NEURAL INFORMATION PROCESSING SYSTEMS

上海科技大学
ShanghaiTech University

Diffusion-based Reinforcement Learning via Q-weighted Variational Policy Optimization

Shutong Ding[1,3] Ke Hu[1] Zhenhao Zhang[1] Kan Ren[1,3] Weinan Zhang[2]
Jingyi Yu[1,3] Jingya Wang[1,3] Ye Shi[1,3]

[1]ShanghaiTech University [2]Shanghai Jiao Tong University
[3]MoE Key Laboratory of Intelligent Perception and Human Machine Collaboration

NeurIPS 2024

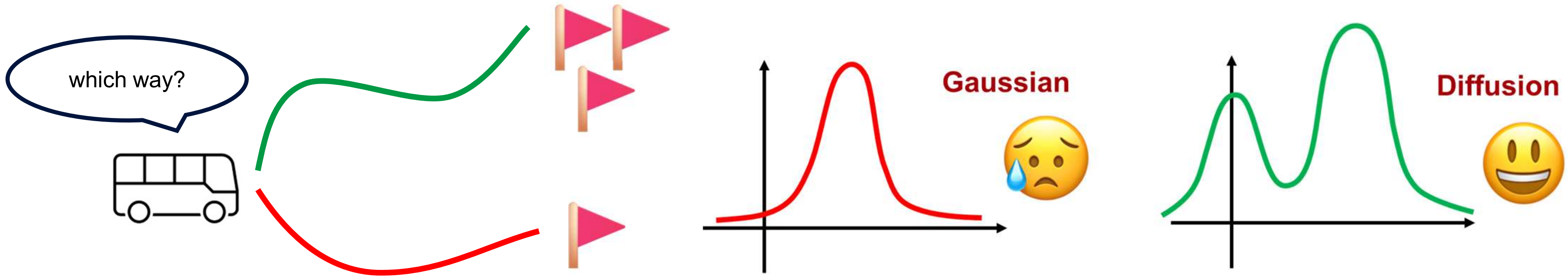
October 7, 2024

立志成才报国裕民

Background: Diffusion in Online RL

1. **Exploration capability** of Gaussian policy or deterministic policy is limited

2. **Expressiveness and multimodality** of diffusion avoid policy falling into the local optimality



Background: Diffusion in Online RL

Directions of applying diffusion in Online RL:

1. **Use the variational loss of diffusion to do policy improvement**

$$\mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[||\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t,t)||^2\right].$$

2. **Apply deterministic policy loss to train the diffusion model (like Diffusion-QL [1])**

$$\pi=\underset{\pi_\theta}{\arg\min}\mathcal{L}(\theta)=\mathcal{L}_d(\theta)+\mathcal{L}_q(\theta)=\mathcal{L}_d(\theta)-\boxed{\alpha\cdot\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D},\boldsymbol{a}^0\sim\pi_\theta}\left[Q_\phi\left(\boldsymbol{s},\boldsymbol{a}^0\right)\right]}$$

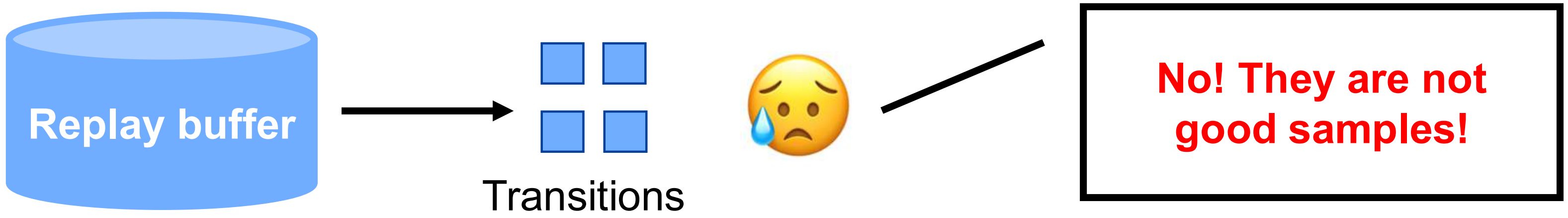
[1] Wang Z, Hunt J J, Zhou M. Diffusion policies as an expressive policy class for offline reinforcement learning[J]. arXiv preprint arXiv:2208.06193, 2022..

