Mingling Foresight with Imagination: Model-Based Cooperative Multi-Agent Reinforcement Learning

Zhiwei Xu, Dapeng Li, Bin Zhang, Yuan Zhan, Yunpeng Bai, Guoliang Fan

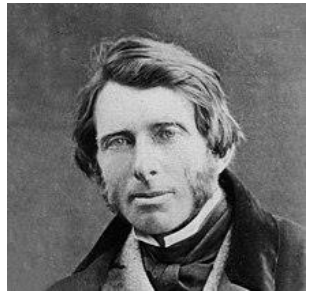
Institute of Automation, Chinese Academy of Sciences

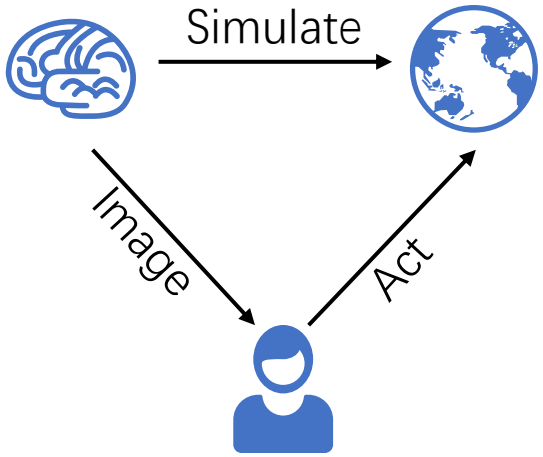
School of Artificial Intelligence, University of Chinese Academy of Sciences

Why to have the foresight?

To mingle prudence and foresight with imagination and admiration, and you have the perfect human soul.

——John Ruskin

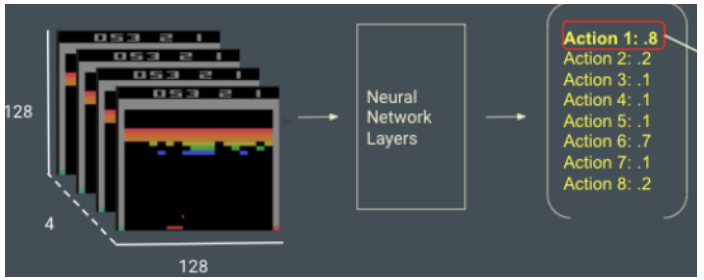




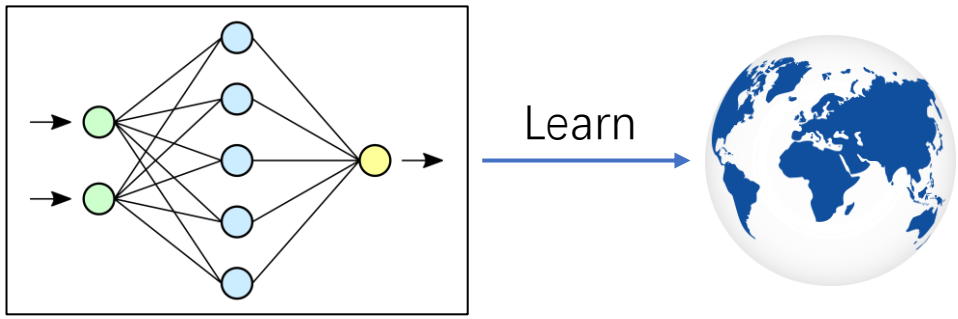
When humans make decisions, they not only rely on the current state but also consider the future state obtained after several interactions with the environment following their current strategies. The phenomenon is called "long-term vision" or foresight.

How to have the foresight?

