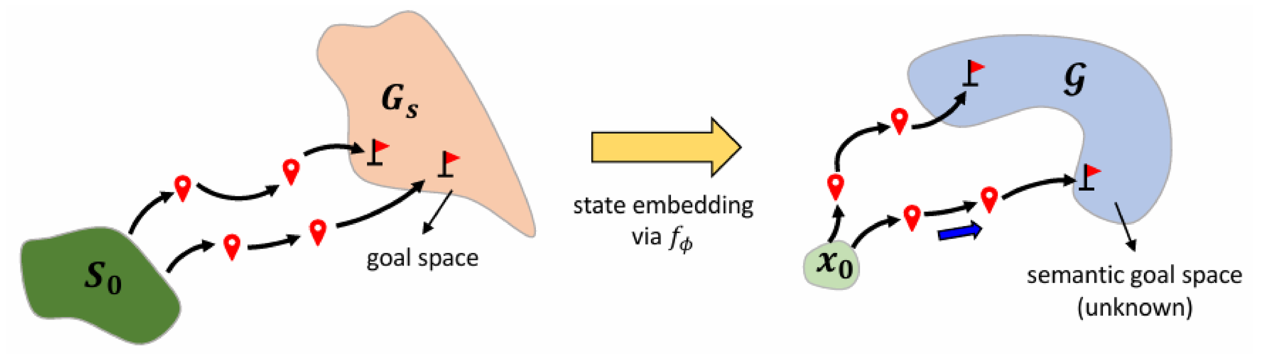
LAGMA: LAtent Goal-guided Multi-Agent Reinforcement Learning

Hyungho Na[1] and Il-Chul Moon[1,2]
[1]KAIST, [2]summary.ai

Paper Code

Motivation: Efficient Training for MARL

- Goal-conditioned RL (GCRL) has shown good performance in a single agent task, such as complex pathfinding with sparse reward
- GCRL concept has been limitedly applied to MARL task
  - ➢ Goal is not explicitly known
  - ➢ Partial observation and decentralized execution
  - ➢ complex coordination rather than the shortest path finding



- Cooperative MARL problem
  → finding trajectories toward semantic goals in latent space

Key Concept: Goal-reaching Trajectory

- Achieving a common goal in cooperative tasks → **Goal-reaching**
  - $\mathcal{T}$ satisfying the following is considered **Goal-reaching denoted as** $\mathcal{T}^*$

$$\mathcal{T} := \{s_0, \boldsymbol{a_0}, r_0, s_1, \boldsymbol{a_1}, r_1, \ldots, s_T\} \quad \text{such that} \quad \Sigma_{t=0}^{T-1} r_t = R_{\max}$$

  - For $\forall s \in \mathcal{T}^*$, $\tau^*_{s_t} := \{s_t, s_{t+1}, \ldots s_T\}$ is a goal-reaching trajectory

Embedding via modified VQ-VAE

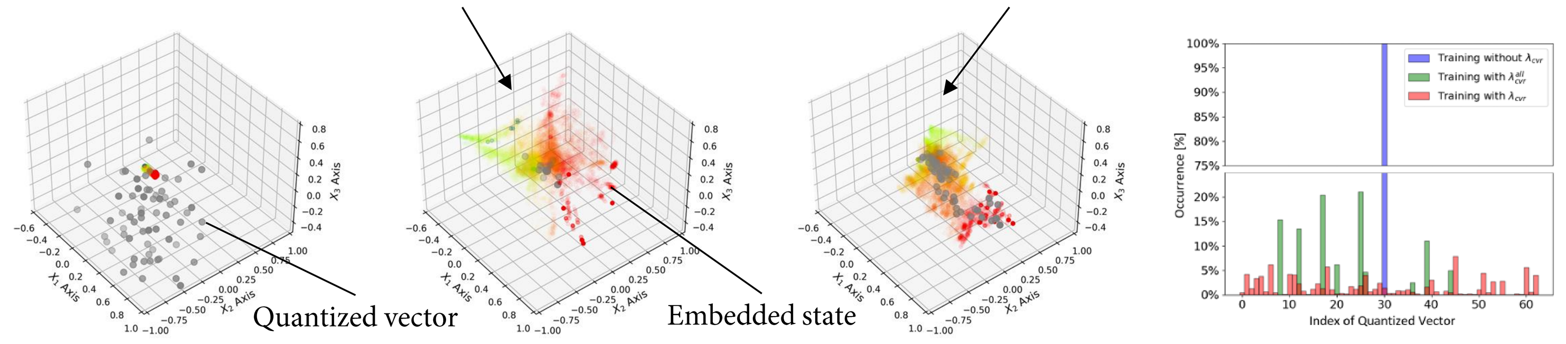
- For quantized embedding space construction
- ➢ VQ-VAE: $\mathcal{L}_{VQ}(\phi, \psi, \boldsymbol{e}) =$