Background: Diffusion in Online RL

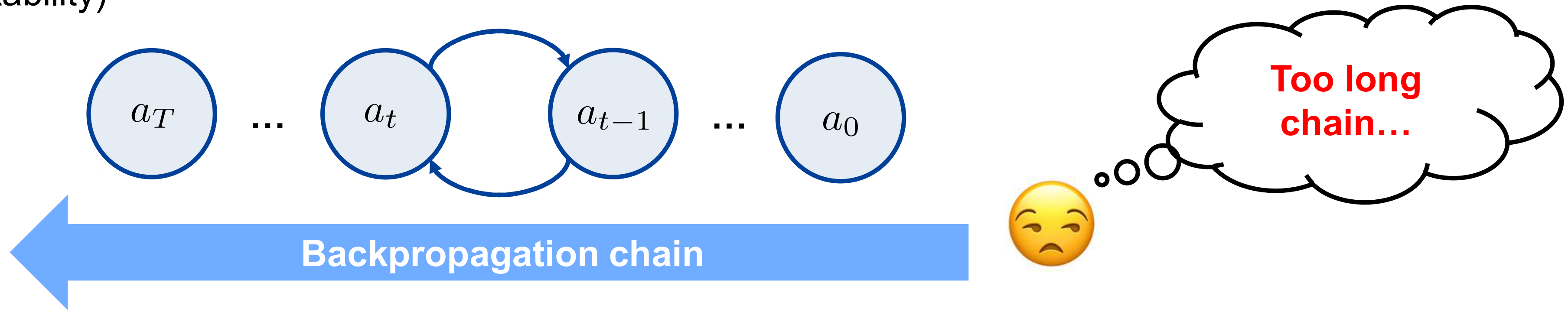
上海科技大学 ShanghaiTech University

Difficulties:

Direction 1: Unavailability of **good** actions (not like imitation learning or offline RL)



Direction 2: Too long backpropagation chain (leads to high training cost and training instability)

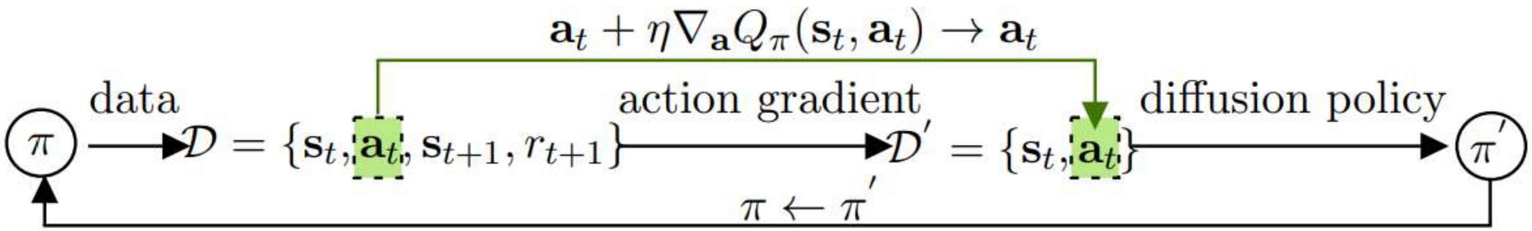


立志成才报国裕民

Background: Diffusion in Online RL

Previous works:

DIPO [1]: Directly perform gradient update on action sample (**affect the multimodality**)



QSM [2]: Do score matching with the gradient of Q function (**Doubled approximation error**)

Update score model:

$$\phi = \operatorname{argmin}_\phi N^{-1} \sum (\Psi_\phi(x_t, a_t) - \nabla_a Q(x_t, a_t))^2;$$

[1] Yang L, Huang Z, Lei F, et al. Policy representation via diffusion probability model for reinforcement learning[J]. arXiv preprint arXiv:2305.13122, 2023.
[2] Psenka M, Escontrela A, Abbeel P, et al. Learning a diffusion model policy from rewards via q-score matching[J]. arXiv preprint arXiv:2312.11752, 2023.

Solution: Q-weighted Variational Policy Optimization

Can we use the variational loss to train diffusion policy?