- The state-action value / state value



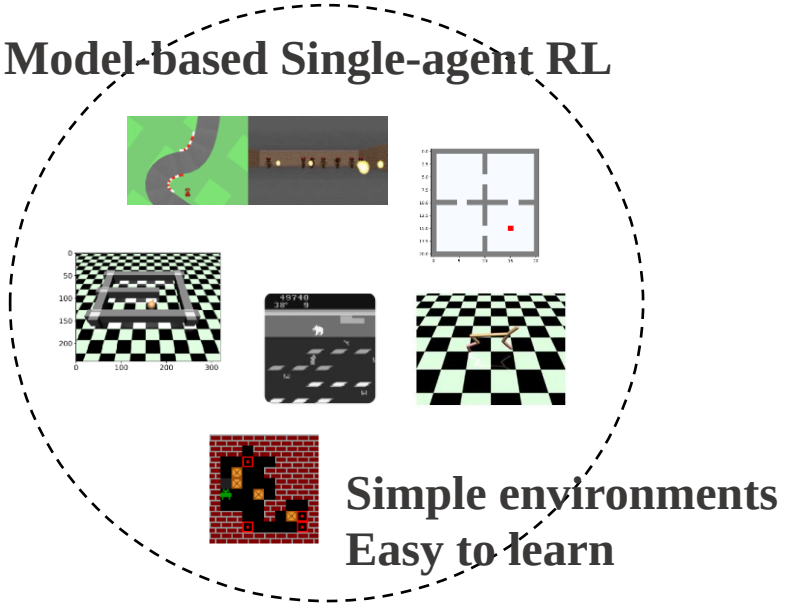
- Auxiliary tasks for predicting the next signal / World Model



Apply Model-based Single-agent RL to Multi-agent RL?

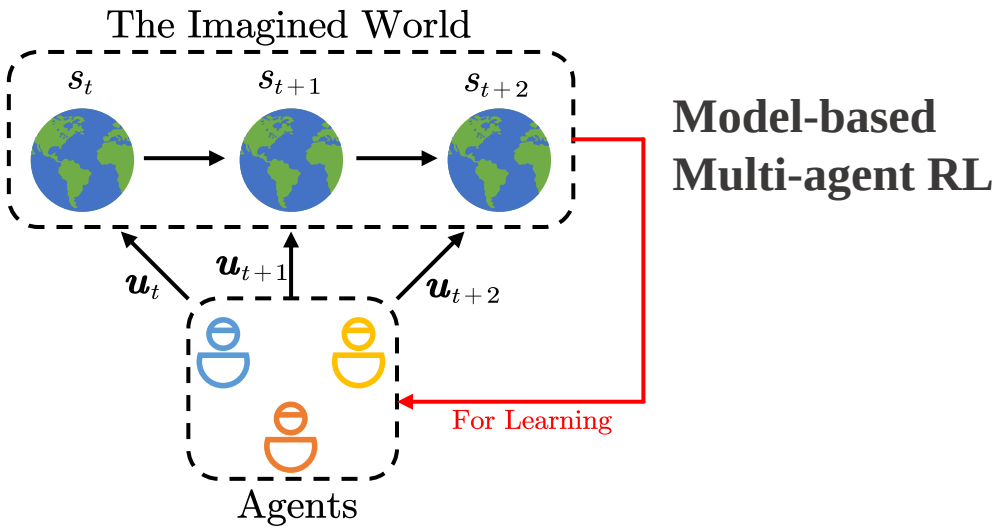
Model-based Single-agent RL

- MuZero
- Recurrent World Models
- Model-based Value Expansion
- Value Prediction Network
- Dreamer DreamerV2
- …



Model-based Multi-agent RL?

- Environmental models are not available
- Environmental models are difficult to learn
- Complexity of multi-agent problem



Environmental Models

- Auto-regressive models

1. Intuitive and simple.
2. Calculated items of the generative process cannot be reused.
3. Explicitly render high-dimensional observations.

- State-space models

1. Abstract the environment.
2. Find a compact latent state space $\mathcal{S}$ that can contain all important information.
3. In the multi-agent system, each $\hat{s} \in \mathcal{S}$ is an abstract representation of the local observation $\boldsymbol{z}_t$ of all agents

$$p(\boldsymbol{z}_t \mid \hat{s}_{0:t}, \boldsymbol{u}_{0:t-1}) = p(\boldsymbol{z}_t \mid \hat{s}_t)$$

$$p(\boldsymbol{z}_{0:T}, \boldsymbol{u}_{0:T}) = p_{\text{init}}(\hat{s}_0) \cdot \int \prod_{t=1}^{T} \left(p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right) p(\boldsymbol{u}_t \mid \boldsymbol{z}_t) p(\boldsymbol{z}_t \mid \hat{s}_t)\right) d\hat{s}_{1:T}$$

