Tackling Continual Offline RL through Selective Weights Activation on Aligned Spaces

Jifeng Hu, Sili Huang, Li Shen, Zhejian Yang, Shengchao Hu, Shisong Tang, Hechang Chen, Lichao Sun, Yi Chang, Dacheng Tao

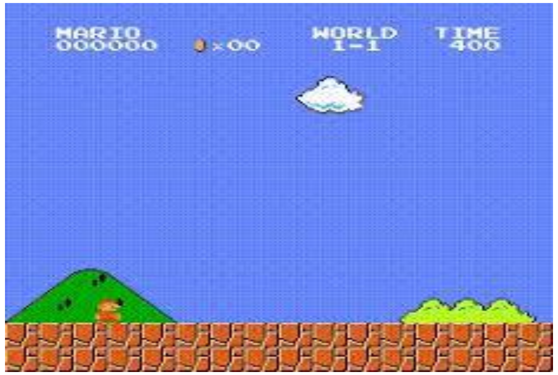
Jifeng Hu
2025

School of Artificial Intelligence, Jilin University, Changchun, China

1. Artificial General Intelligence

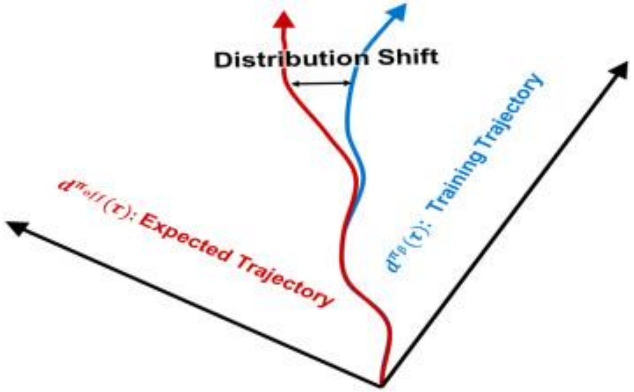
Offline learning

Continual learning

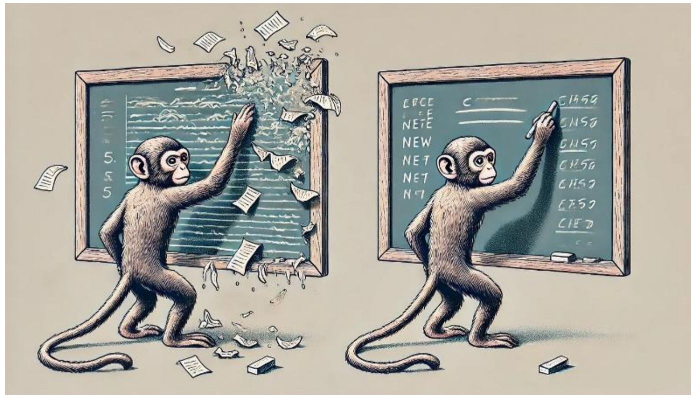