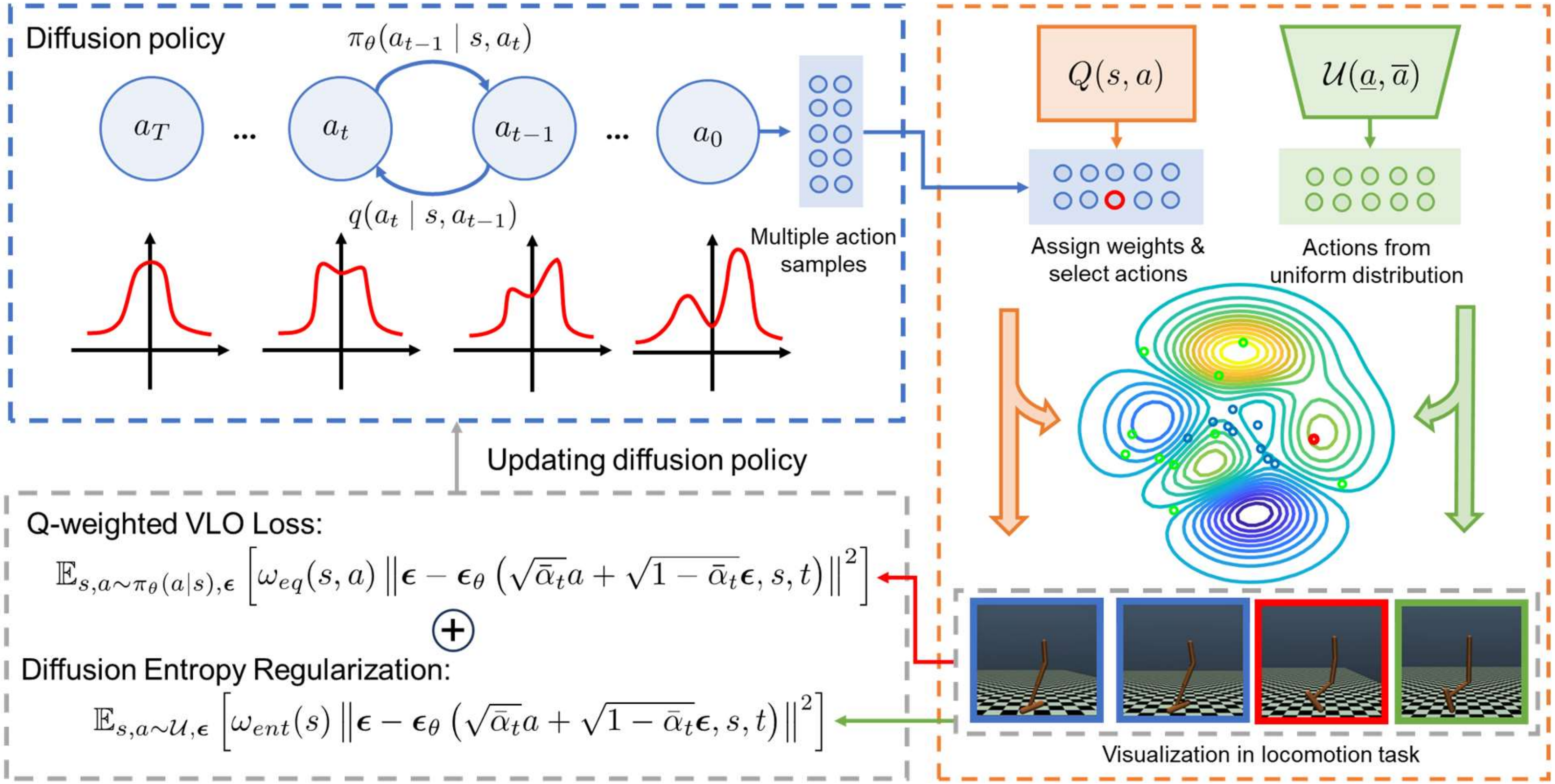
Distinguish bad and good samples according to their **Q-value**?

But samples from replay buffer are still not good enough!

What if use samples from **current policy**?