We replace $p(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t)$ with $q_\theta(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t)$, then we get the evidence lower bound (ELBO):

$$\begin{aligned}
&\log p\left(\boldsymbol{z}_{0:T}, \boldsymbol{u}_{0:T}\right) \\
&- \log \mathbb{E}_{q_\theta(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T})}\left[\frac{p\left(\hat{s}_{1:T}, \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}{q_\theta\left(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}\right] \\
&\geq \mathbb{E}_{q_\theta(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T})} \log \left[\frac{p\left(\hat{s}_{1:T}, \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}{q_\theta\left(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}\right] \\
&- \int q_\theta\left(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right) \log \left[\frac{p\left(\hat{s}_{1:T}, \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}{q_\theta\left(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right)}\right] d\hat{s}_{1:T} \\
&- \int \sum_{t=1}^{T} q_\theta\left(\hat{s}_{1:T} \mid \boldsymbol{u}_{0:T}, \boldsymbol{z}_{1:T}\right) \log \left[\frac{p(\boldsymbol{u}_t \mid \boldsymbol{z}_t)\, p(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1})\, p(\boldsymbol{z}_t \mid \hat{s}_t)}{q_\theta(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t)}\right] d\hat{s}_{1:T} \\
&- \sum_{t=1}^{T} \left\{ \int q_\theta\left(\hat{s}_{1:t} \mid \boldsymbol{u}_{0:t}, \boldsymbol{z}_{1:t}\right) \log \left[p(\boldsymbol{u}_t \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_t \mid \hat{s}_t)\right] d\hat{s}_{1:t} \right. \\
&\quad \left. + \int q_\theta\left(\hat{s}_{1:t} \mid \boldsymbol{u}_{0:t}, \boldsymbol{z}_{1:t}\right) \log \left[\frac{p(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1})}{q_\theta(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t)}\right] d\hat{s}_{1:t} \right\} \\
&- \sum_{t=1}^{T} \left\{ \int q_\theta\left(\hat{s}_{1:t} \mid \boldsymbol{u}_{0:t}, \boldsymbol{z}_{1:t}\right) \log \left[p(\boldsymbol{u}_t \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_t \mid \hat{s}_t)\right] d\hat{s}_{1:t} \right. \\
&\quad \left. - \int q_\theta\left(\hat{s}_{1:t-1} \mid \boldsymbol{u}_{0:t-1}, \boldsymbol{z}_{1:t-1}\right) \mathcal{D}_{\mathrm{KL}}\left[q_\theta\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right) \| p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)\right] d\hat{s}_{1:t} \right\} \\
&- \sum_{t=1}^{T} \left\{ \log \left[p(\boldsymbol{u}_t \mid \boldsymbol{z}_t)\right] + \log \left[p(\boldsymbol{z}_t \mid \hat{s}_t)\right] \right. \\
&\quad \left. - \mathcal{D}_{\mathrm{KL}}\left[q_\theta\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right) \| p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)\right] \right\},
\end{aligned}$$

Model-Based Value Decomposition

The ELBO is:

$$\mathcal{L}\left(\boldsymbol{z}^{0:T}, \boldsymbol{u}^{0:T}\right)$$
$$= \sum_{t=1}^{T} \{\log\left[p(\boldsymbol{u}_t \mid \boldsymbol{z}_t)\right] + \log\left[p(\boldsymbol{z}_t \mid \hat{s}_t)\right]$$
$$-\mathcal{D}_{\mathrm{KL}}\left[q_\theta\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right) \| p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)\right]\}$$

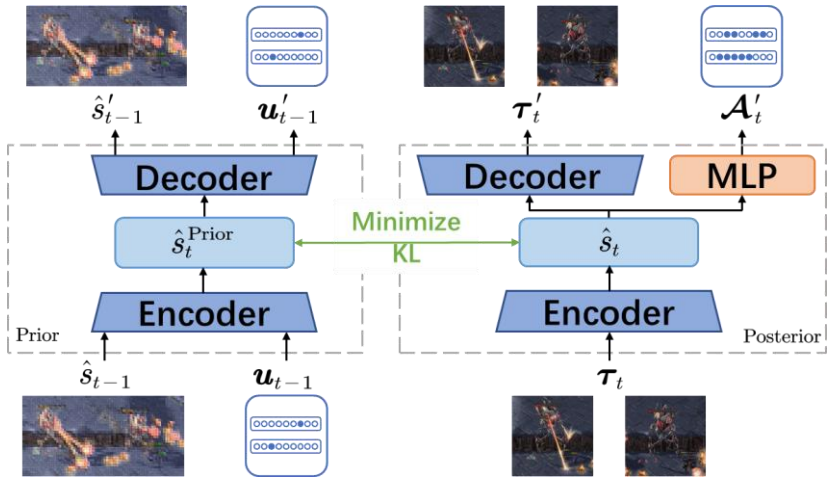
$p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)$ is not available. So we replace $p\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)$ with $p_\phi^{\mathrm{Prior}}\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)$.