Cross-task learning and deployment

1. Continual Offline RL



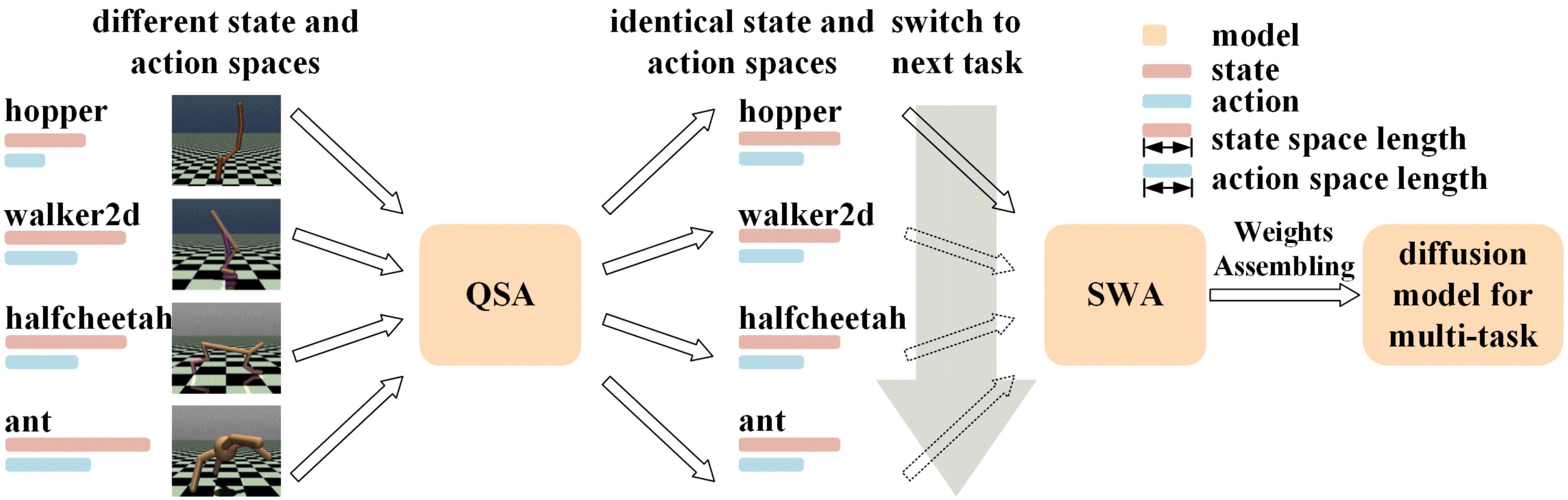
Distribution shift



Catastrophic forgetting

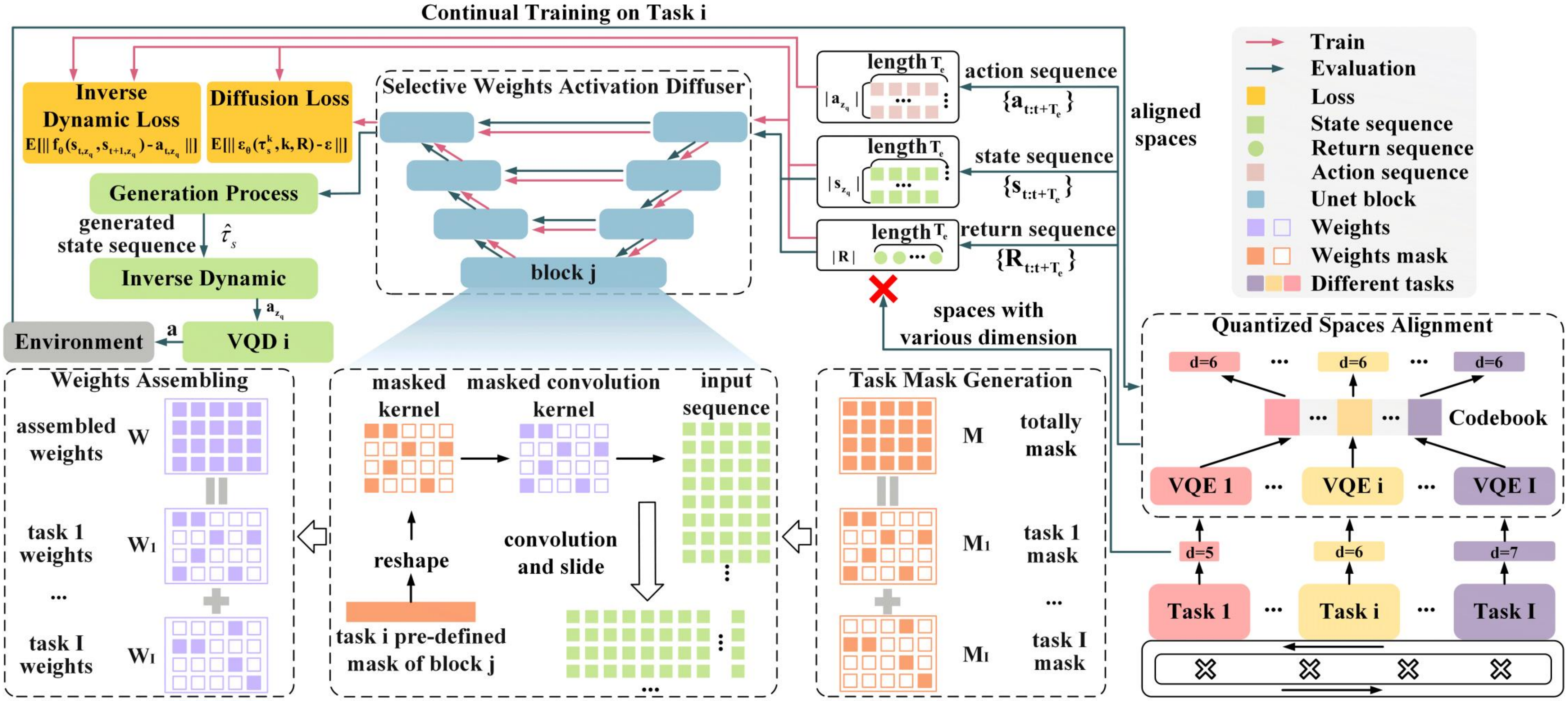
Inconsistency of state and action spaces

2. Method: high-level intuition of VQ-CD



2. Method: detailed structure

- QSA provides the unified space
- SWA preserves plasticity for new tasks and retains knowledge for previous tasks
- Diffusion model structure posses the ability to fit complex offline data distribution



