A Implementation

Codes to reproduce the results can be found at `https://github.com/super864/Natural-Robustness-RL`

B Broader Impact

This study investigates and highlights the potential vulnerabilities of commonly used benchmark RL algorithms to a suite of different white-box and black-box attacks. As such, there is a potential for the results of these study to be used with malicious intent to mount attacks on existing RL algorithms that has been deployed in production. Nonetheless, we believe that the results of this study may also be used as a guideline to select a more robust RL policy or as a stepping stone to developing a more robust RL algorithm. Hence, we truly believe that the benefits of the results of this study will outweigh the potential negative societal impact.

C Additional results on other environments

This section presents the comparison plots for the other environments that were not shown in the main manuscript.

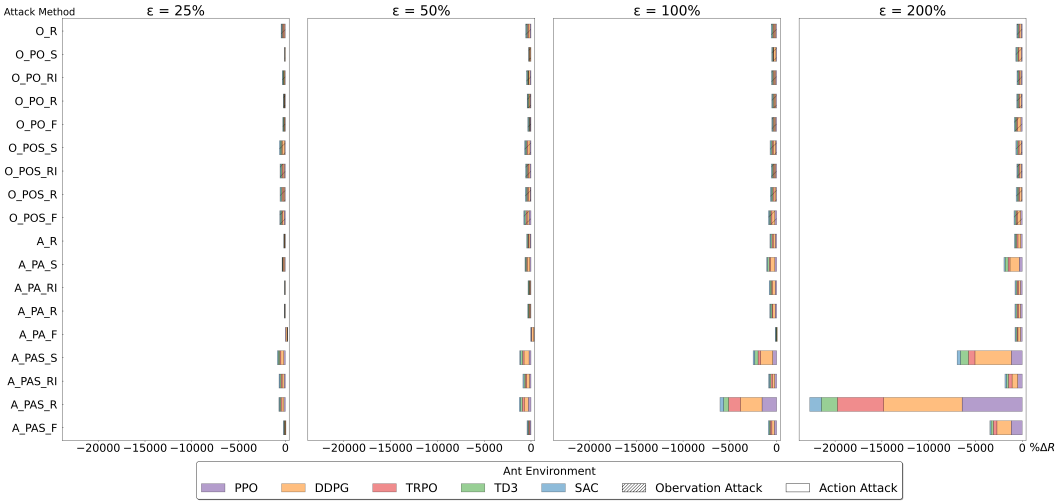

Figure 8: Ant Black-Box attack comparison: All black-box strategies are shown on the y-axis, and the x-axis represents the cumulative $\%\Delta R$ across all RL algorithms. The algorithms are present by the colors. The shaded bar and solid bar show the observation and the action channel. Each subplot represents a particular attack budget ϵ.