$$\underbrace{||f_\psi([x = f_\phi(s)]_q) - s||_2^2}_{\text{reconstruction loss}} + \lambda_{\text{vq}}\underbrace{||\text{sg}[f_\phi(s)] - x_q||_2^2}_{\text{VQ loss}} + \lambda_{\text{commit}}\underbrace{||f_\phi(s) - \text{sg}[x_q]||_2^2}_{\text{commitment loss}}$$

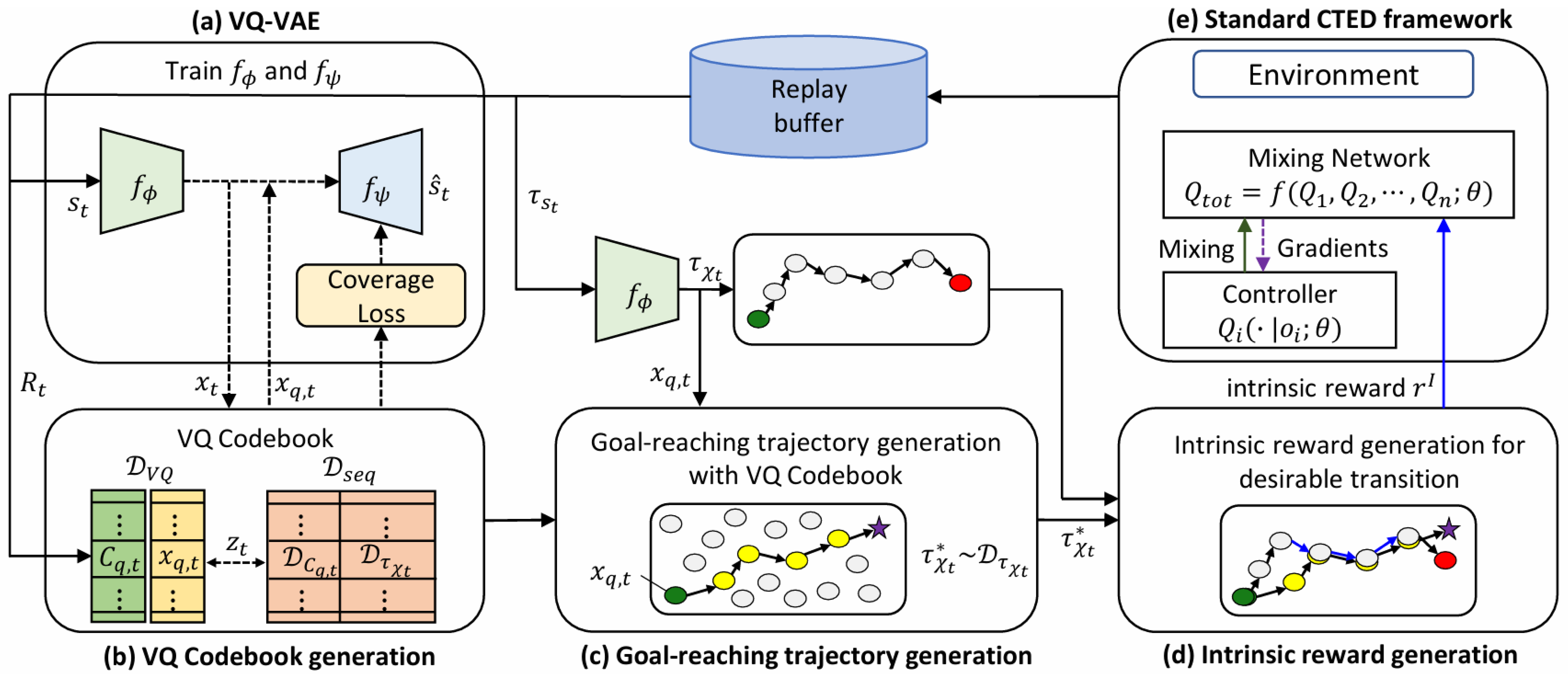
- ➢ Modified VQ-VAE: $\mathcal{L}^{tot}_{VQ}(\phi, \psi, \boldsymbol{e})f = \mathcal{L}_{VQ}(\phi, \psi, \boldsymbol{e}) + \lambda_{\text{cvr}}\boxed{\mathcal{L}_{\text{cvr}}(\boldsymbol{e})}$

