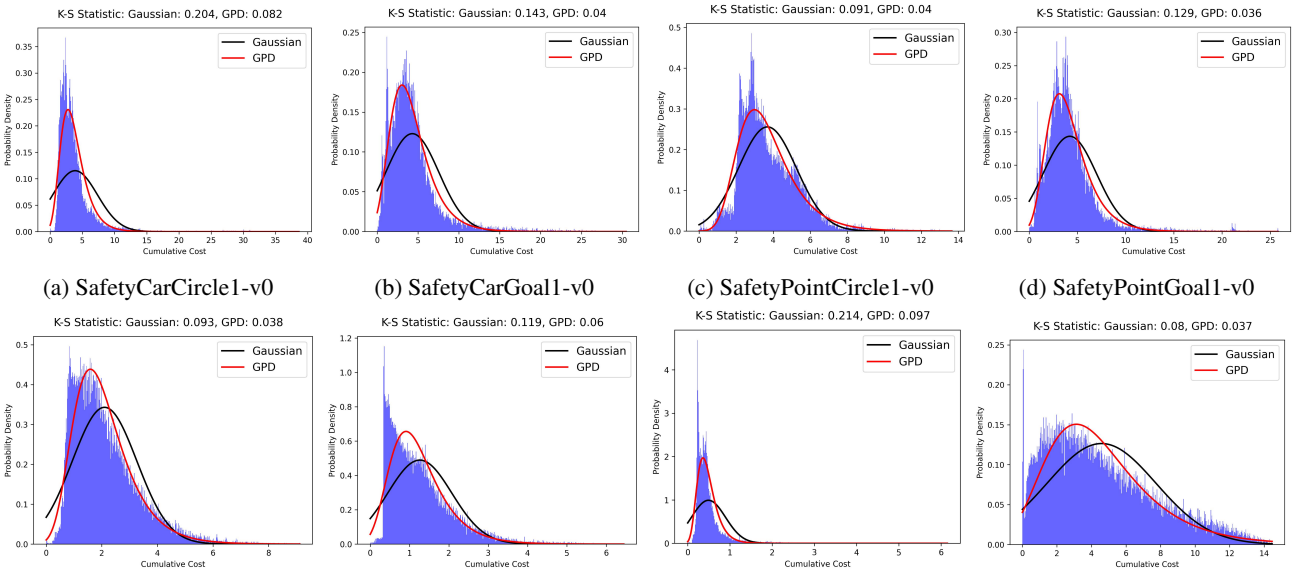

(a) SafetyCarCircle1-v0 (b) SafetyCarGoal1-v0 (c) SafetyPointCircle1-v0 (d) SafetyPointGoal1-v0

(e) SafetyAntVelocity-v1 (f) SafetyHalfCheetahVelocity-v1 (g) SafetyHumanoidVelocity-v1 (h) SafetySwimmerVelocity-v1

Figure 1: The distribution curves fitted by GPD and Gaussian on the training data in EVO. The fitting accuracy is quantified using the Kolmogorov-Smirnov test, as shown in the title, where lower values indicate higher fitting accuracy. GPD presents accurate fitting performance across various data distributions.