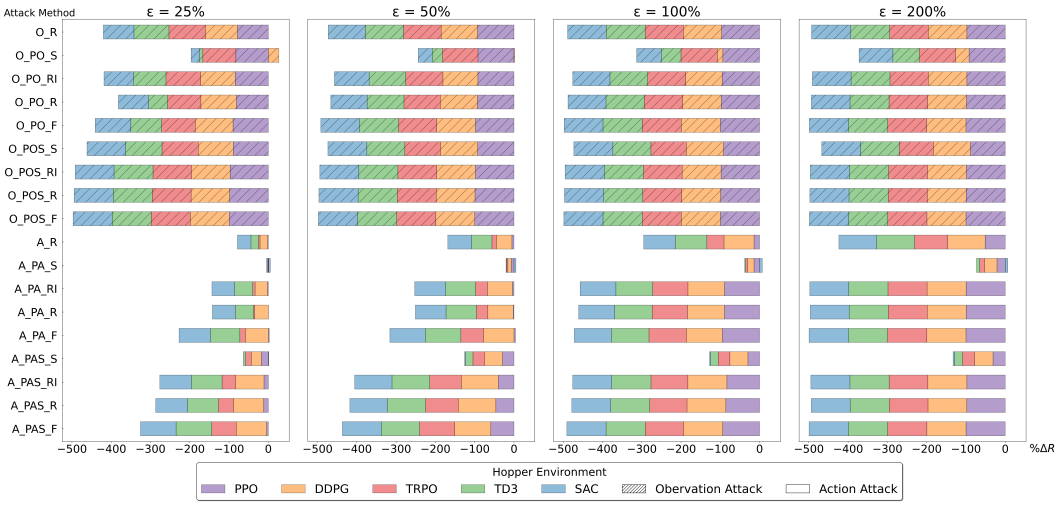

Figure 9: Hopper Black-Box attack comparison

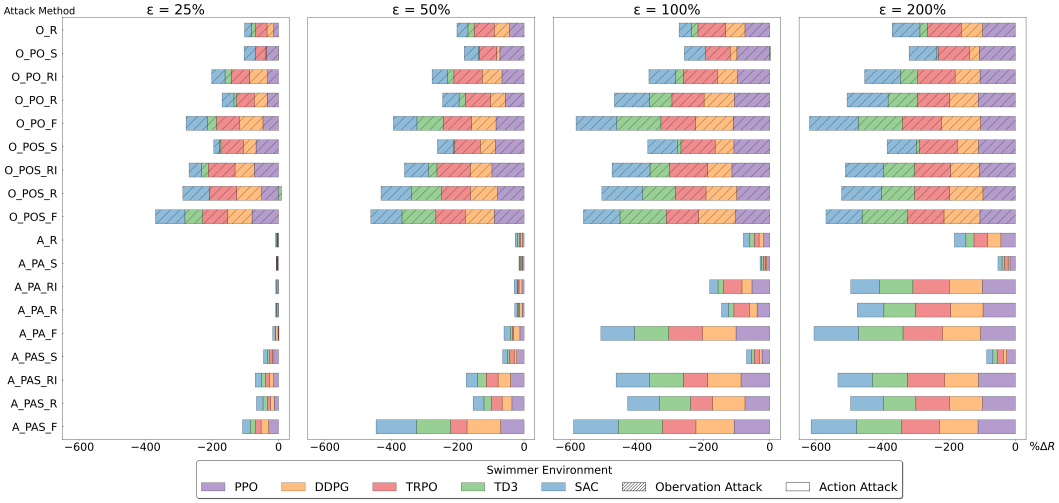

Figure 10: Swimmer Black-Box attack comparison