coverage loss

$$\mathcal{L}^{\text{all}}_{\text{cvr}}(\boldsymbol{e}) = \frac{1}{n_c}\sum_{j=1}^{n_c}||\text{sg}[f_\phi(s)] - e_j||_2^2 \qquad \mathcal{L}_{\text{cvr}}(\boldsymbol{e}) = \frac{1}{|\mathcal{J}(t)|}\sum_{j\in\mathcal{J}(t)}||\text{sg}[f_\phi(s)] - e_j||_2^2$$



LAGMA generates a goal-reaching trajectory in latent space and incentivizes transitions towards this reference trajectory



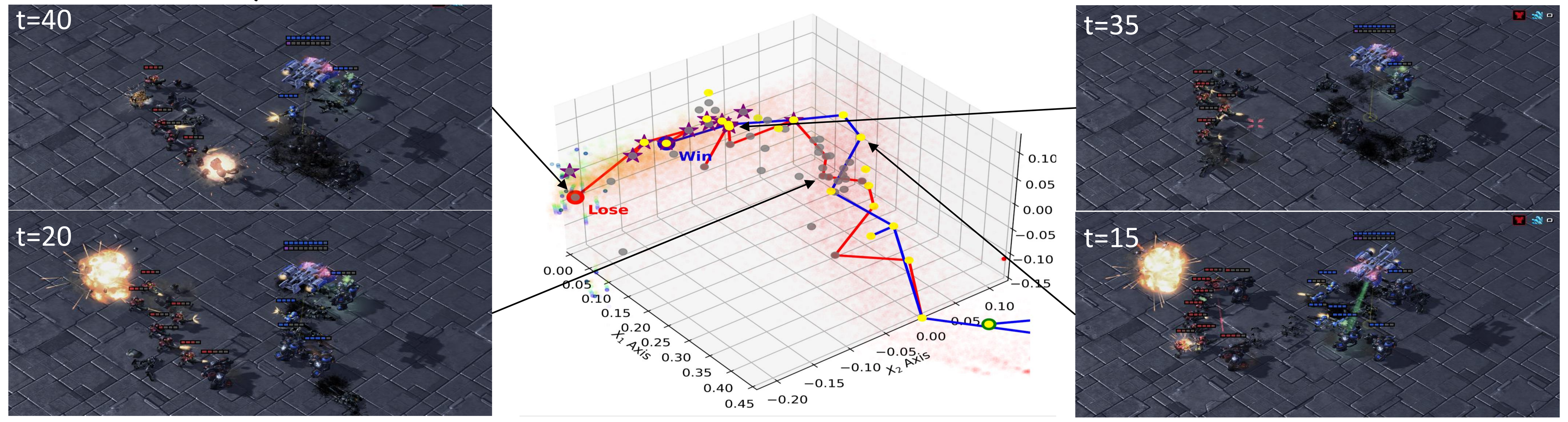
- ➢ **(1) Modified VQ-VAE** is developed for quantized embedding space construction
- ➢ **(2) Goal-reaching trajectory** is generated via **extended codebook**
- ➢ **(3) Latent Goal-guided intrinsic reward** guarantees a better convergence on optimal policy