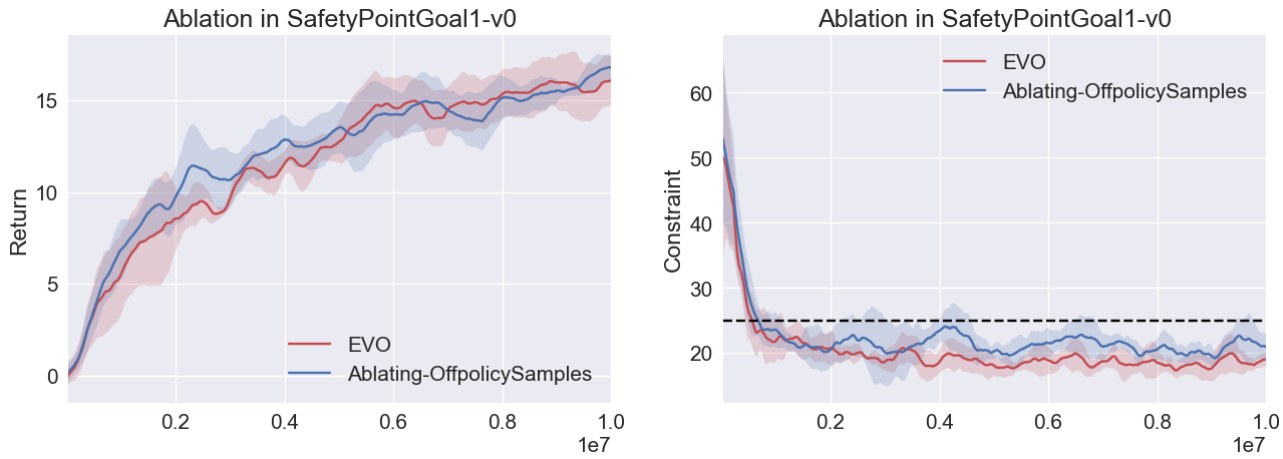

Figure 2: Training curves and convergence speed before and after ablating off-policy resampling in EVO. Constraint satisfaction in EVO converges after 1×10^6 steps.

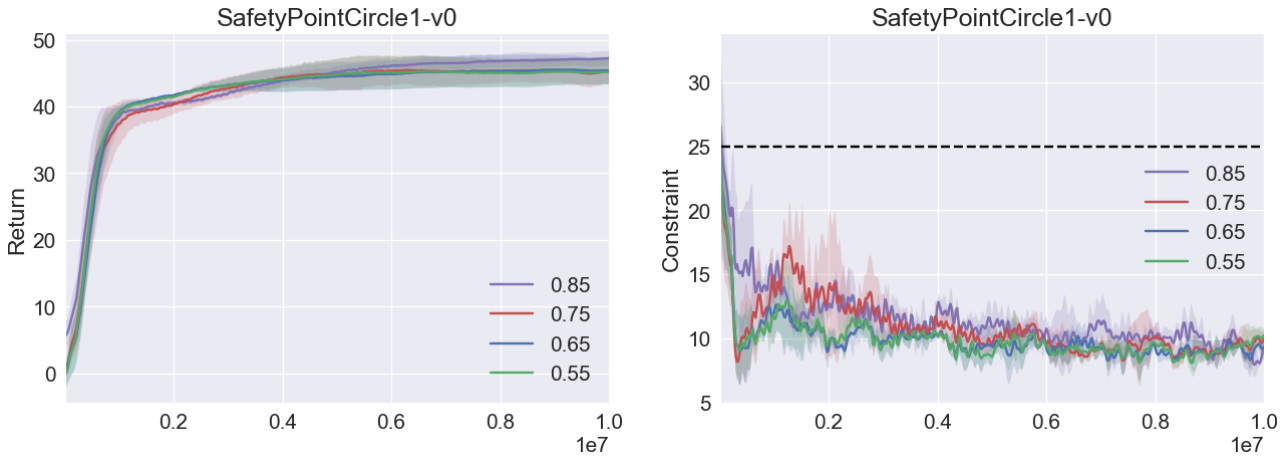

Figure 3: Sensitivity analysis experiments for different ν in EVO. EVO is robust to the initial choice of ν.

Environment	CPO	EVO(10)	EVO(20)	EVO(50)	EVO(100)
SafetyPointCircle1	11h 23m 28s	11h 53m 19s (8s)	11h 52m 28s (8s)	11h 53m 31s (8s)	11h 55m 22s (9s)
SafetyPointGoal1	11h 29m 42s	11h 58m 51s (8s)	11h 59m 47s (7s)	11h 58m 33s (8s)	11h 59 m 8s (9s)

Table 1: EVO training time with different sample sizes compared to CPO. The time data includes the total training time for EVO, with the GPD fitting time shown in parentheses "()". EVO increases limited training time compared to CPO. GPD fitting takes only a few seconds in total.