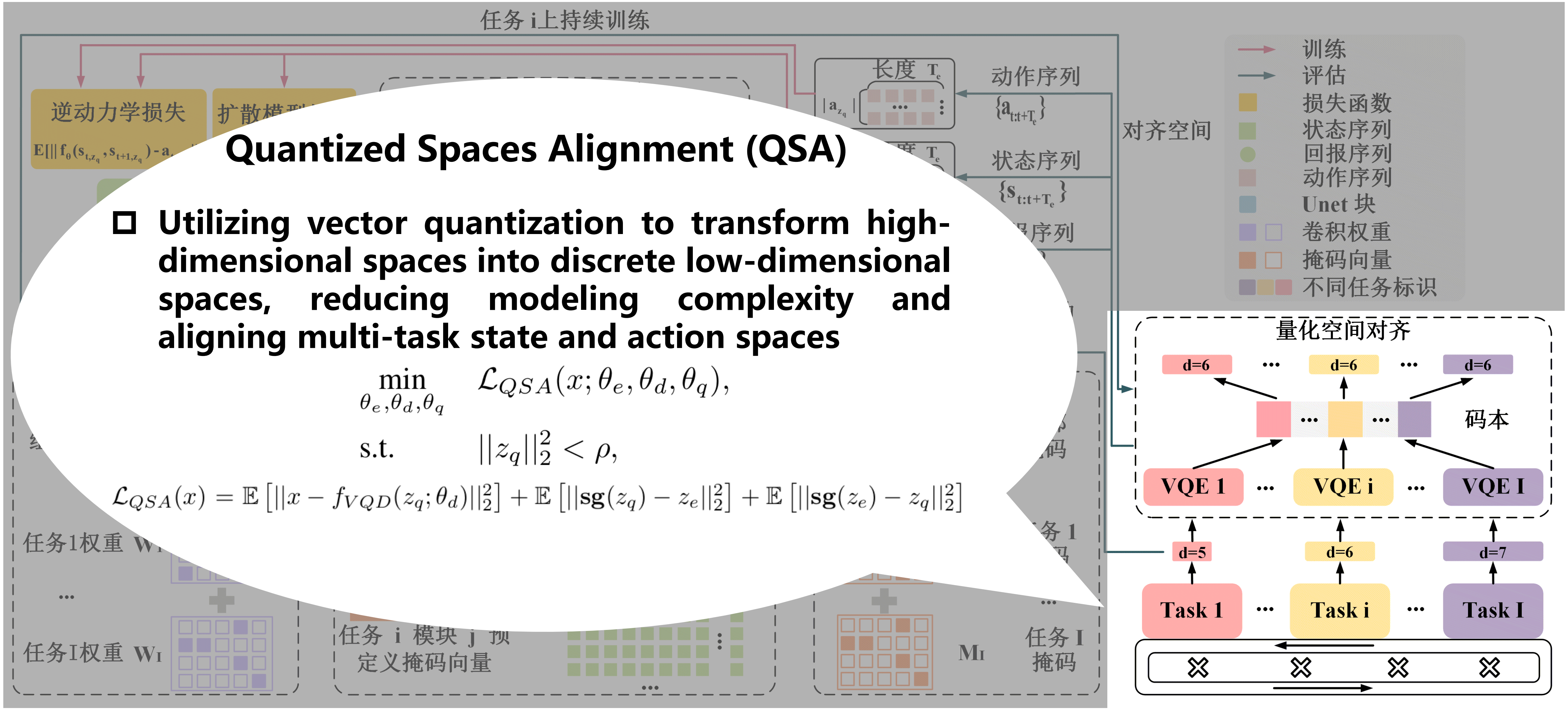
2. Method: QSA

Quantized Spaces Alignment (QSA)

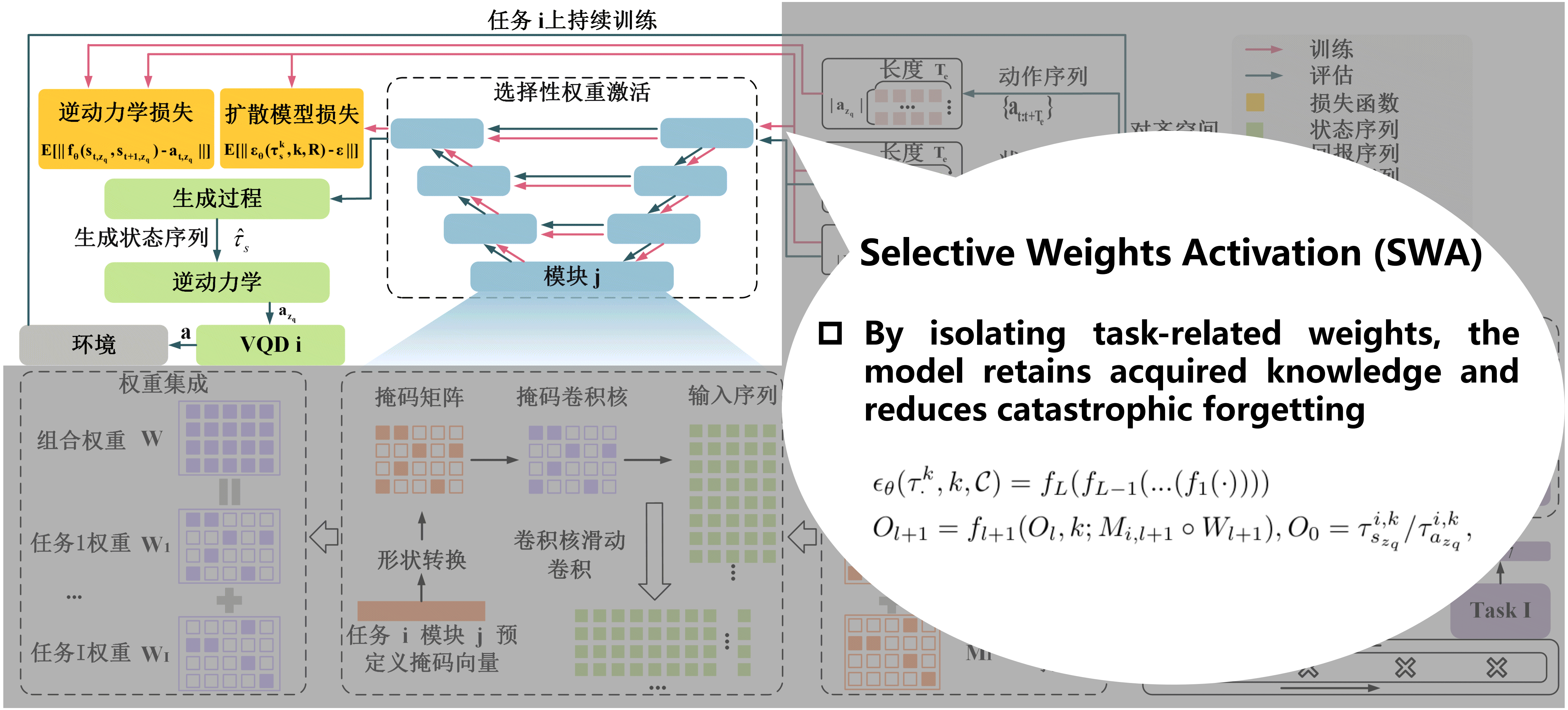
- Utilizing vector quantization to transform high-dimensional spaces into discrete low-dimensional spaces, reducing modeling complexity and aligning multi-task state and action spaces

