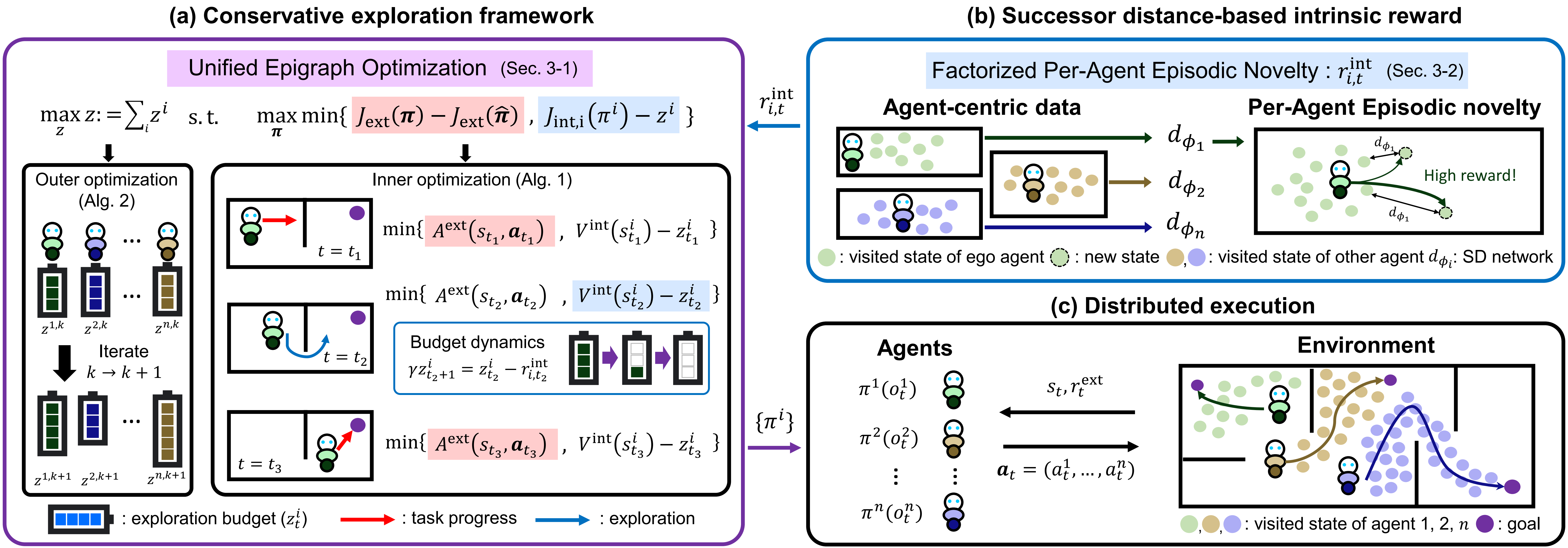
(a) Conservative exploration framework
Unified Epigraph Optimization (Sec. 3-1)
$\max_z z := \sum_i z^i$ s. t. $\max_{\pi} \min\{ J_{\text{ext}}(\pi) - J_{\text{ext}}(\hat{\pi}), J_{\text{int,i}}(\pi^i) - z^i \}$
Outer optimization (Alg. 2)
...
$z^{1,k}$
$z^{2,k}$
$z^{n,k}$
Iterate
$k \rightarrow k+1$
$z^{1,k+1}$
$z^{2,k+1}$
$z^{n,k+1}$
Inner optimization (Alg. 1)
$t = t_1$
$\min\{ A^{\text{ext}}(s_{t_1}, a_{t_1}), V^{\text{int}}(s^i_{t_1}) - z^i_{t_1} \}$
$t = t_2$
$\min\{ A^{\text{ext}}(s_{t_2}, a_{t_2}), V^{\text{int}}(s^i_{t_2}) - z^i_{t_2} \}$
Budget dynamics
$\gamma z^i_{t_2+1} = z^i_{t_2} - r^{\text{int}}_{i,t_2}$
$t = t_3$
$\min\{ A^{\text{ext}}(s_{t_3}, a_{t_3}), V^{\text{int}}(s^i_{t_3}) - z^i_{t_3} \}$
: exploration budget (z^i_t)
: task progress
: exploration
$r^{\text{int}}_{i,t}$
$\{\pi^i\}$
(b) Successor distance-based intrinsic reward
Factorized Per-Agent Episodic Novelty : $r^{\text{int}}_{i,t}$ (Sec. 3-2)
Agent-centric data
Per-Agent Episodic novelty
d_{ϕ_1}
d_{ϕ_2}
d_{ϕ_n}
High reward!
: visited state of ego agent
: new state
: visited state of other agent
d_{ϕ_i}: SD network
(c) Distributed execution
Agents
Environment
$\pi^1(o^1_t)$
$\pi^2(o^2_t)$
$\pi^n(o^n_t)$
s_t, r^{ext}_t
$a_t = (a^1_t, \ldots, a^n_t)$
: visited state of agent 1, 2, n
: goal