As mentioned above, we need to approximate two models:

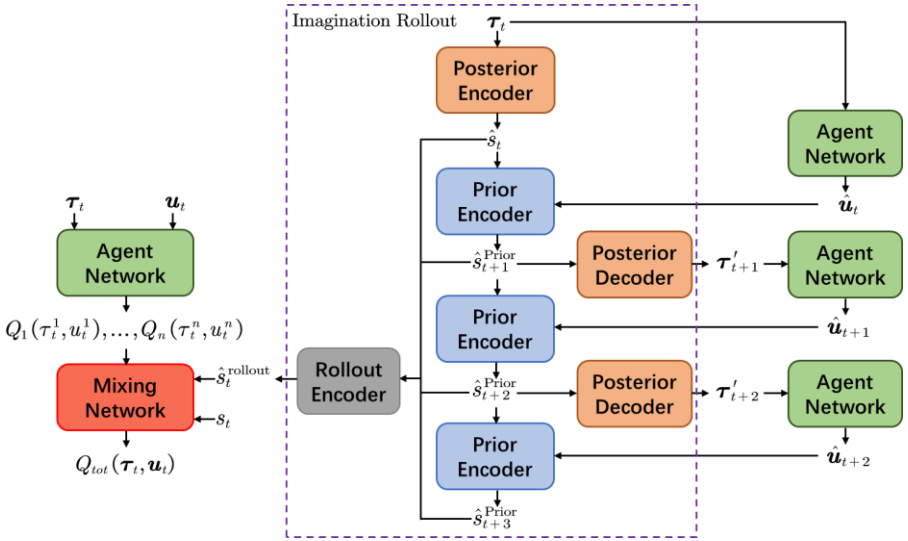
$$\hat{s}_t^{\mathrm{Prior}} \sim p_\phi^{\mathrm{Prior}}\left(\cdot \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right),$$
$$\hat{s}_t \sim q_\theta\left(\cdot \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right).$$

Our proposed **M**odel-**B**ased **V**alue **D**ecompostion (**MBVD**) follows the idea of planning in latent spaces, so we use the Variational Autoencoder (VAE) to maximize the ELBO. For the last term in the ELBO, we regard the two items in Kullback-Leibler (KL) divergence as the posterior and the prior.

The framework of the imagination module in MBVD:



The workflow of MBVD:



Loss Function

The loss function for reinforcement learning can be obtained:

$$\mathcal{L}_{\mathrm{RL}} = \left(y^{tot} - Q_{tot}(\boldsymbol{\tau}_t, \boldsymbol{u}_t, s_t, \hat{s}_t^{\mathrm{Rollout}}; \psi)\right)^2,$$

where $y^{tot} = r_t + \gamma \max_{\boldsymbol{u}_{t+1}} Q_{tot}\left(\boldsymbol{\tau}_{t+1}, \boldsymbol{u}_{t+1}, s_{t+1}, \hat{s}_{t+1}^{\mathrm{Rollout}}; \ \psi^{-}\right)$.

The loss function of the posterior can be divided into reconstruction loss and KL divergence loss. The posterior model infers the current latent state after the observations of all agents are given, which requires that the posterior model extract helpful information from the original input. It can be achieved by narrowing the difference between the model output τ' and the model input τ. The reconstruction loss function and the KL loss can be written as:

$$\mathcal{L}_{\mathrm{RC}} = \mathrm{MSE}\left(\boldsymbol{\tau}_t, \boldsymbol{\tau}_t'; \theta\right), \qquad \mathcal{L}_{\mathrm{RC}}^{\mathrm{Prior}} = \mathrm{MSE}\left(\left(\hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right), \left(\hat{s}_{t-1}', \boldsymbol{u}_{t-1}'\right); \phi\right),$$

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}} = {} & \mathcal{D}_{\mathrm{KL}}\left[p_\phi^{\mathrm{Prior}}\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right) \| \mathcal{N}(0,1)\right] \\ & + \mathcal{D}_{\mathrm{KL}}\left[q_\theta\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{z}_t\right) \| p_\phi^{\mathrm{Prior}}\left(\hat{s}_t \mid \hat{s}_{t-1}, \boldsymbol{u}_{t-1}\right)\right].\end{aligned}$$