Visualization in Latent Space

- Overall Objective:

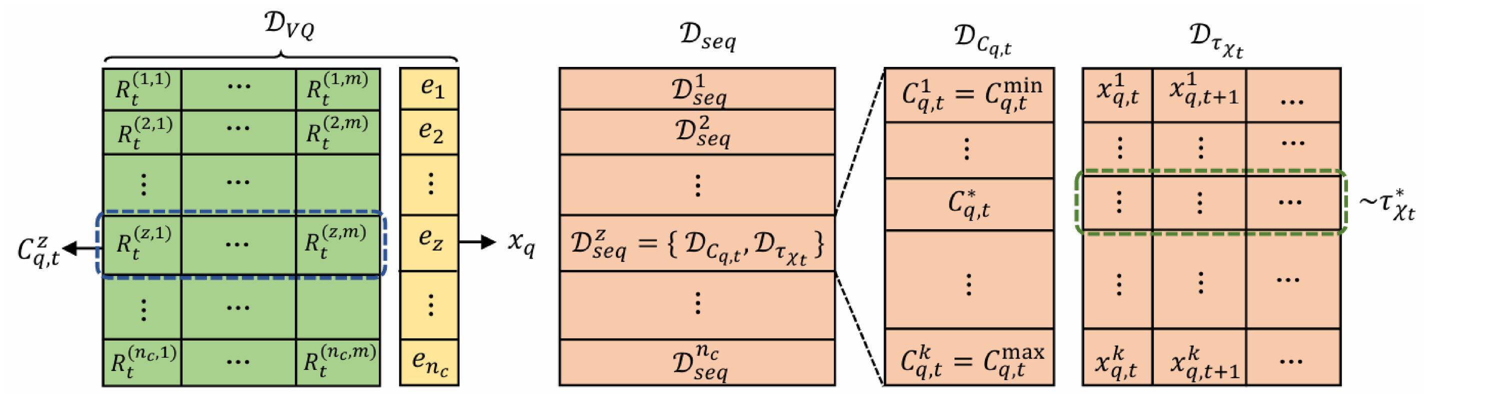
$$\mathcal{L}(\theta) = \left(r^{\text{ext}} + r^I + \gamma\max_{\boldsymbol{a}'} Q^{tot}_{\theta^-}(s', \boldsymbol{a}') - Q^{tot}_\theta(s, \boldsymbol{a})\right)^2$$

- Qualitative Analysis:



Goal-reaching Trajectory Generation

- The value of quantized vector is computed as $C_{q,t}(x_{q,t}) = \frac{1}{N_{x_{q,t}}}\sum_{j=1}^{N_{x_{q,t}}} R^j_t(x_{q,t})$
- Note that: $R_t = \Sigma_{i=t}^{T-1}\gamma^{i-t}r_i$ and $\tau_{\chi_t} = [f_\phi(\tau_{x_t})]_q = \{x_{q,t}, x_{q,t+1}, x_{q,t+2}, \ldots, x_{q,T}\}$