$$\min_{\theta_e,\theta_d,\theta_q} \quad \mathcal{L}_{QSA}(x;\theta_e,\theta_d,\theta_q),$$
$$\text{s.t.} \quad ||z_q||_2^2 < \rho,$$

$$\mathcal{L}_{QSA}(x) = \mathbb{E}\left[||x - f_{VQD}(z_q;\theta_d)||_2^2\right] + \mathbb{E}\left[||\mathbf{sg}(z_q) - z_e||_2^2\right] + \mathbb{E}\left[||\mathbf{sg}(z_e) - z_q||_2^2\right]$$



2. Method: SWA



Selective Weights Activation (SWA)

- By isolating task-related weights, the model retains acquired knowledge and reduces catastrophic forgetting

$$\epsilon_\theta(\tau_\cdot^k, k, \mathcal{C}) = f_L(f_{L-1}(...(f_1(\cdot))))$$

$$O_{l+1} = f_{l+1}(O_l, k; M_{i,l+1} \circ W_{l+1}), O_0 = \tau_{s_{zq}}^{i,k} / \tau_{a_{zq}}^{i,k},$$

3. Experiments

Experiments:
Environments with identical state and action spaces: MuJoCo Ant-dir and Continual World (CW)
Environments with various state and action spaces: D4RL (Hopper, Walker2d, and HalfCheetah)

Evaluation Metrics:
Mean performance on all tasks: $P = \text{mean}(\sum_i \Psi_i)$
Normalized score: $\Phi_i = \frac{R_i - R_{random}}{R_{expert} - R_{random}} * 100$