In addition, we have other optional training objectives for auxiliary tasks in complex scenarios. In this paper, we make predictions of the feasible action set only in StarCraft II:

$$\mathcal{L}_{\mathrm{FA}} = \mathrm{BCE}\left(\mathcal{A}_t, \mathcal{A}_t'; \phi\right)$$

where BCE denotes the binary cross-entropy error. Thus, the total loss function can be written as:

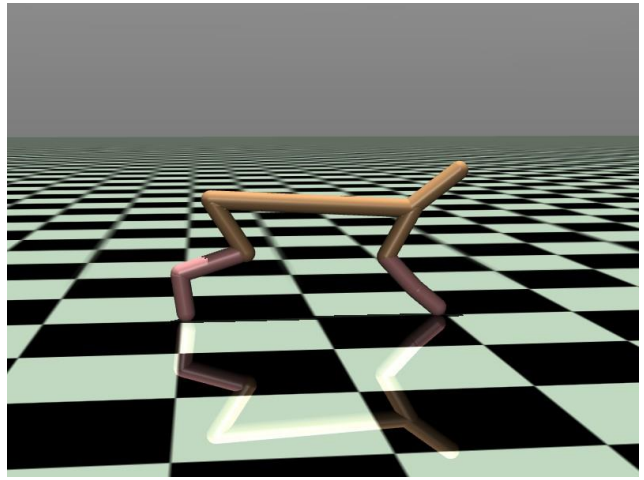
$$\mathcal{L} = \mathcal{L}_{\mathrm{RL}} + \mathcal{L}_{\mathrm{RC}} + \mathcal{L}_{\mathrm{RC}}^{\mathrm{Prior}} + \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{FA}}$$

By minimizing the total loss function $\mathcal{L}$, we can guide MBVD to accelerate reinforcement learning with the help of the imagination module.

