Mixture-of-Experts Meets In-Context Reinforcement Learning

Wenhao Wu (吴文浩)

Department of Control Science and Intelligent Engineering
Nanjing University

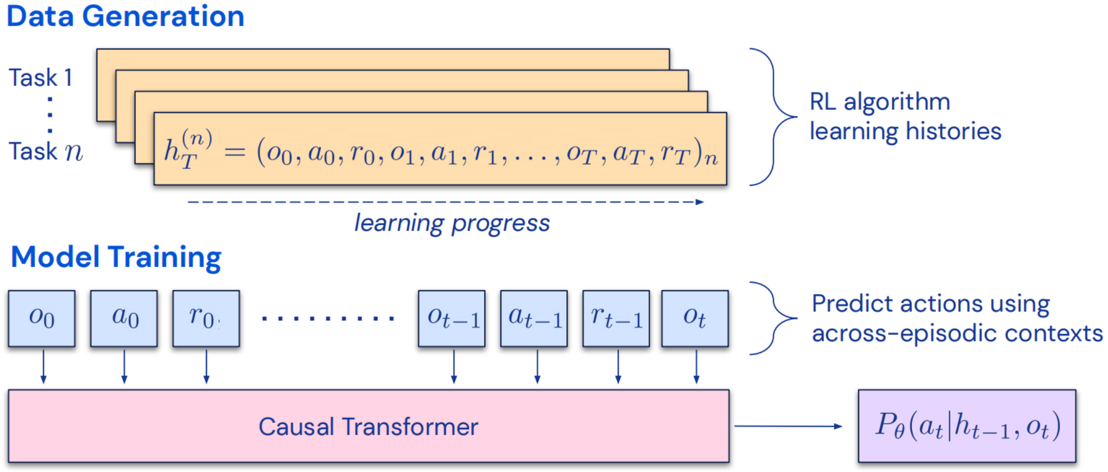
Email: wenhaowu@smail.nju.edu.cn

Code: https://github.com/NJU-RL/T2MIR

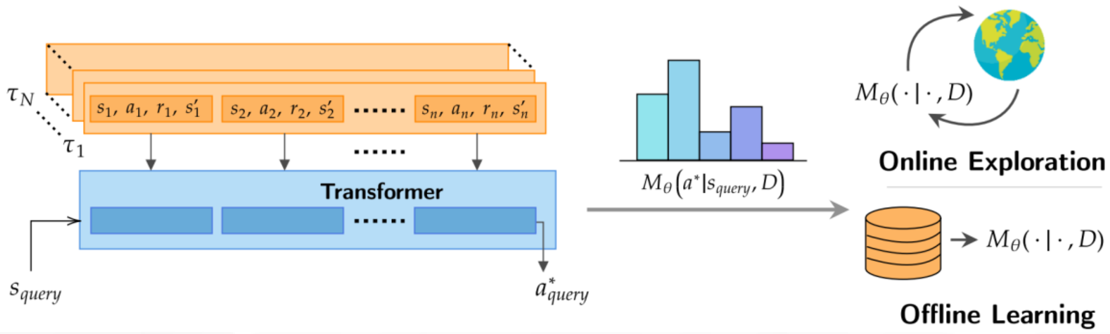
Contents

1. In-context RL

Algorithm Distillation (AD)



Decision-Pretrained Transformer (DPT)



- **Algorithm Distillation**
  - train a causal transformer to predict actions given preceding learning histories as context
  - cross-episodic trajectories
  - a dataset of learning histories is generated by a source RL algorithm

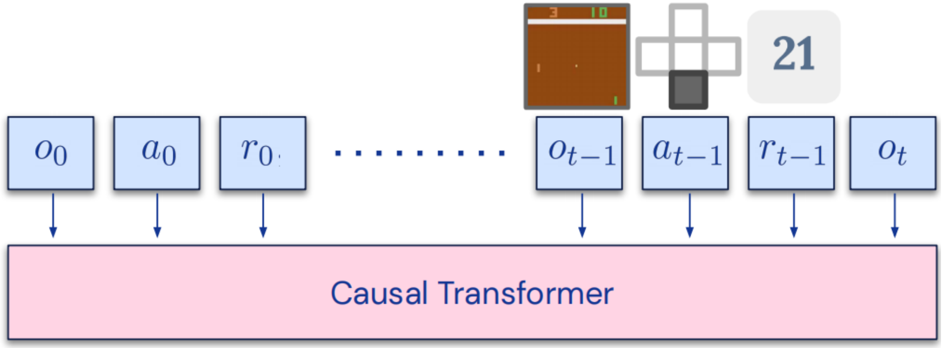
$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\sum_{t=0}^{T-1} \log P_\theta\left(a = a_t^n \middle| \tau_{\text{pro}}^n, s_t^n\right), \tau_{\text{pro}}^n = h_{t-1}^n$$