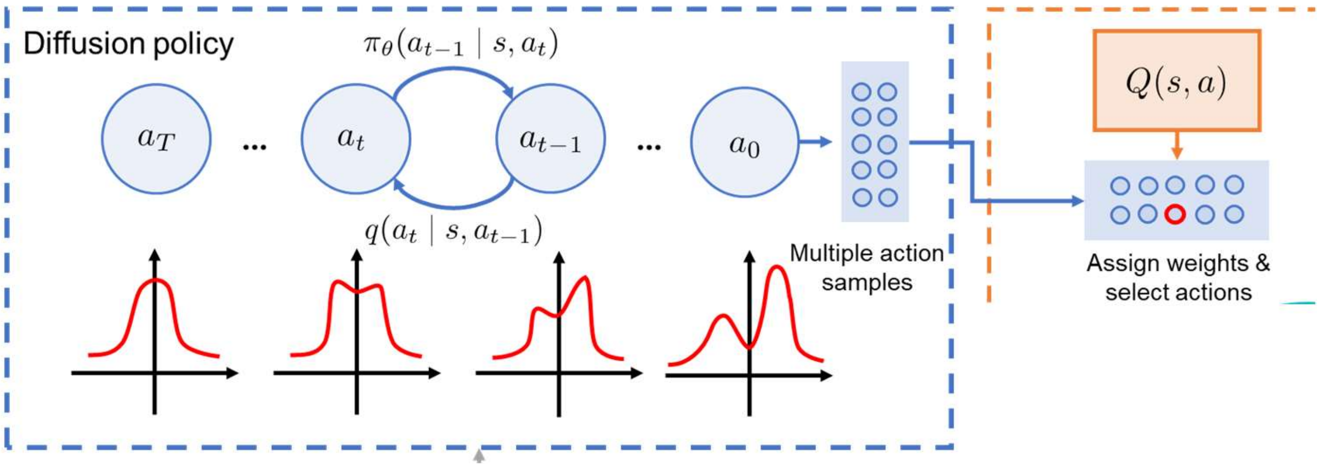
Solution: Q-weighted Variational Policy Optimization

Motivated by these two ideas, we propose Q-weighted variational policy optimization.



Q-weighted Variational Loss

上海科技大学
ShanghaiTech University



How to Further improve the training action quality?

Only choose the best generated sample for training.

Q-weighted VLO Loss (tight lower bound of RL policy objective):

$$\mathbb{E}_{s,a\sim\pi_\theta(a|s),\epsilon}\left[\omega_{eq}(s,a)\left\|\epsilon-\epsilon_\theta\left(\sqrt{\bar{\alpha}_t}a+\sqrt{1-\bar{\alpha}_t}\epsilon,s,t\right)\right\|^2\right]$$

Since the weight must be nonnegative, we define the weight as

$$\omega_{eq}(s,a)\triangleq\omega_{qadv}(s,a)=\begin{cases}A(s,a), & A(s,a)\geq 0\\ 0, & A(s,a)<0\end{cases},$$

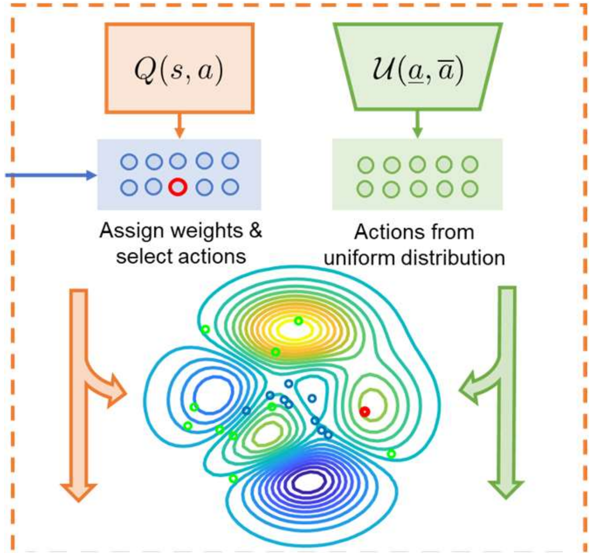
立志成才报国裕民

Diffusion Entropy Regularization

We also add Diffusion Entropy Regularization term in objective for enhancing exploration capability:

$$\mathbb{E}_{\boxed{s,a\sim\mathcal{U},\boldsymbol{\epsilon}}}\left[\omega_{ent}(s)\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}a+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},s,t\right)\right\|^2\right]$$

- Maximizing entropy can be viewed as **minimizing the distance** between current policy and the **uniform distribution**.
- Hence, we use samples from the uniform distribution to train diffusion policy for maximizing the diffusion entropy.