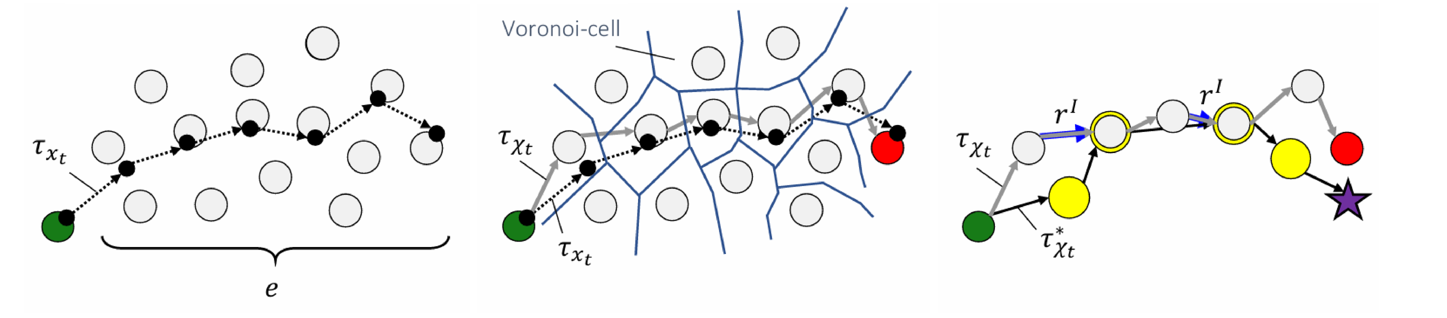


Training with Latent Goal-guided Incentive

For $s' \in \tau^*_{\chi_t}$, $\boxed{r^I(s') := \gamma(C_{q,t}(s') - \max_{\boldsymbol{a}'} Q_{\theta^-}(s', \boldsymbol{a}'))}$ guarantees

$$y = r(s, \boldsymbol{a}) + \boxed{r^I(s')} + \gamma V_{\theta^-}(s') \to y^*$$

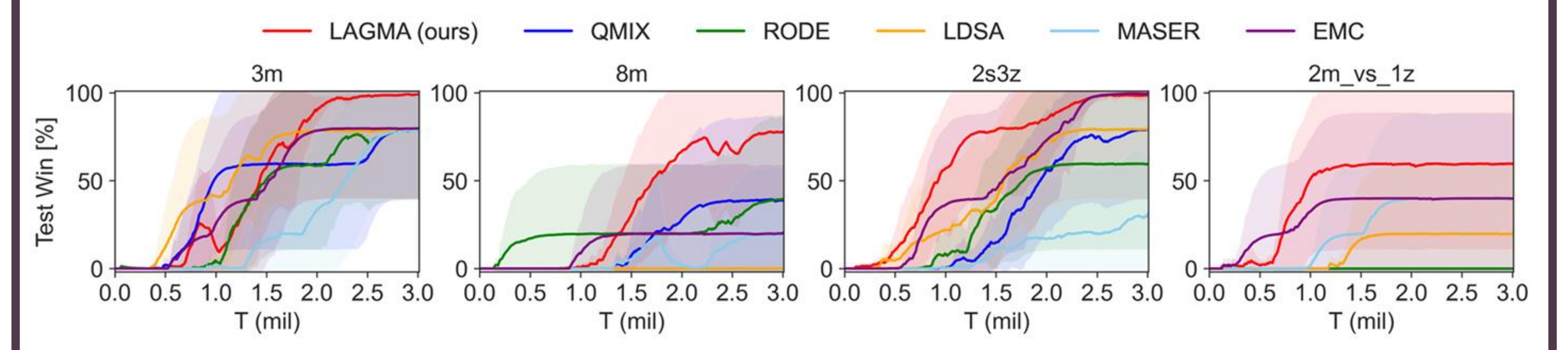
where $y^* = r(s, \boldsymbol{a}) + \gamma V^*(s')$ if $x_{q,t+1} \in \tau^*_{\chi_t}$ and $x_{q,t+1} \neq x_{q,t}$



(a) Trajectory embedding. (b) Trajectory in quantized latent space. (c) Intrinsic reward generation.

Experiments

- Performance comparison in SMAC (sparse reward)



- Ablation on key components
- ➢ Evaluated with controlled parameters