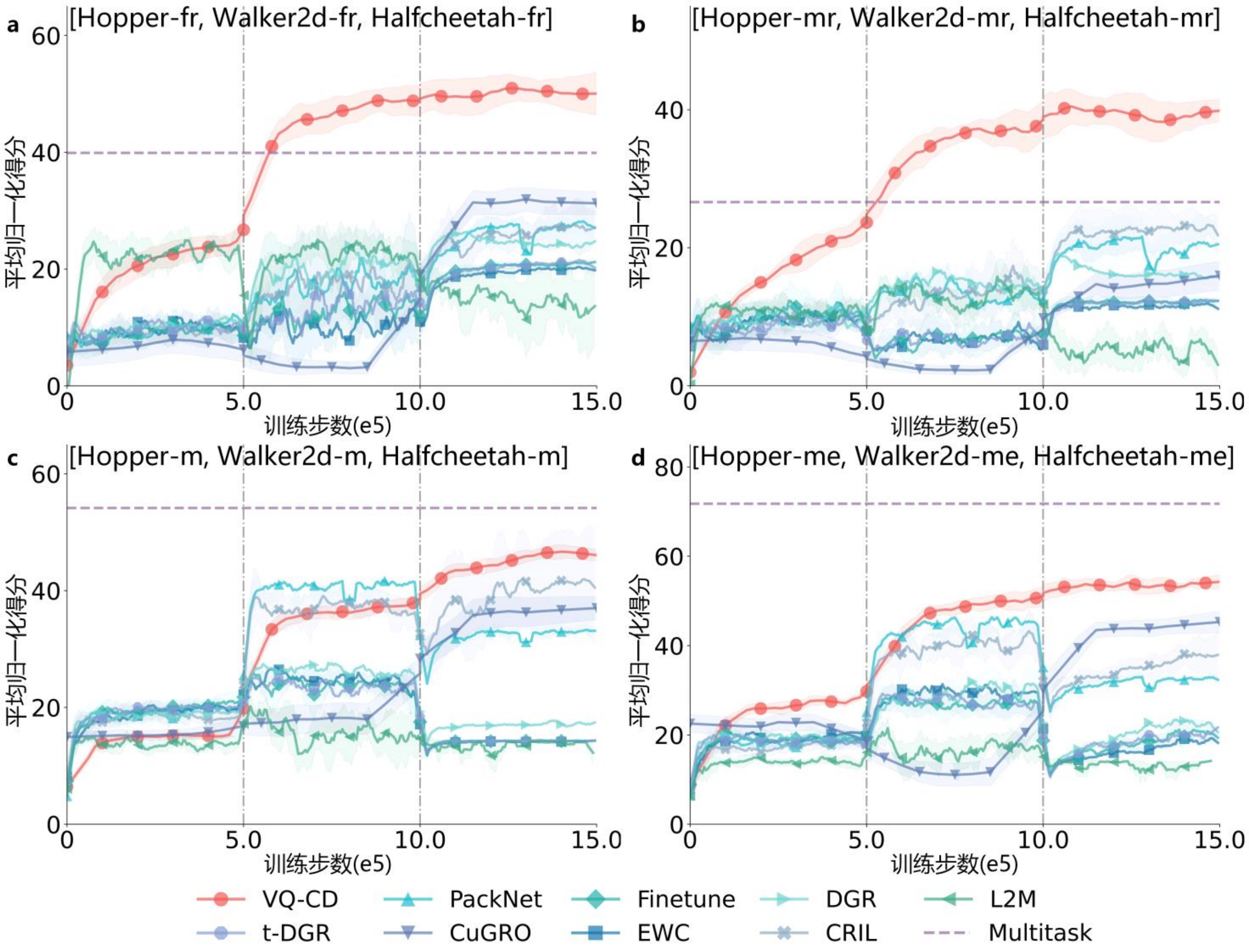
Baselines:
Diffusion-based methods: CRIL, DGR, t-DGR, MTDIFF, CuGRO, and CoD
Non-diffusion-based methods: L2M, EWC, PackNet, Finetune, IL-rehearsal, and Multitask

Rehearsal-based methods: CRIL, DGR, t-DGR, CoD, and IL-rehearsal
Regularization-based methods: L2M, EWC, CuGRO, and Finetune
Structure-based methods: PackNet, Multitask, and MTDIFF

3. Experiments: main results

Experimental setting:
To test the model's learning capability in scenarios with inconsistent state and action spaces

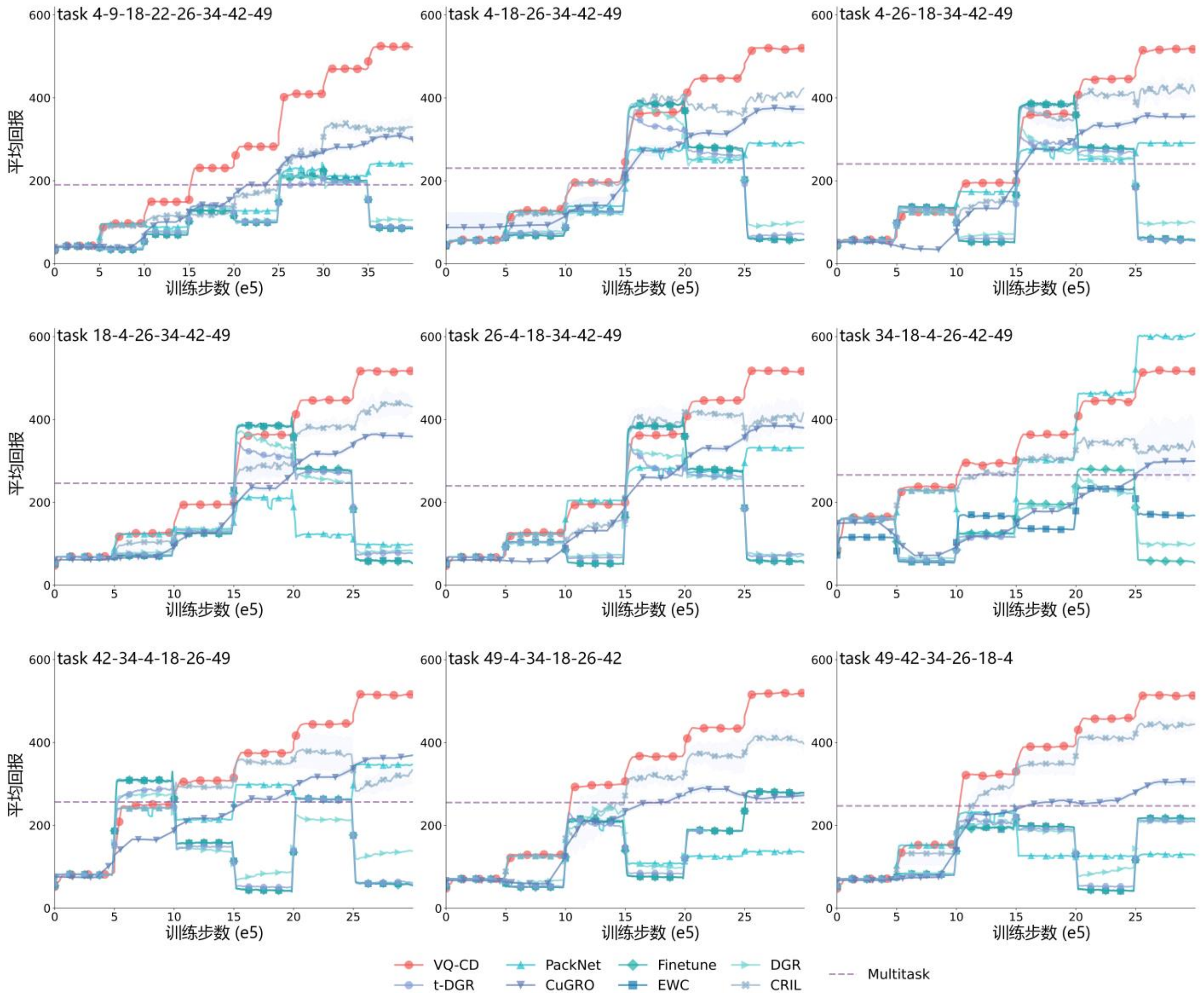
Discussion:
Under task settings with diverse state and action spaces, our method can effectively learn in the aligned space and achieve optimal performance