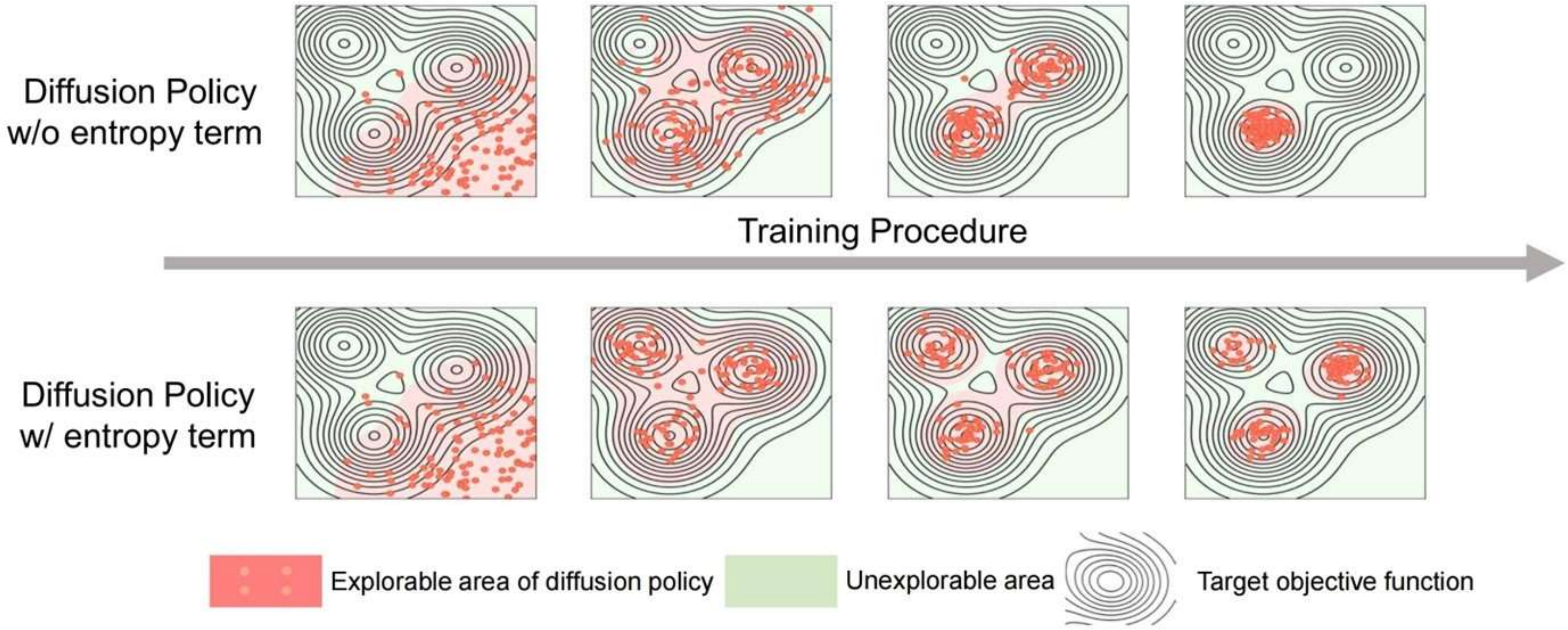


Comparison w/ & w/o Entropy Term

上海科技大学
ShanghaiTech University

We also add Diffusion Entropy Regularization term in objective for enhancing exploration capability:

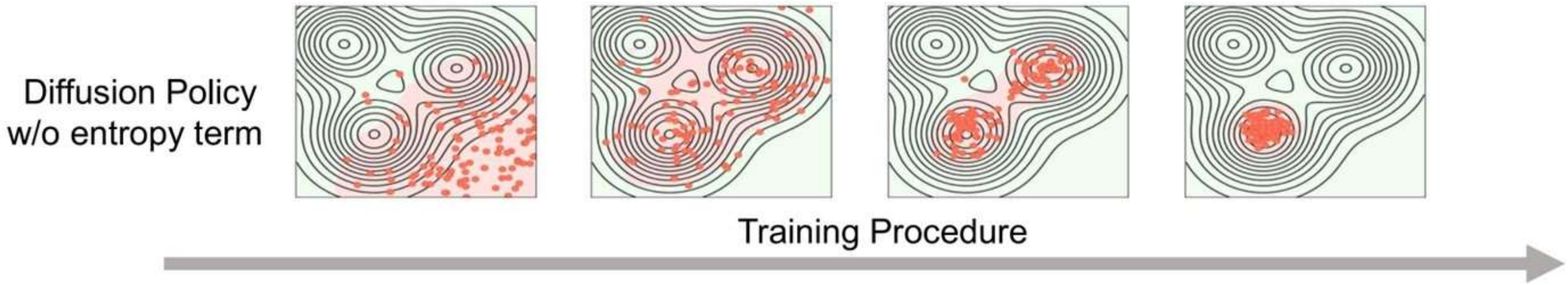
$$\mathbb{E}_{s,a\sim\mathcal{U},\boldsymbol{\epsilon}}\left[\omega_{ent}(s)\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}a+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},s,t\right)\right\|^2\right]$$



