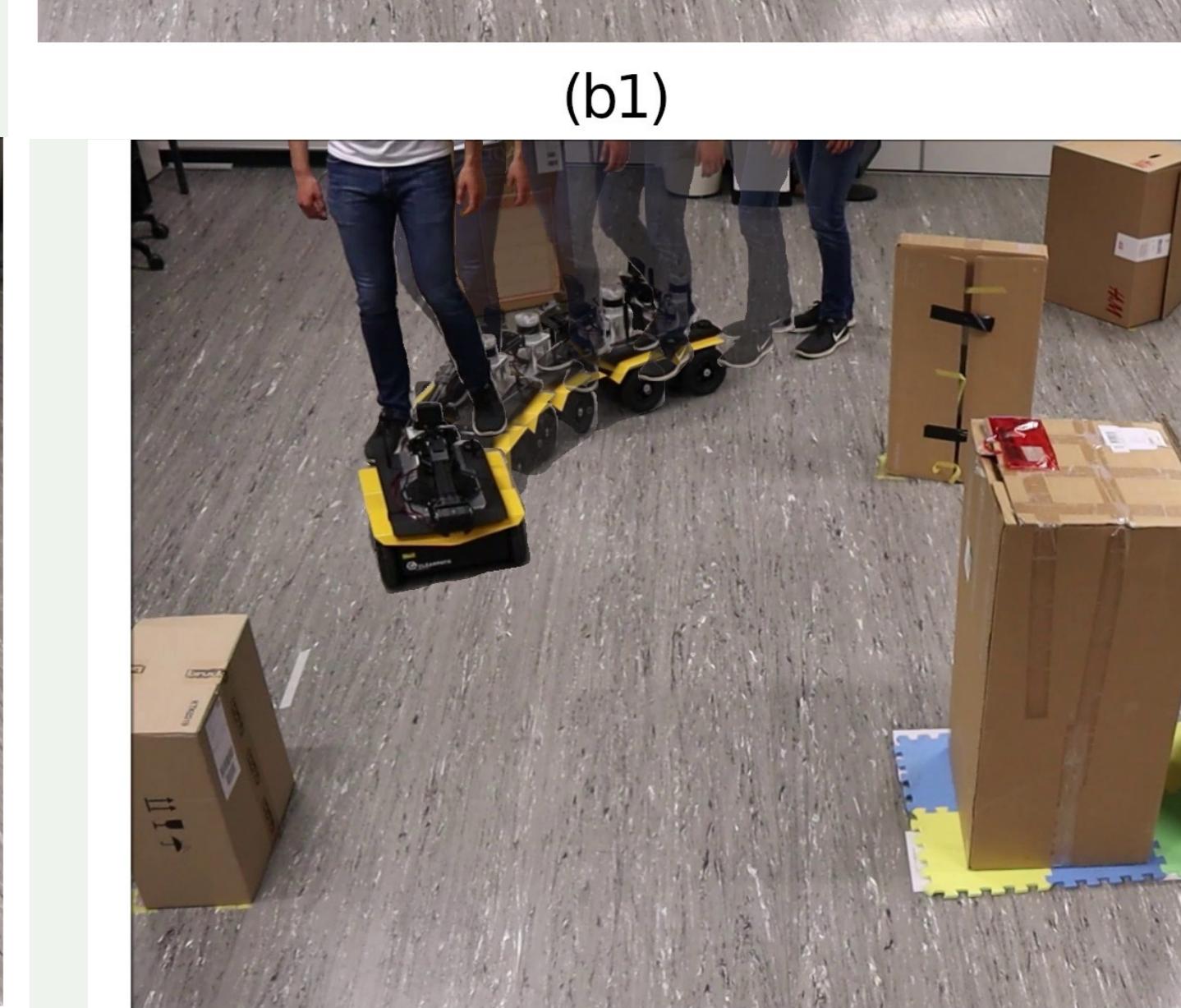
(a)



(b2)



(b1)



(c)

- Differential drive robot with 1D-Lidar.
- Training in Gazebo Simulation.
- Zero-Shot Sim-to-Real.
- Safe Interaction with Dynamic Objects and Human.

Summary

Multiplicative Value Function

- Safety Critic: Binary decision problem.
- Reward Critic: Constraint-free RL.

Integration into SAC and PPO:

- Increased sample efficiency and learning stability.

Future works:

- Theoretical justification
- Code: github.com/nikeke19/Safe-Mult-RL
- Homepage with Videos: zhejz.github.io/saferl