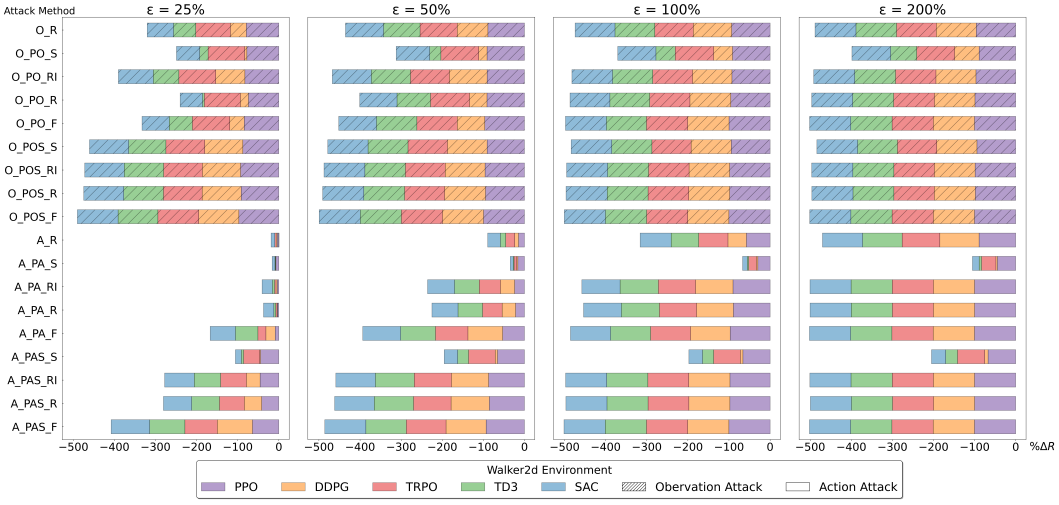

Figure 11: Walker Black-Box attack comparison

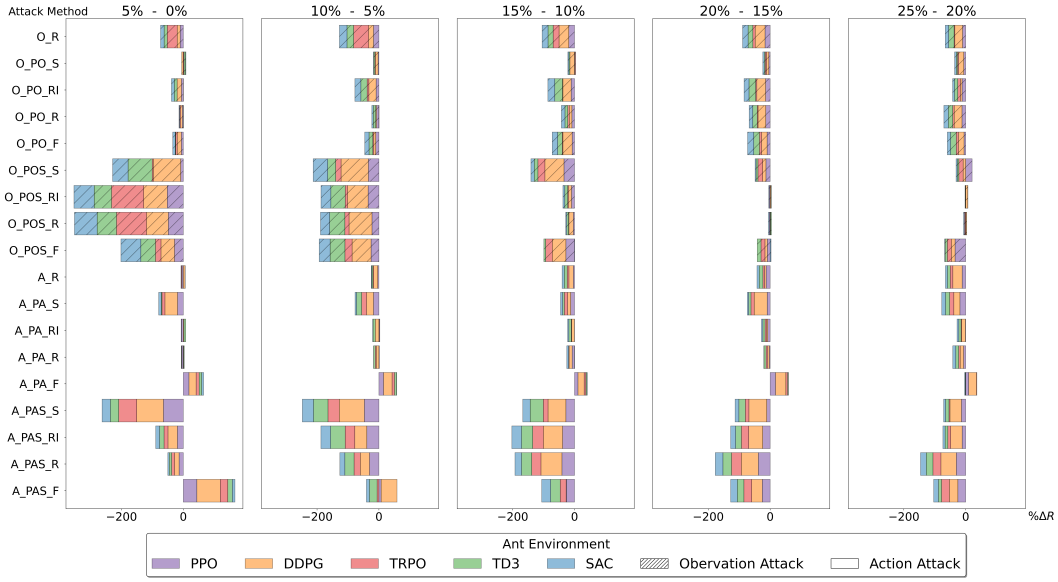

Figure 12: Ant attack differences between percentages:

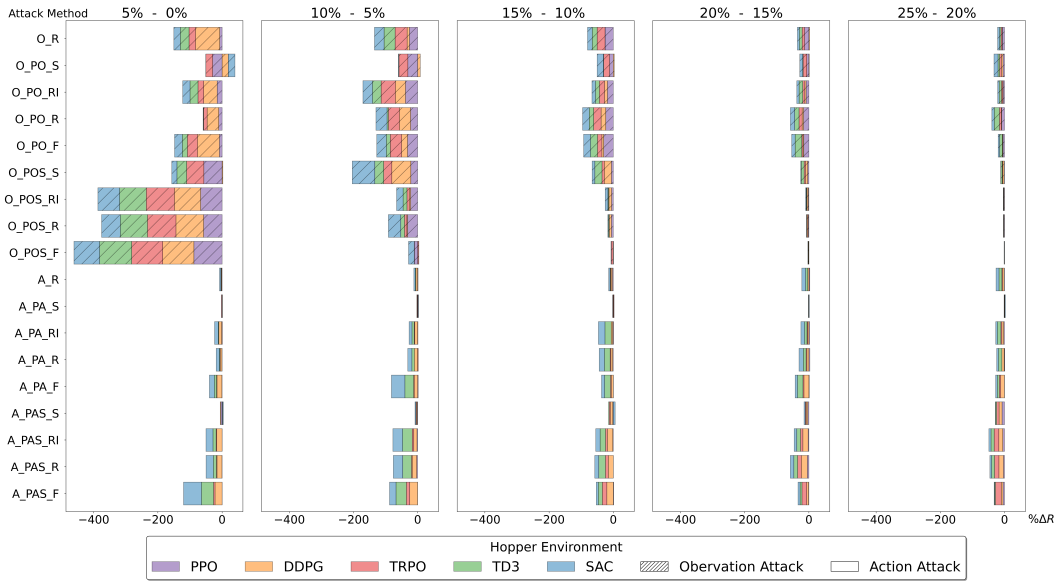

Figure 13: Hopper attack differences between percentages

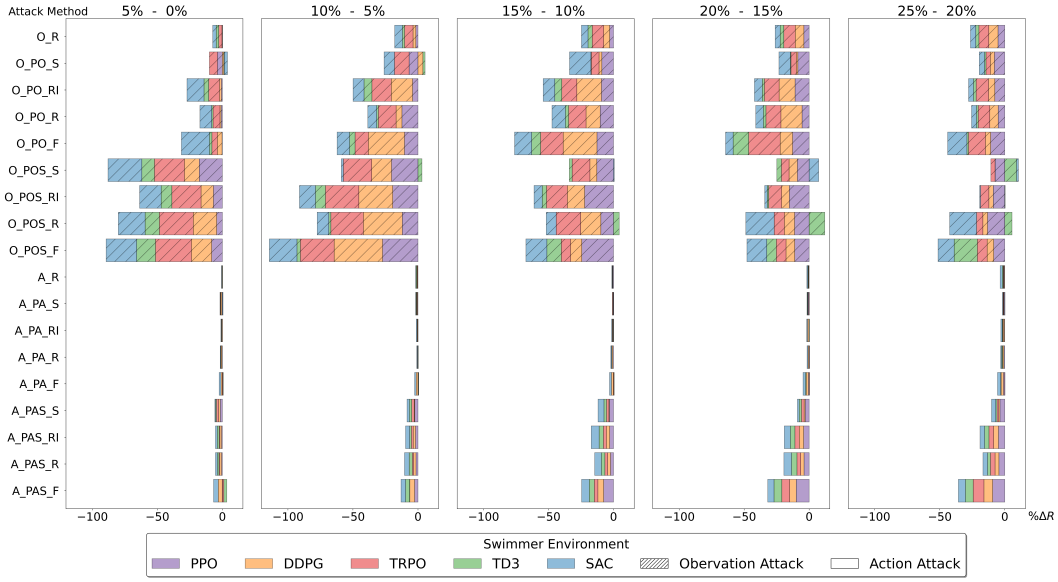

Figure 14: Swimmer attack differences between percentages

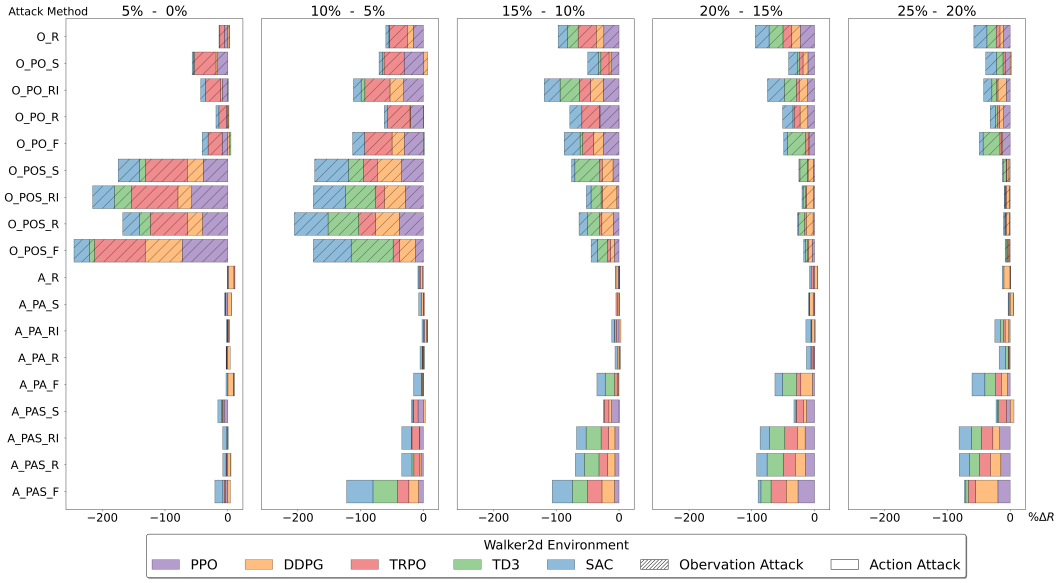

Figure 15: Walker attack differences between percentages