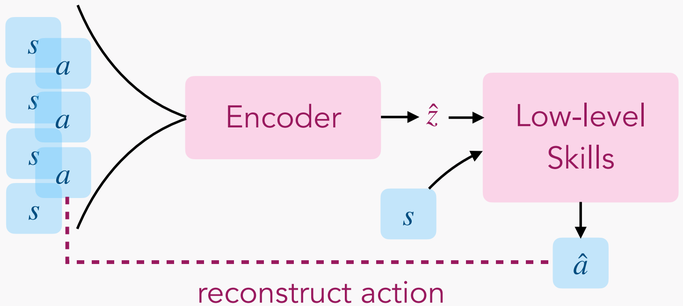
Offline

Online

(1) Extract **low-level task-agnostic behaviors** offline with VAE skill pre-training to reduce task horizon and structure online exploration

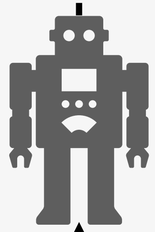


Unlabeled Prior Data

Low-level Skills

pick low-level action to interact with the environment

pick skill latent z as high-level action every H steps



train high-level policy with RL

store online transitions from environment

(s, z, r, s')

Labeled Off-Policy Data

(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** to encourage online exploration

$(s, \hat{z}, r^{+}(s, \hat{z}), s')$

estimated skill latent

optimistic reward label

(RND + online reward model)