Experiment

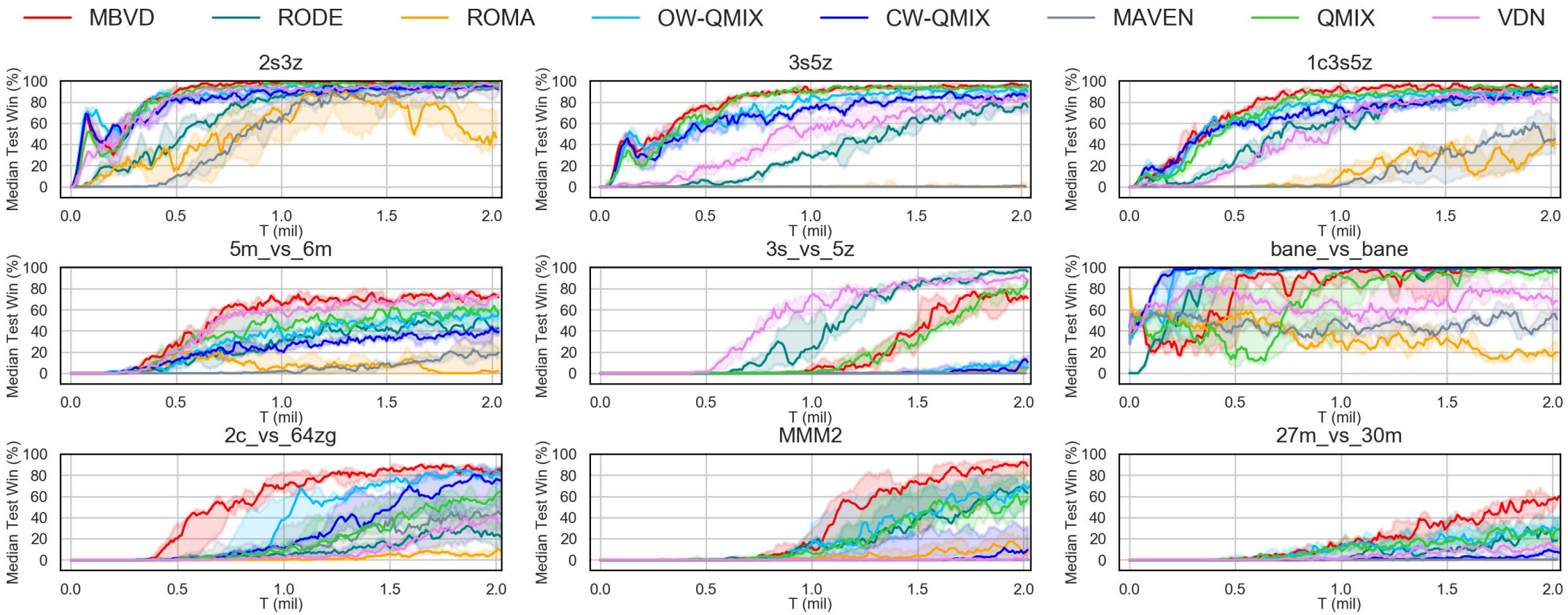
- StarCraft II

- Google Research Football



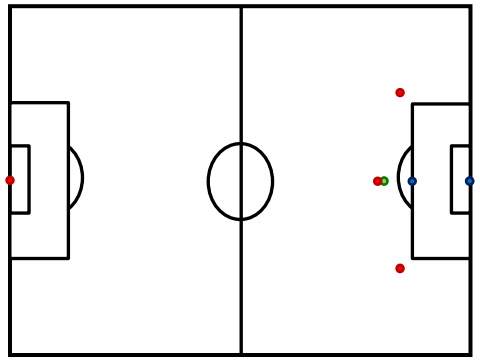
- Multi-agent MuJoCo

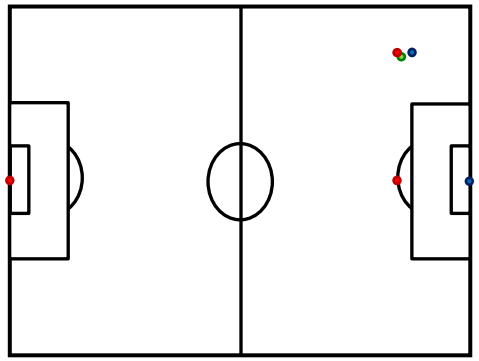
StarCraft II (SMAC)

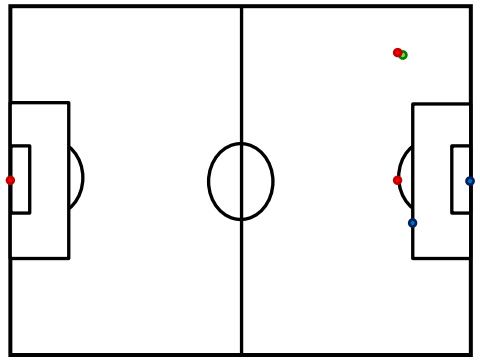


Google Research Football

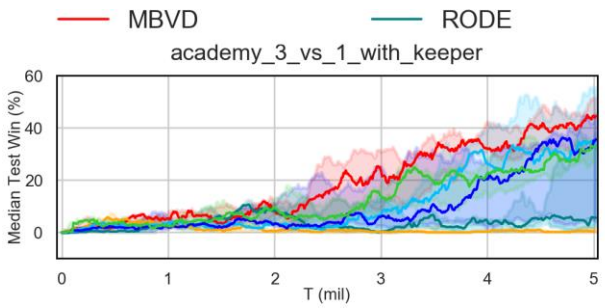
The red dots represent our players, and the blue dots denote the opposing players. The football is represented by green dots.