- **Decision-Pretrained Transformer**
  - predict the optimal action given a query state and a prompt of interactions
  - need expert policy to label actions
  - robust to different dataset qualities

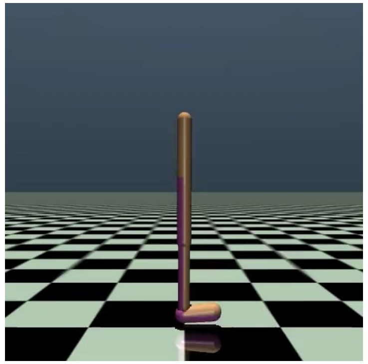
$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\sum_{t=0}^{T} \log P_\theta\left(a = a_t^{n*} \middle| \tau_{\text{pro}}^n, s_t^n\right), \tau_{\text{pro}}^n \sim \mathcal{D}^n$$

1. Limitations in In-context RL

multi-modality in state-action-reward sequence

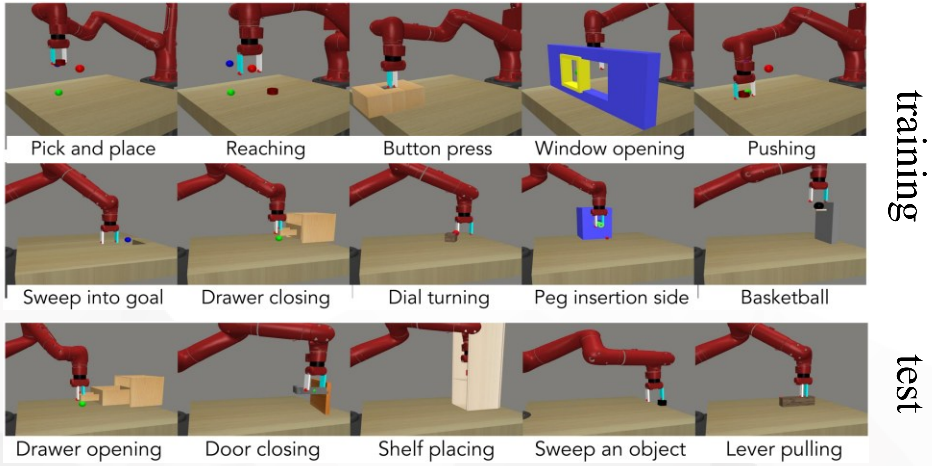


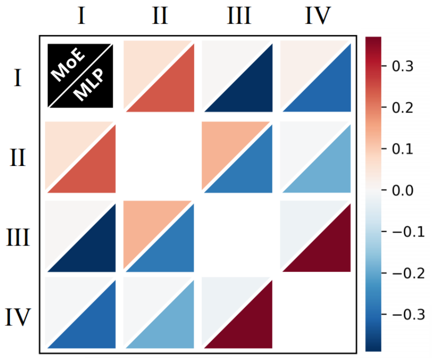
$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T, a_T, r_T)$$



Walker-Param

task diversity and heterogeneity





cosing similarity

- **multi-modality in state-action-reward sequence**
  - states: physical quantities (position, velocity, and acceleration)
  - actions: joint torques or discrete commands
  - rewards: simple scalars
- **task diversity and heterogeneity**
  - some tasks are similar and others differ significantly
  - intrinsic gradient conflicts in challenging scenarios with significant task variation