立志成才 报国裕民

Reducing Diffusion Policy Variance via Action Selection

- Diffusion policy has a **large policy variance**
- Reducing this variance can improve **the quality** of collected transitions

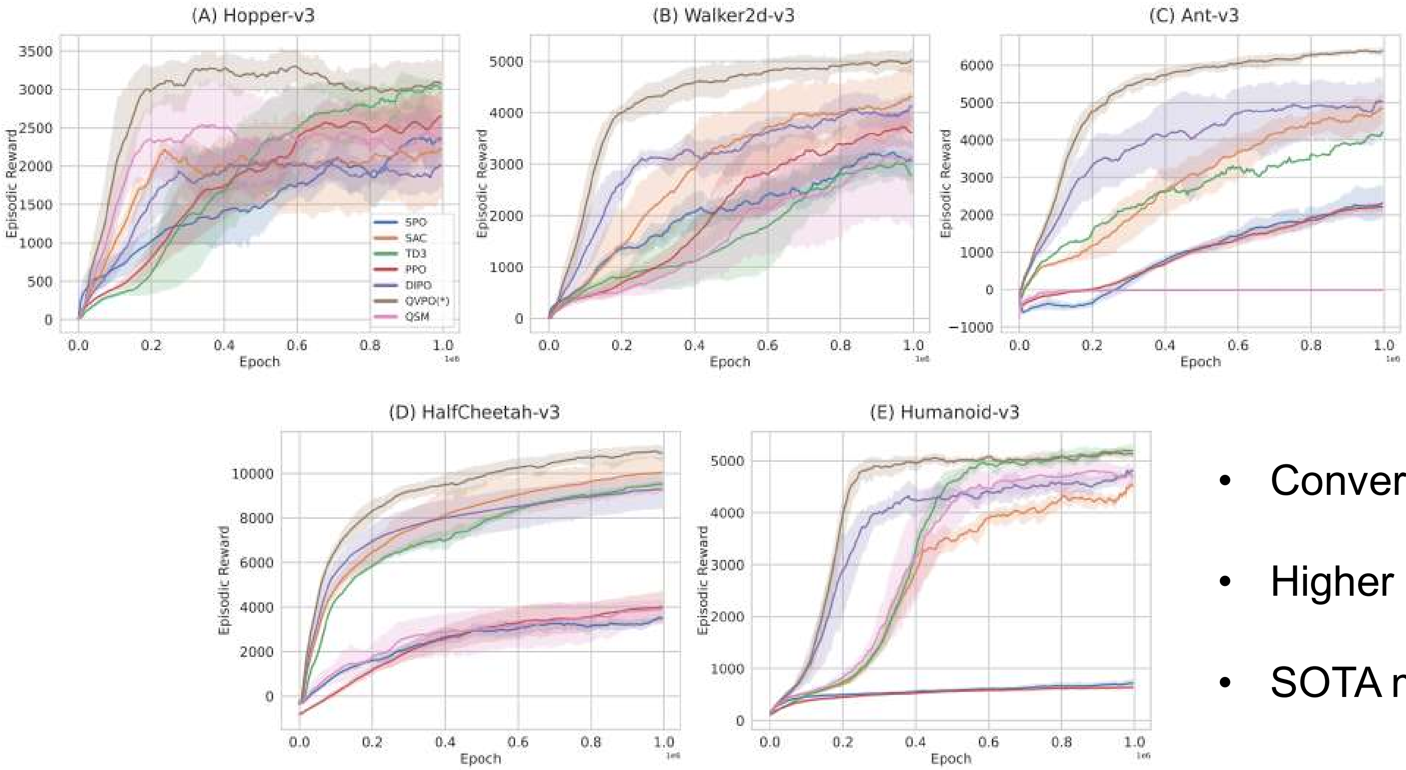


We design **efficient behavior policy** via action selection for sample efficiency:

$$\pi_\theta^K(a \mid s) \triangleq \underset{a \in \{a_1, \cdots, a_K \sim \pi_\theta(a|s)\}}{\operatorname{argmax}} Q(s, a).$$

