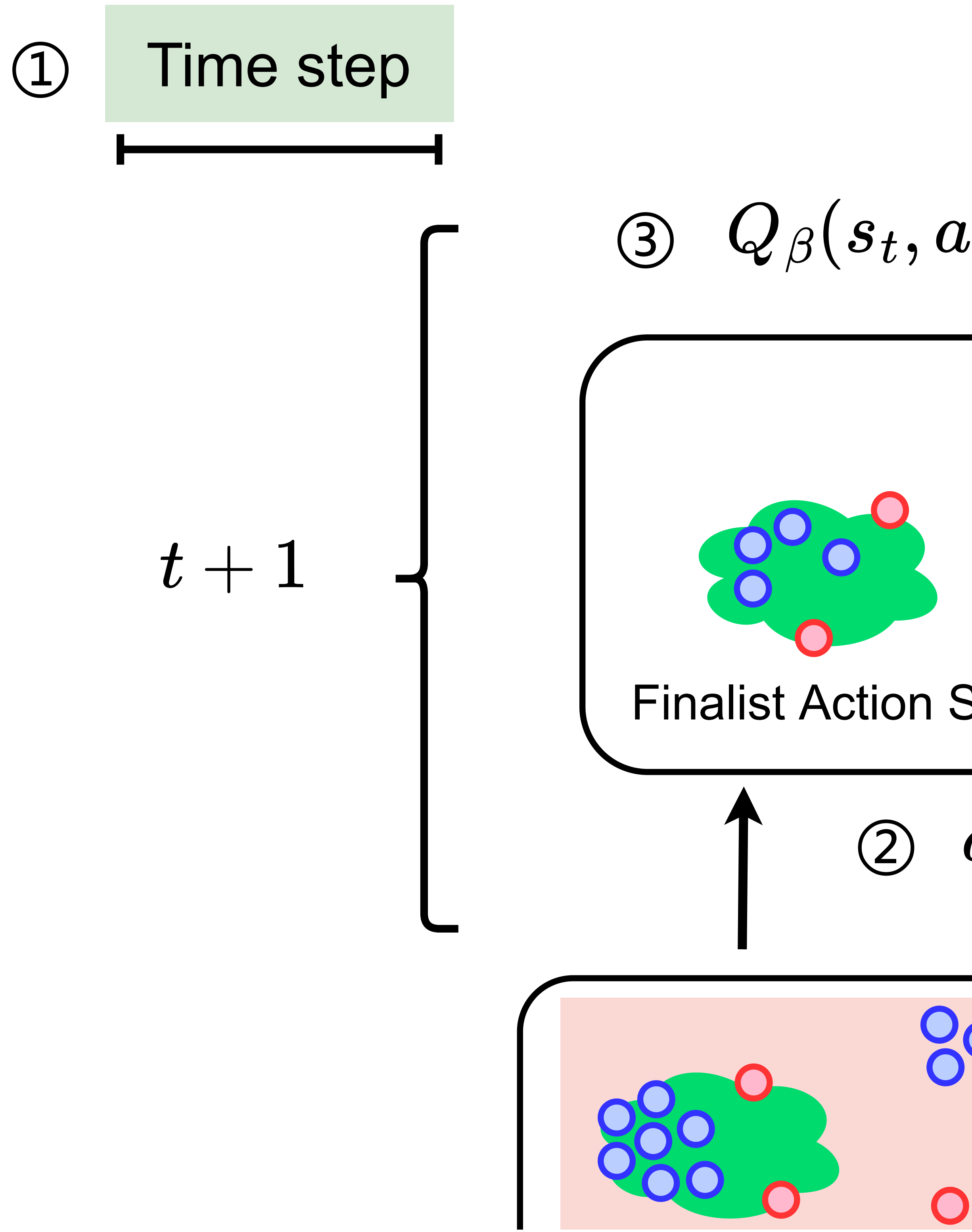
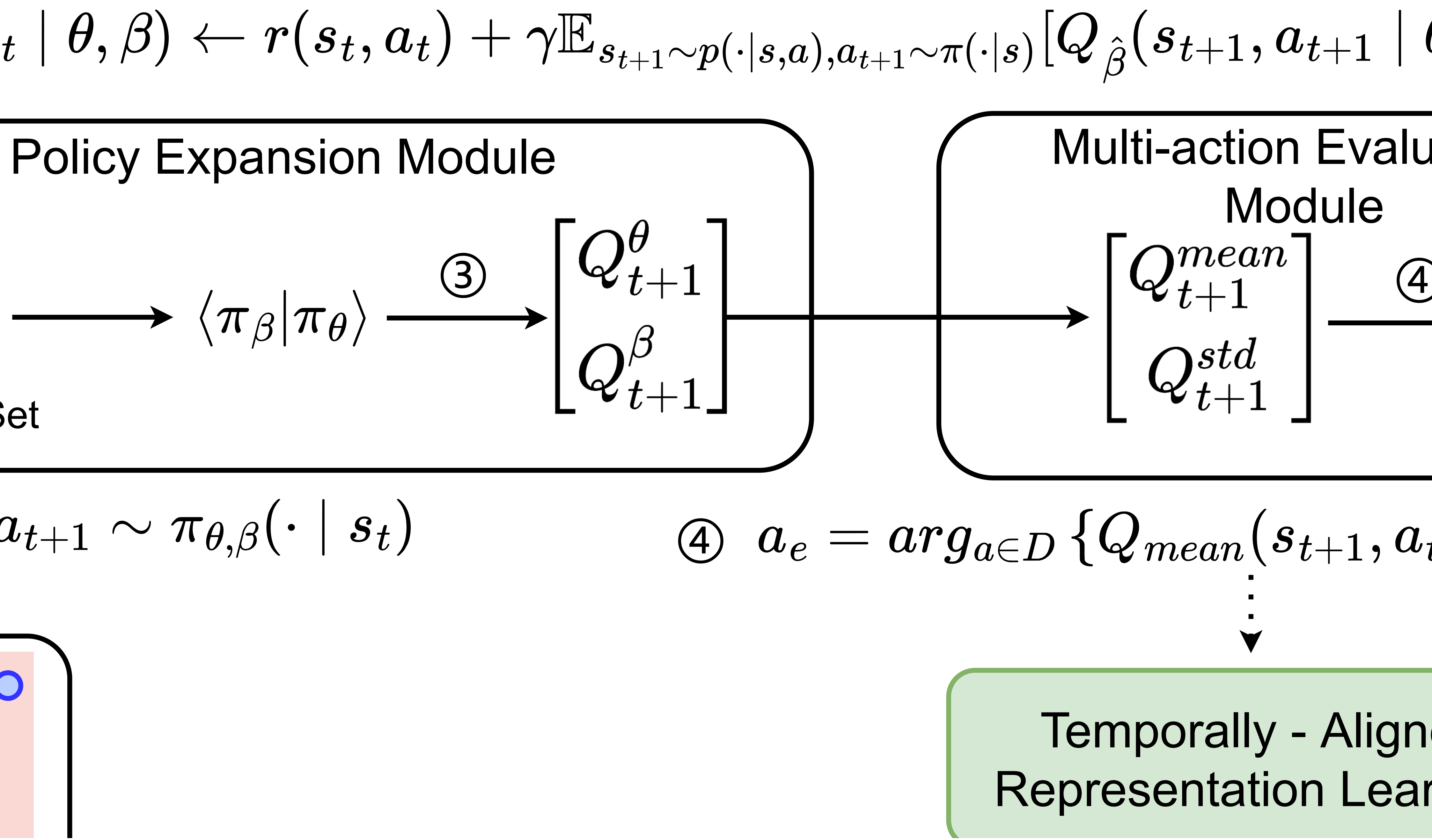
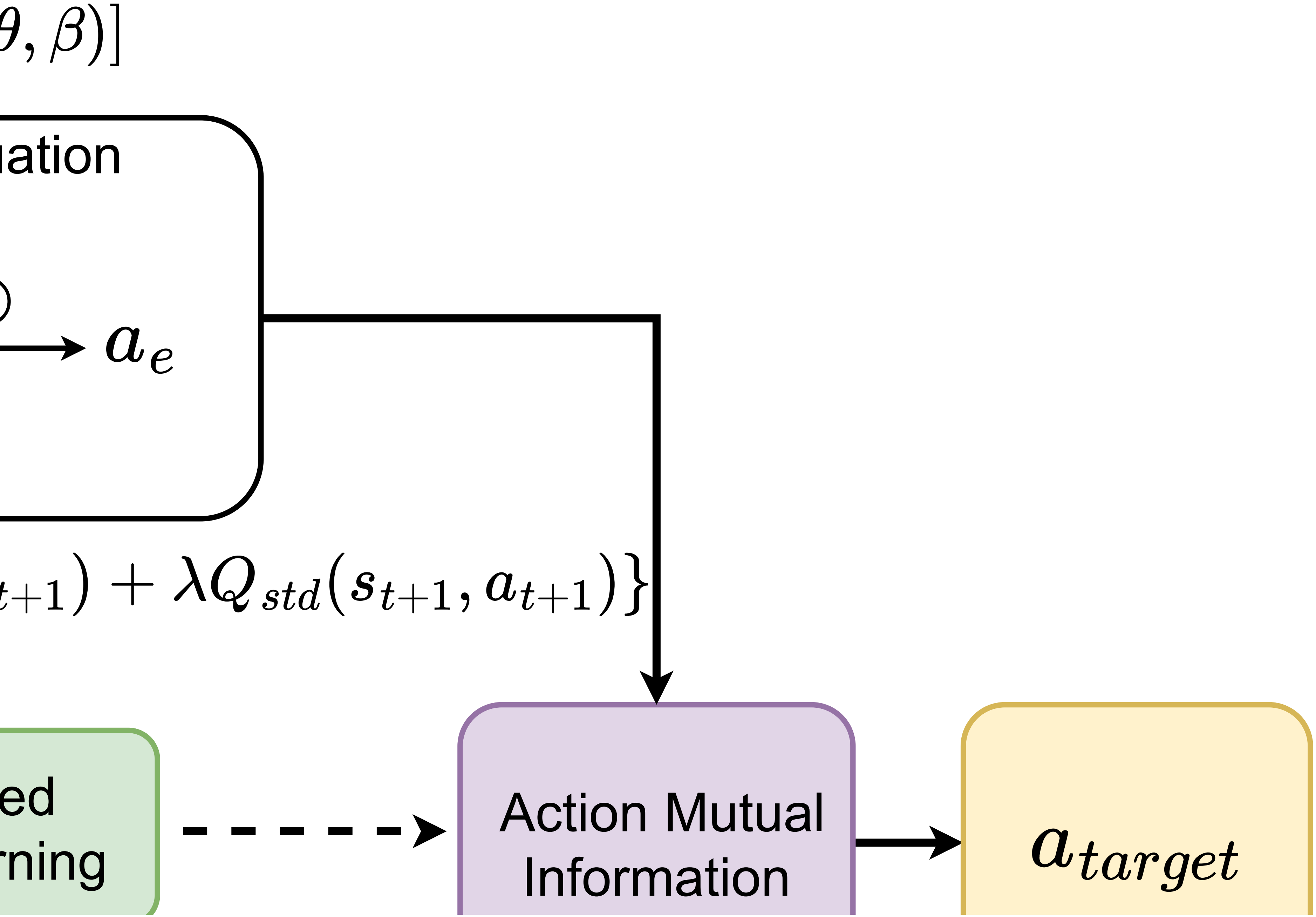
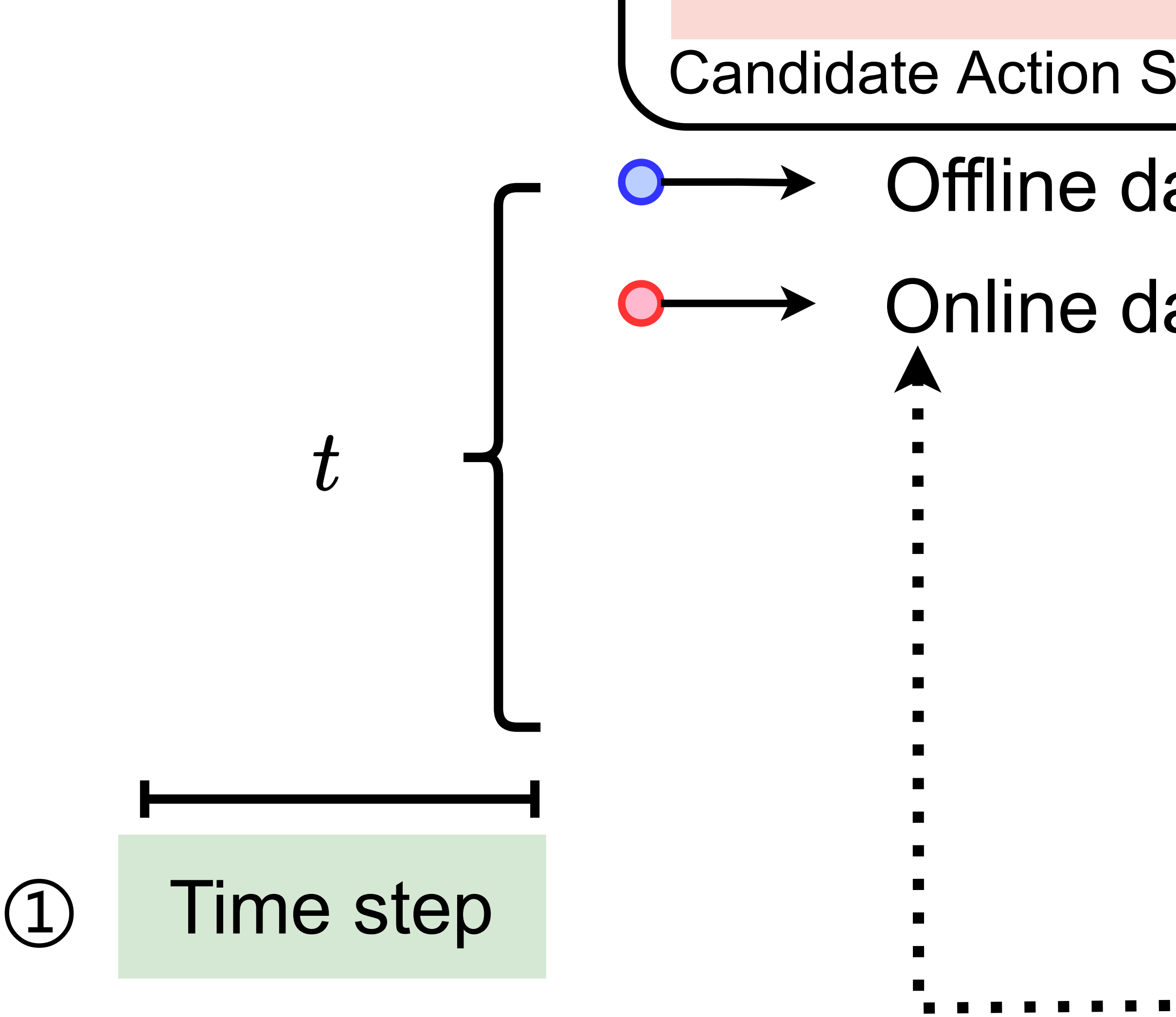
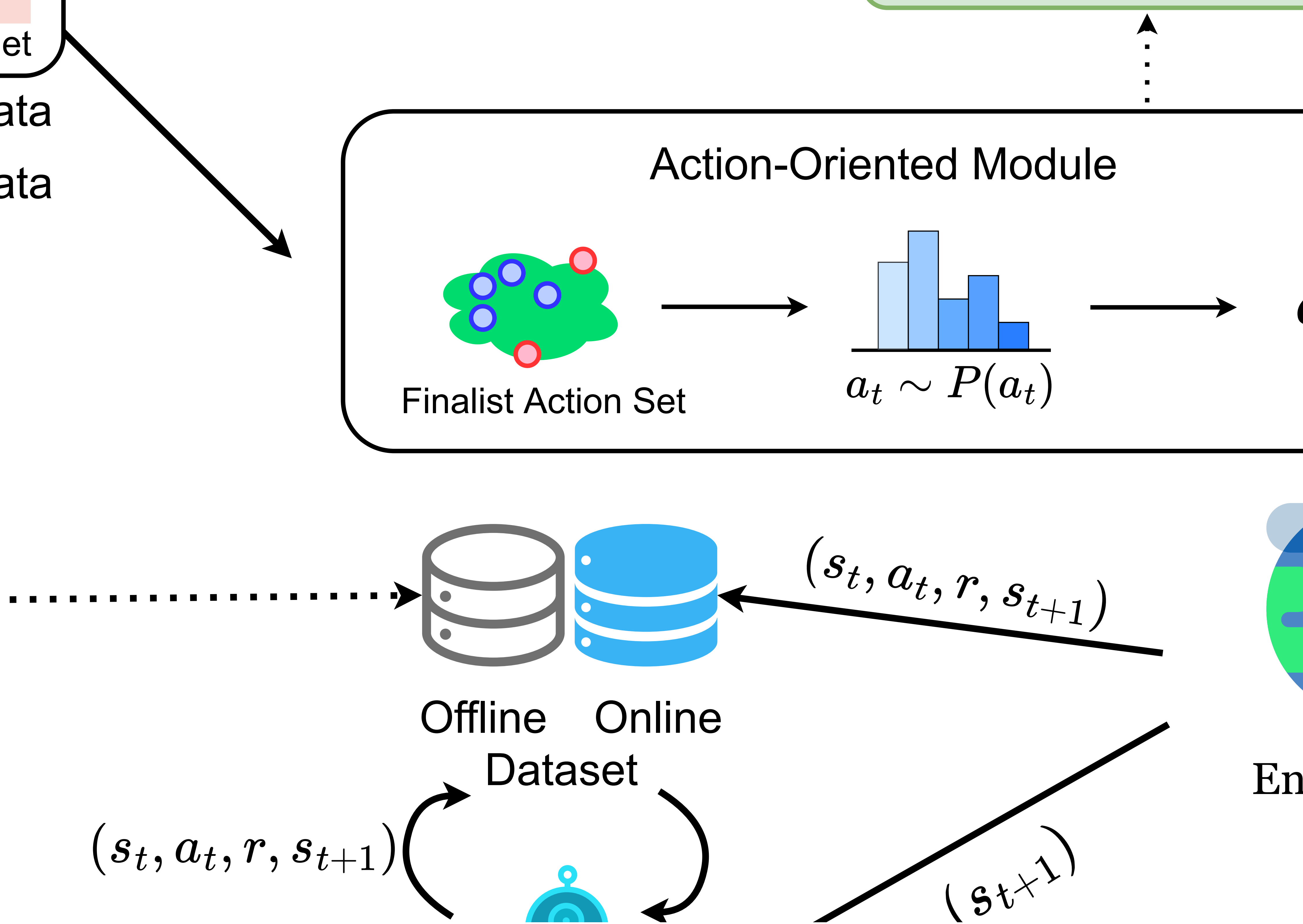
