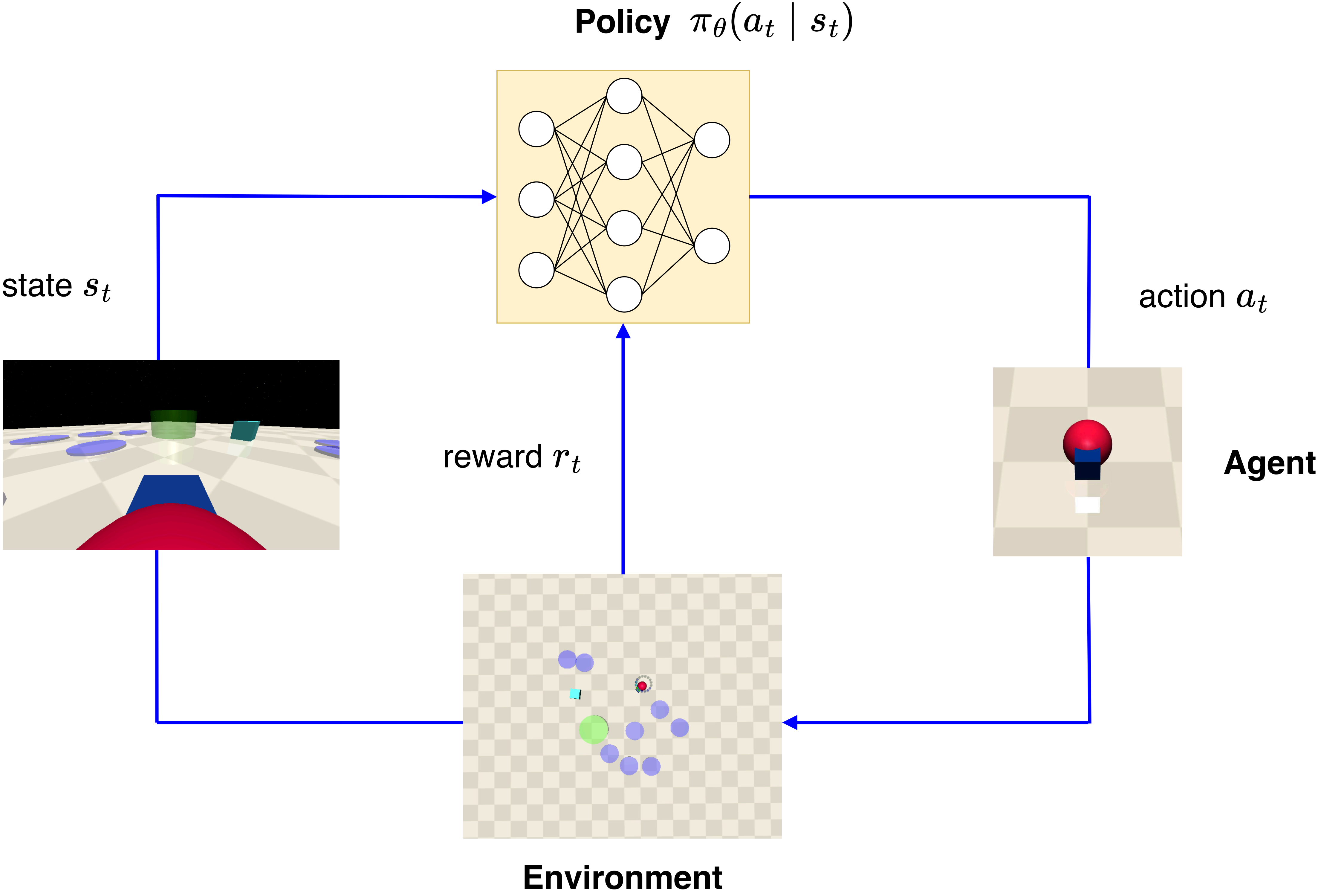
Policy $\pi_\theta(a_t \mid s_t)$
state s_t
action a_t
reward r_t
Agent
Environment