Results

上海科技大学
ShanghaiTech University



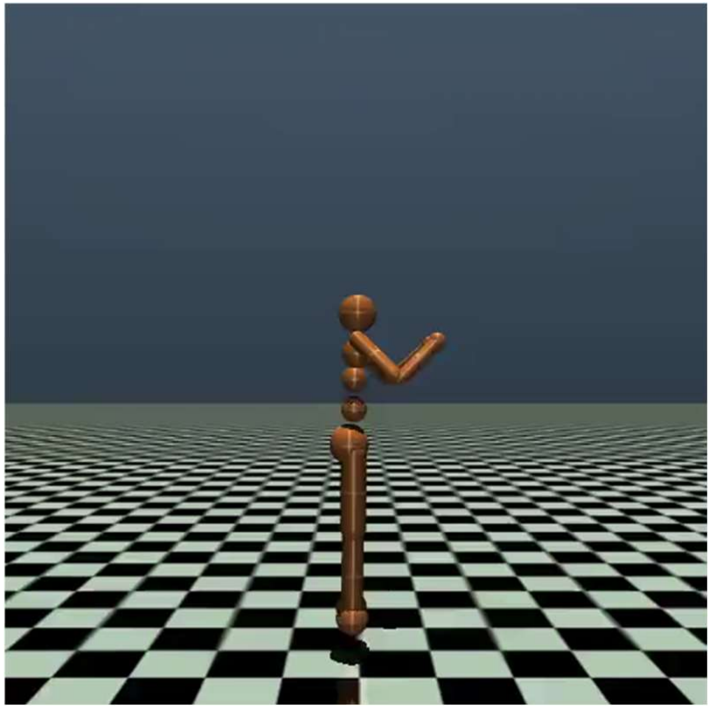
- Converge faster (sample efficiency)
- Higher cumulative reward
- SOTA method in online RL