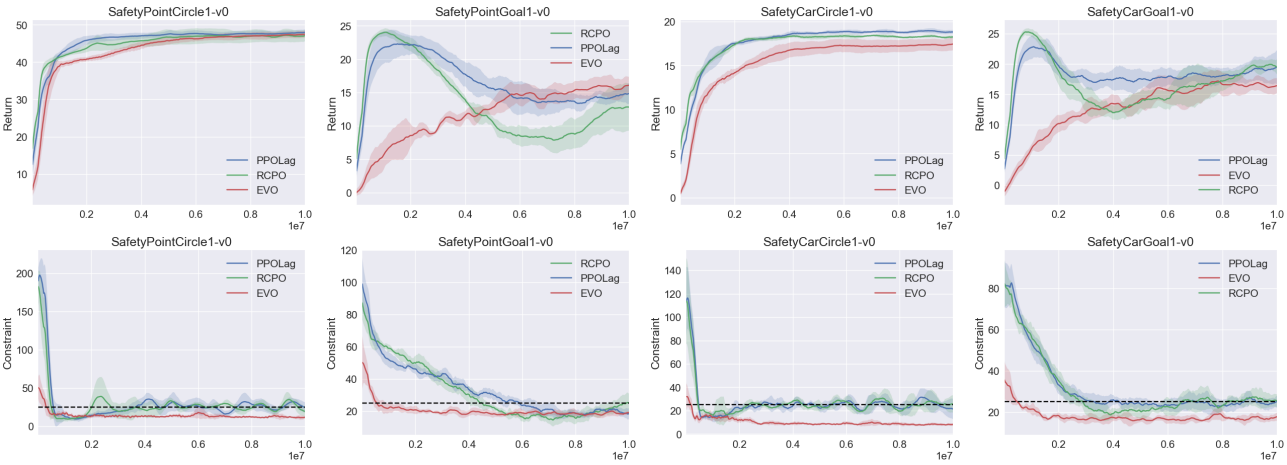

Figure 4: Comparison of EVO to PPO-Lagrangian and RCPO on Safety Gym. The x-axis is the total number of training steps, the y-axis is the average return or constraint. The solid line is the mean and the shaded area is the standard deviation. The dashed line is the constraint threshold which is 25.

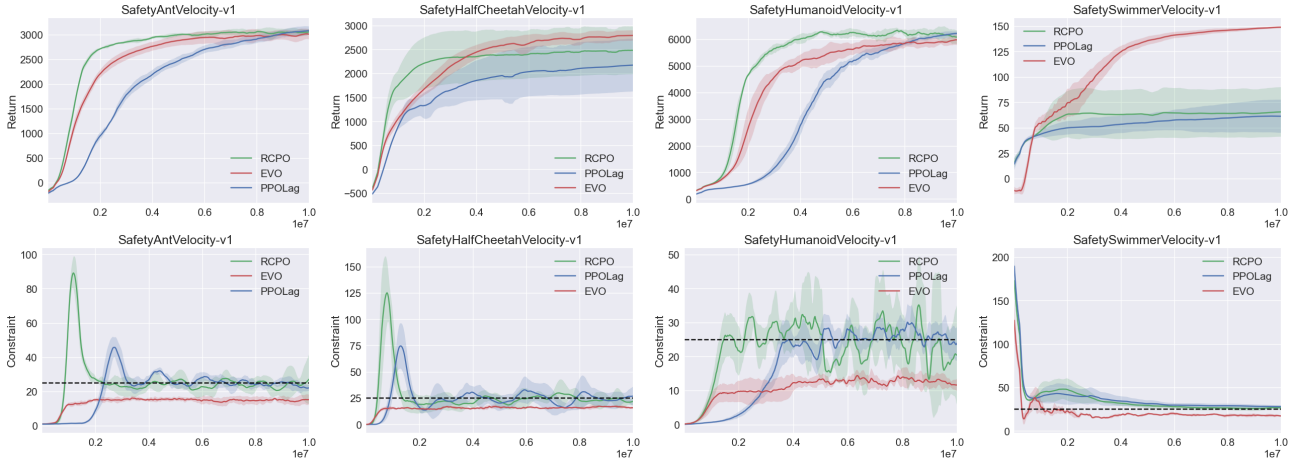

Figure 5: Comparison of EVO to PPO-Lagrangian and RCPO on Safety MuJoCo. The x-axis is the total number of training steps, the y-axis is the average return or constraint.