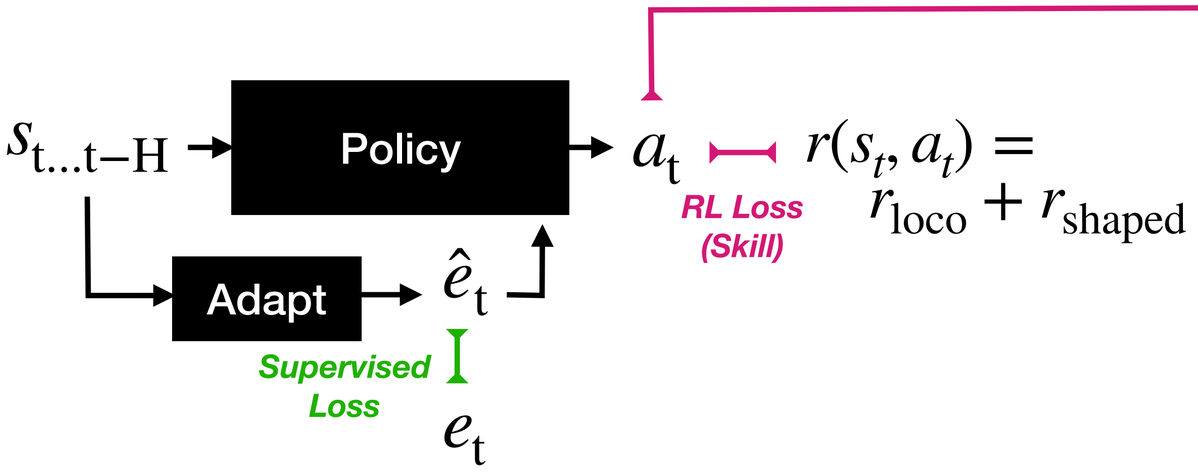
Stage 1: Motor Skill Learning

Policy Optimization with Concurrent State Estimation

$$L(\theta) = \hat{\mathbb{E}}_t \left[\log \pi_\theta(a_t \mid s_t, \hat{e}_t) \hat{A}_t \right] + \|e_t - \hat{e}_t\|^2$$



Stage 2: Adapt for Active Estimation

Policy Optimization with State Estimation Reward + Imitation Loss

$$L(\theta) = \hat{\mathbb{E}}_t \left[\log \pi_\theta^{\text{est}}(a_t \mid s_t, \hat{e}_t) \hat{A}_t^{\text{est}} \right] + \|e_t - \hat{e}_t\|^2 + \beta D_{\text{KL}} \left[\pi_\theta^{\text{est}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t) \right]$$