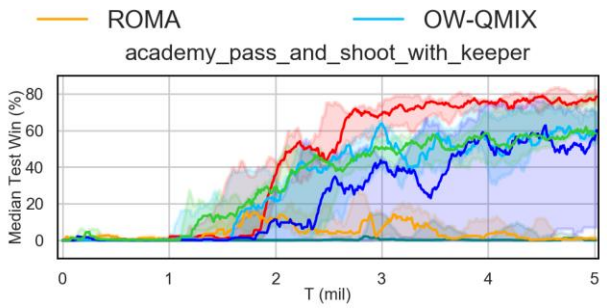


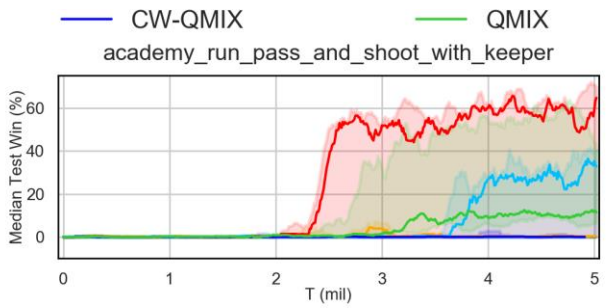




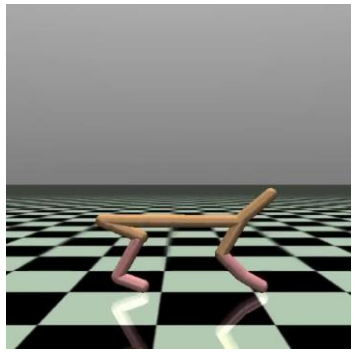
- Academy_3_vs_1_with_keeper
- Academy_pass_and_shoot_with_keeper
- Academy_run_pass_and_shoot_with_keeper



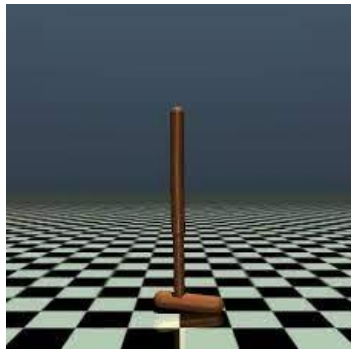




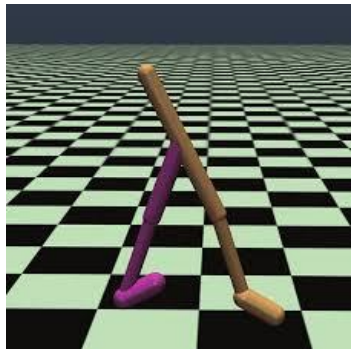
Multi-agent Mujoco



- Half_cheetah (6)

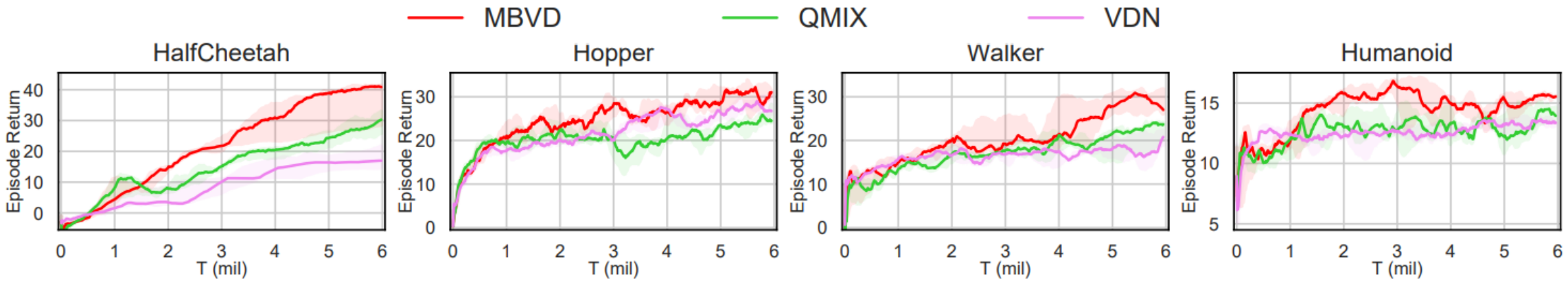


- Hopper (3)

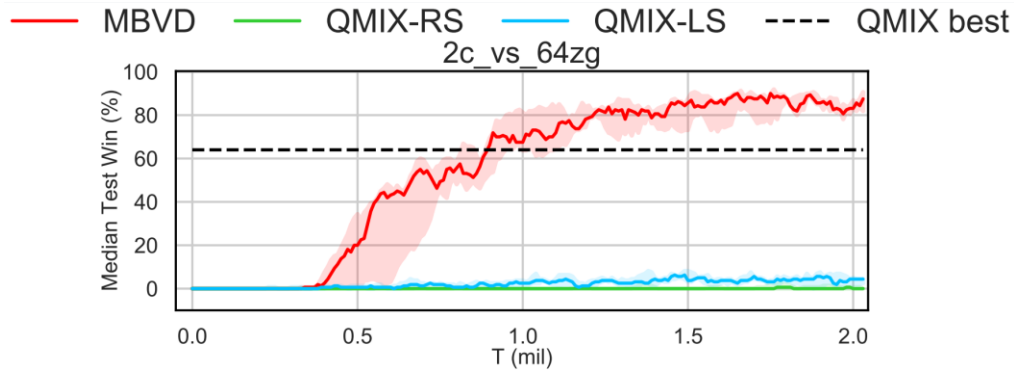


- Walker (6)

