Training Metrics with a CNN Policy Mean Episode Reward Mean Episode Length Critic Loss 1e+15 560 A2C A2C A2C 700 1e+14 SAC SAC SAC 1e+13 480 PPO PPO PPO 1e+12 600 Mean Episode Reward 1e+11 Mean Episode Length 400 1e+10 500 1e+09 1e+08 400 1e+07 1e+06 240 300 1e+05 1e+04 160 200 1e+03 1e+02 80 100 1e+01 1e+00 0 0 0 150000 300000 450000 150000 300000 450000 150000 300000 450000 **Environment Step Environment Step Environment Step Actor Loss Entropy Loss FPS** 1e+16 1e+14 A2C A2C A2C 48 SAC SAC SAC 1.5 1e+12 PPO **PPO** PPO 1e+10 42 1e+08 1e+06 0.0 **Entropy Loss** 1e+04 36 1e+02 FPS -18‡88 -1.530 -1e+02 -1e+04 -3.0-1e+06 24 -1e+08 -1e+10 18 -1e+12 -4.5-1e+14 -1e+16 12 150000 300000 450000 150000 300000 450000 150000 300000 450000 0 **Environment Step Environment Step Environment Step**