3. Experiments: tasks order

Experimental setting:
To test whether the model is sensitive to the order of multi-task sequences

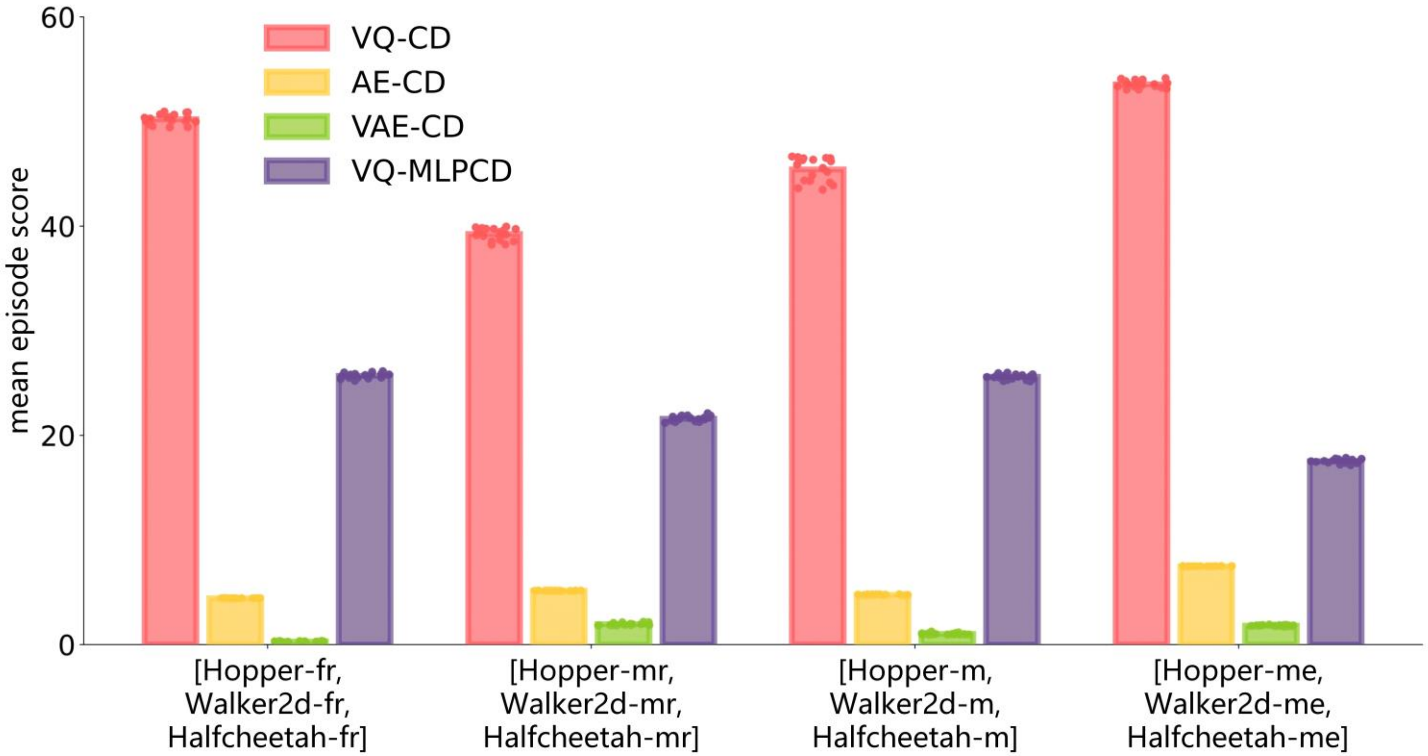
Discussion:
Regardless of how the task order is rearranged, our method can still achieve the optimal performance