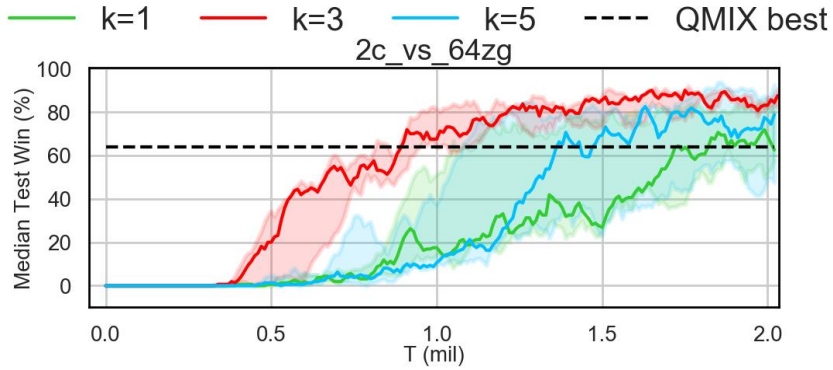
- Humanoid (17)



Ablation Study

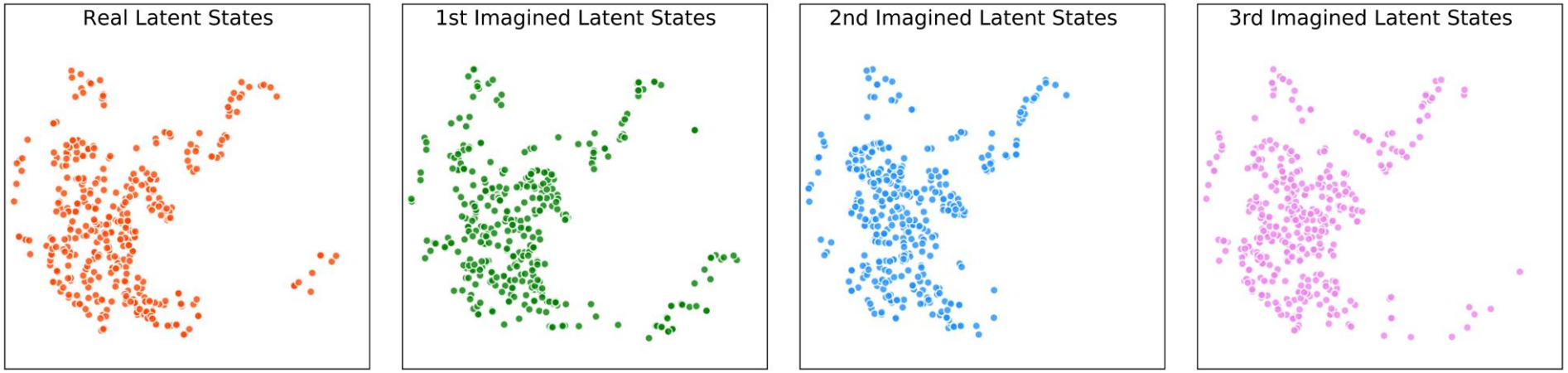


We proposed two variants of QMIX that input the additional information to the mixing network, QMIX-RS and QMIX-LS. Both of them aggregate the information of the next k steps from the actual trajectories rather than imagined rollouts. However, the difference is that QMIX-RS uses the real states and QMIX-LS the latent states. Both QMIX-RS and QMIX-LS failed to solve the task, which means that the transition function that can generate the imagined rollouts is critical



When using short-horizon rollouts, the agents cannot fully interact with the imagined model, resulting in the same performance as vanilla QMIX. However, as k increases significantly, MBVD loses the monotonic improvements because of the compounding error. The effect of k holds the same for the other scenarios, but the optimal choice of k in each task is different.

Visualization



To intuitively explain agents‘ foresight ability in MBVD, we visualized the latent state sequence generated by the interaction between agents and the imagined model. For the trajectories of an episode in the *2c_vs_64zg* scenario, we visualized the t-sne embeddings of the latent states predicted at each step in the rollout and compared them with the real latent state embedding. we find that the difference between the embedding of the imagined latent states and real ones gradually increases as the horizon length grows, but there are similarities between them.

Thanks for listening!