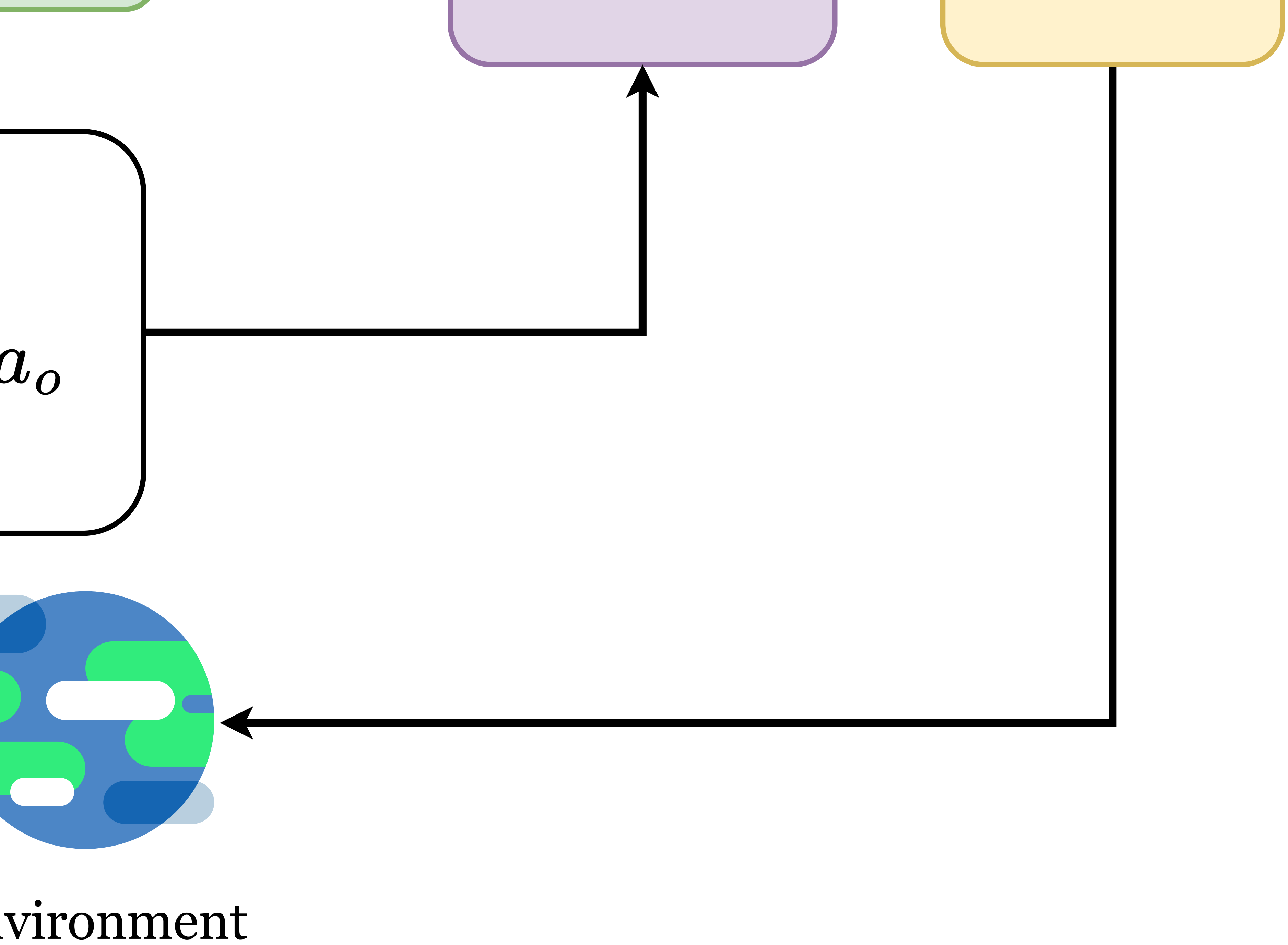

①
Time step
$t+1$
③ $Q_\beta(s_t, a$
Finalist Action S
②

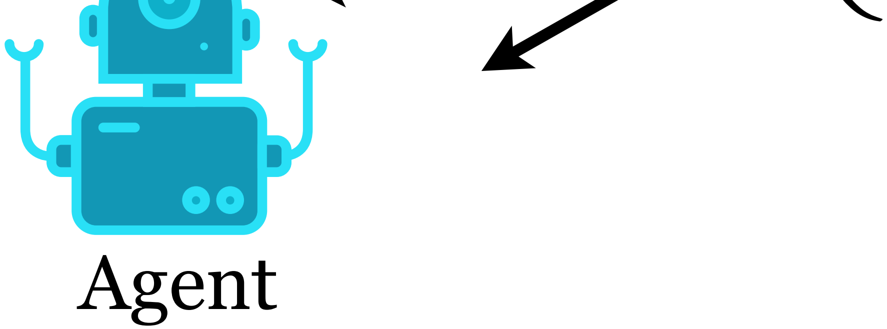

$t \mid \theta, \beta) \leftarrow r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