立志成才 报国裕民

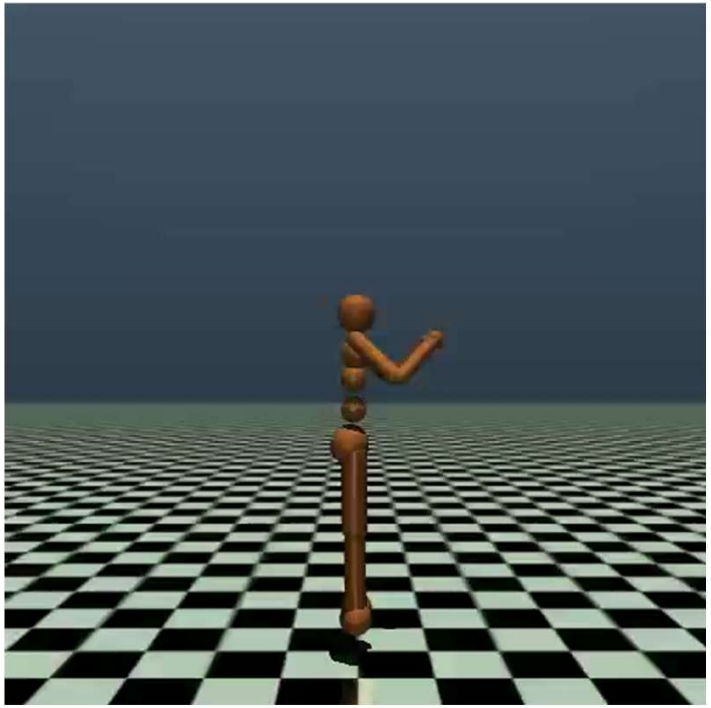
Comparison with SAC in Humanoid-v3

上海科技大学
ShanghaiTech University

SAC



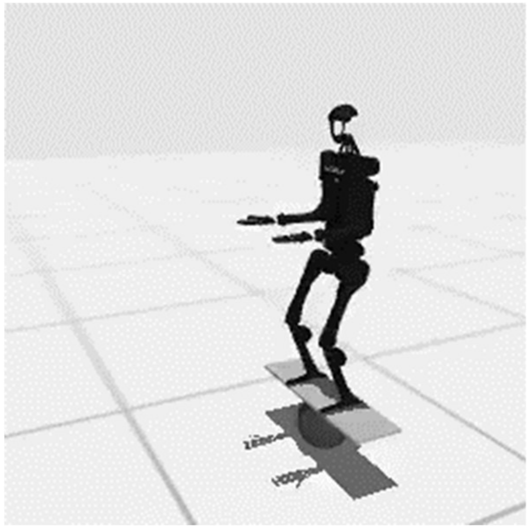
our QVPO



立志成才报国裕民

More Tasks in Unitree Humanoid H1

Balance (TD-MPC2)

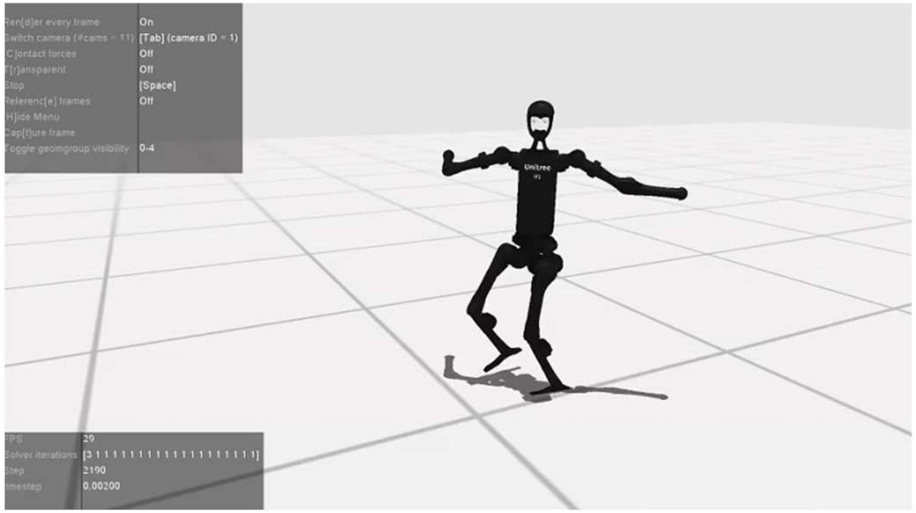


Balance (our QVPO)

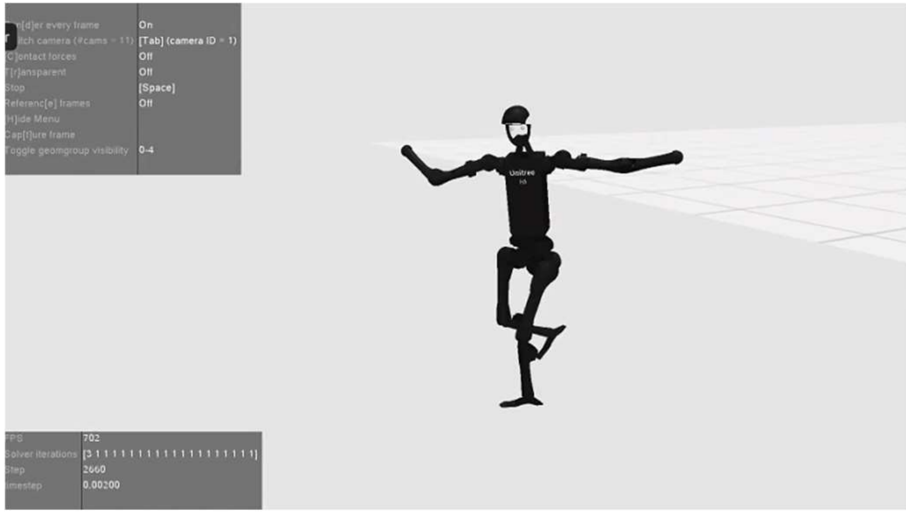
More Tasks in Unitree Humanoid H1

上海科技大学
ShanghaiTech University

Walk (QVPO)



Run (QVPO)



立志成才报国裕民