3. Experiments: ablation

Experimental setting:
To investigate the influence of different modules of VQ-CD

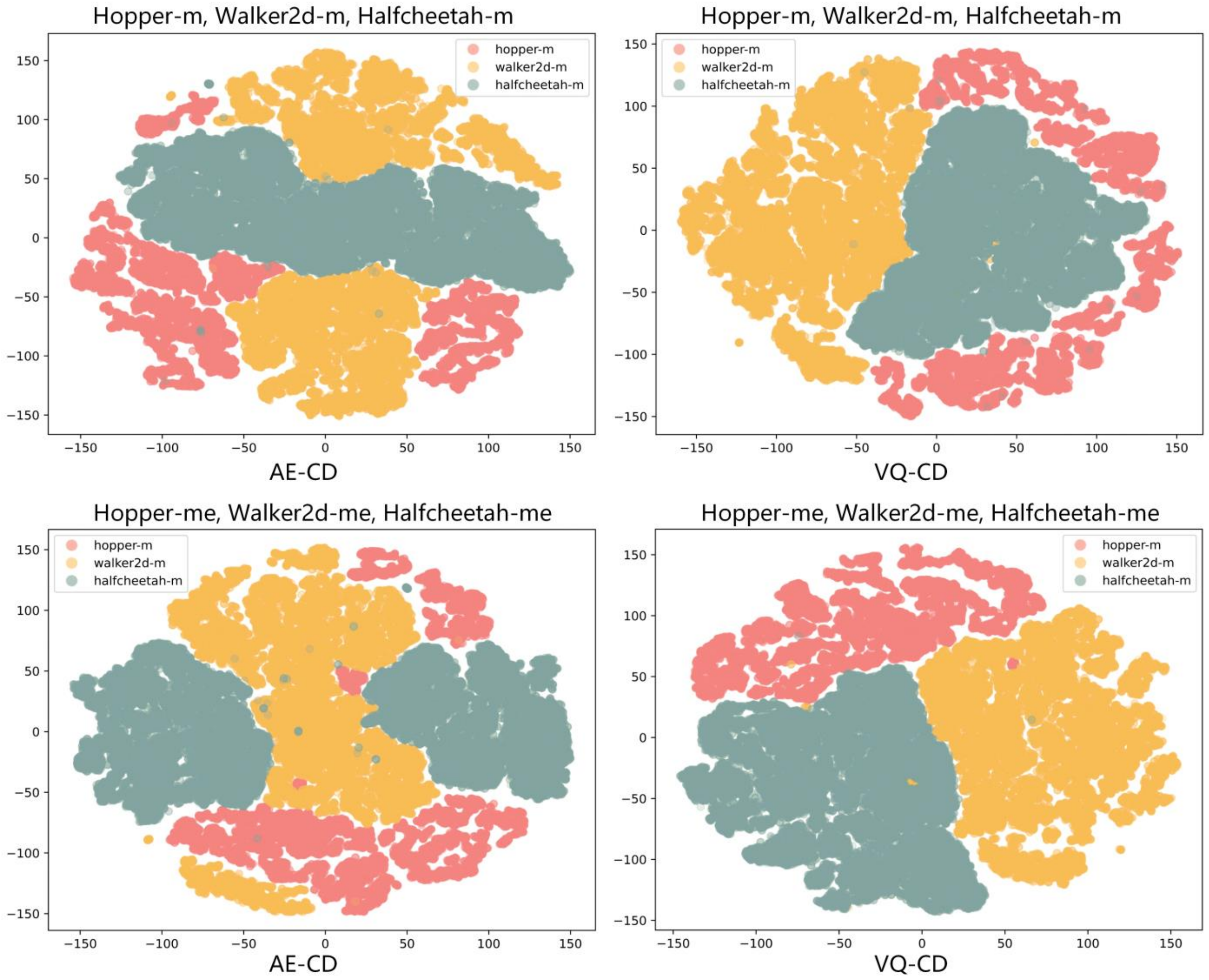
Discussion:
The results show significant improvements in the D4RL CL settings, illustrating the importance and effectiveness of vector quantization in our method.
Compared with AE-CD, VAE-CD performs poorly on all D4RL CL settings. The reason lies in that the implicit Gaussian constraint on each dimension may hurt the space alignment.



3. Experiments: space alignment

Experimental setting:
To test the model's feature modeling capability for multiple state and action spaces