p(\cdot|s,a), a_{t+1} \sim \pi(\cdot|s)} [Q_{\hat{\beta}}(s_{t+1}, a_{t+1} \mid$
Policy Expansion Module
$\langle \pi_\beta | \pi_\theta \rangle$
③
$\begin{bmatrix} Q^{\theta}_{t+1} \\ Q^{\beta}_{t+1} \end{bmatrix}$
Set
Multi-action Evalu
Module
$\begin{bmatrix} Q^{mean}_{t+1} \\ Q^{std}_{t+1} \end{bmatrix}$
④
$a_{t+1} \sim \pi_{\theta,\beta}(\cdot \mid s_t)$
④ $a_e = arg_{a \in D} \{Q_{mean}(s_{t+1}, a$
Temporally - Aligne
Representation Lear


$\theta, \beta)]$
uation
a_e
$_{t+1}) + \lambda Q_{std}(s_{t+1}, a_{t+1})\}$
ed
rning
Action Mutual Information
a_{target}


Candidate Action S
Offline da
Online da
t
① Time step


et
ata
ata
Action-Oriented Module
Finalist Action Set
$a_t \sim P(a_t)$
(s_t, a_t, r, s_{t+1})
Offline
Online
Dataset
En
(s_t, a_t, r, s_{t+1})
(s_{t+1})


a_o
vironment



Agent