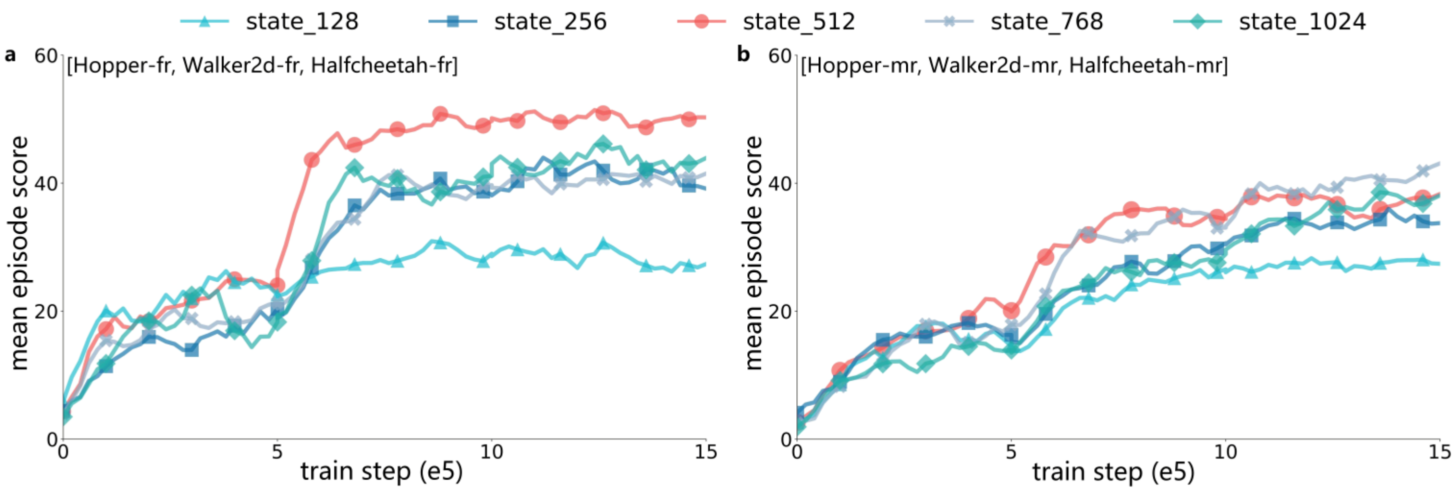
Discussion:
The vector quantization space alignment proposed by this method can significantly distinguish state space representations across different tasks

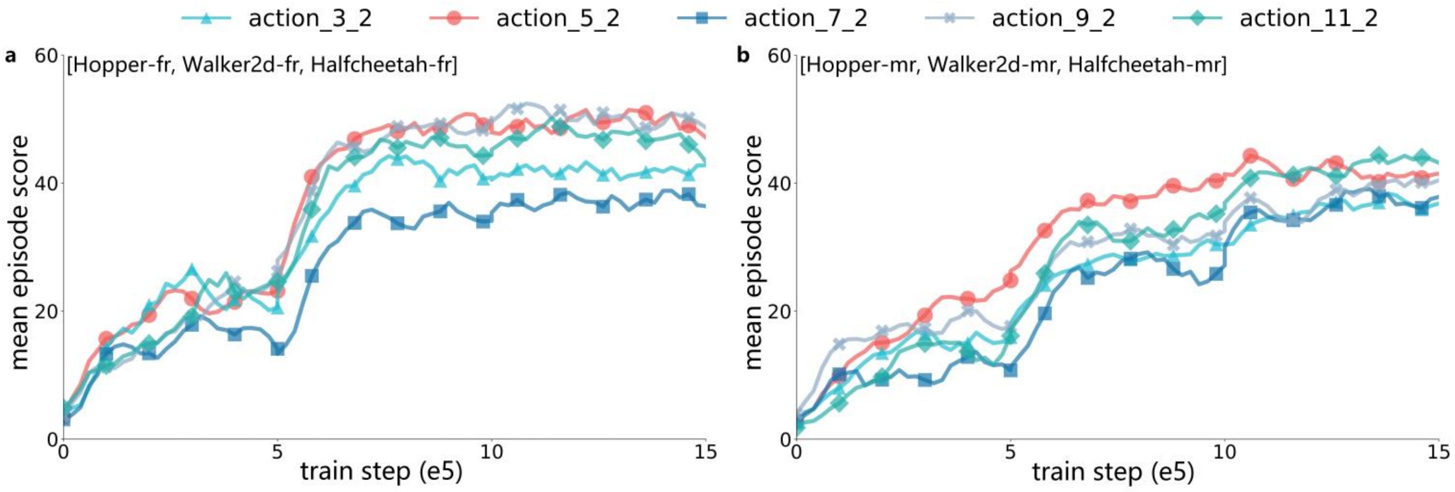


3. Experiments: parameter sensitivity

Experimental setting:
To investigate the influence of codebook size

Discussion:
A small codebook size limits performance, and a negative effect arises when it exceeds a certain value, such as 512.
we also see the same trend in the results, which show that more latent vectors are not always better.





4. Conclusion

- We propose the Vector-Quantized Continual Diffuser (VQ-CD) framework, which can not only be applied to conventional continual tasks but also be suitable for any continual tasks setting, which makes it observably different from the previous CL method.
- In the quantized spaces alignment (QSA) module of VQ-CD, we adopt ensemble vector quantized encoders based on the constrained codebook because it can be expanded expediently. During the inference, we apply task-related decoders to recover the various observation and action spaces.
- In the selective weights activation (SWA) diffuser module of VQ-CD, we first perform task-related task masks, which will then be used to the kernel weights of the diffuser. After training, we propose assembling weights to merge all learned knowledge.
- Finally, we conduct extensive experiments on 15 CL tasks, including conventional CL settings and any CL task sequence settings. The results show that our method surpasses or matches the SOTA performance compared with 17 representative baselines.

Thanks

School of Artificial Intelligence
Jifeng Hu