2. Overview of T2MIR

- **two parallel MoE layers**

- Token-wise MoE

✓ balance loss

✓ enables different experts to process tokens with distinct semantics

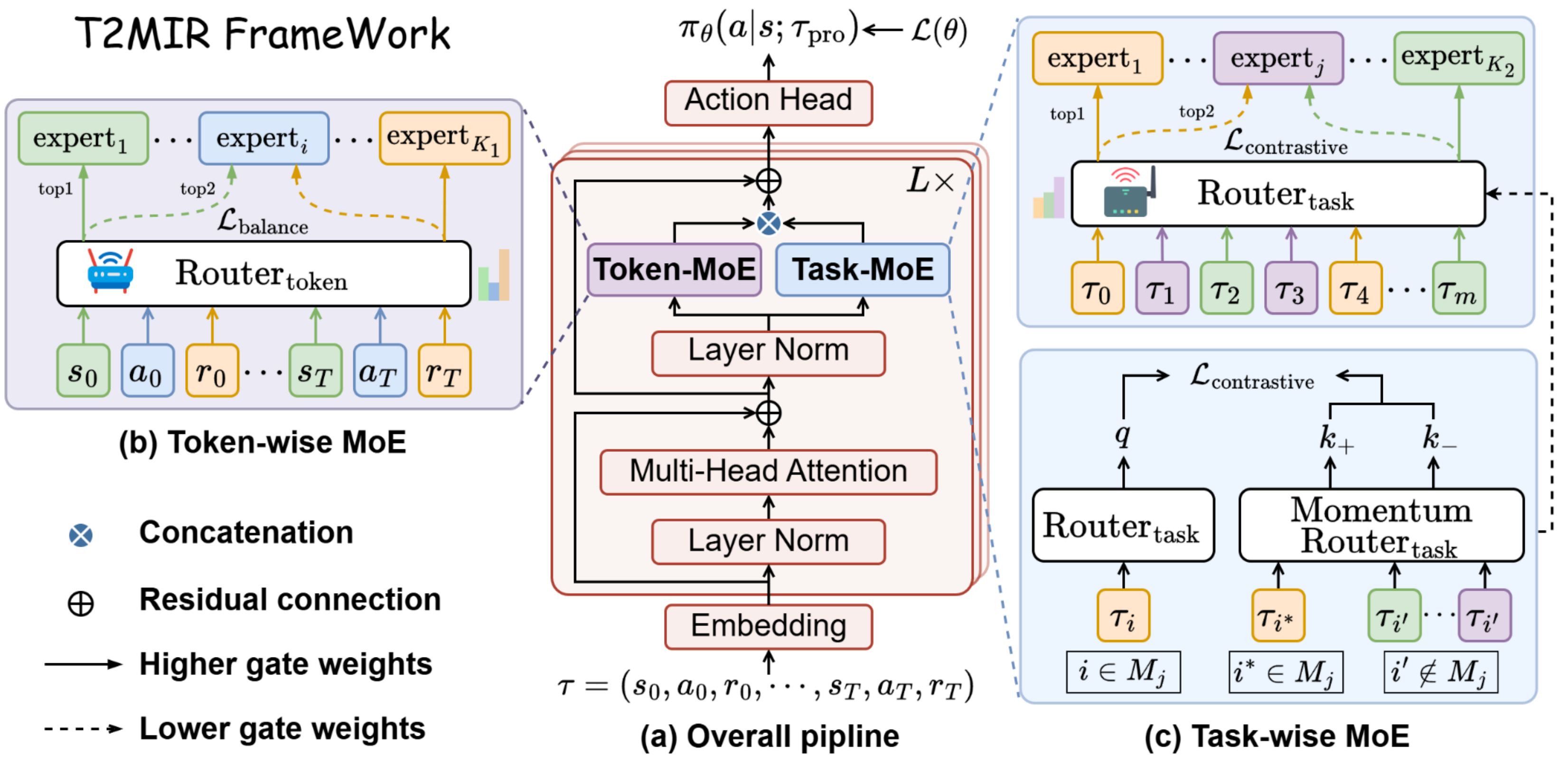
- Task-wise MoE

✓ InfoNCE loss

✓ effectively manages a broad task distribution

✓ alleviate gradient conflicts

The overview of T2MIR



$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\sum_{t=0}^{T-1} \log P_\theta\left(a = a_t^n \middle| \tau_{\text{pro}}^n, s_t^n\right), \tau_{\text{pro}}^n = h_{t-1}^n$$

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\sum_{t=0}^{T} \log P_\theta\left(a = a_t^{n*} \middle| \tau_{\text{pro}}^n, s_t^n\right), \tau_{\text{pro}}^n \sim \mathcal{D}^n$$

2. Token-wise MoE

- **motivation: intrinsic multi-modality in state-action-reward sequence**
  - semantic gap among states, actions and rewards
  - router G_{tok} learns to assign each token to specific experts at the token level

$$y_{\text{tok}} = \sum_{i=1}^{K_1} w_{\text{tok}}(i;h) \cdot E_{\text{tok}}(i|h)$$

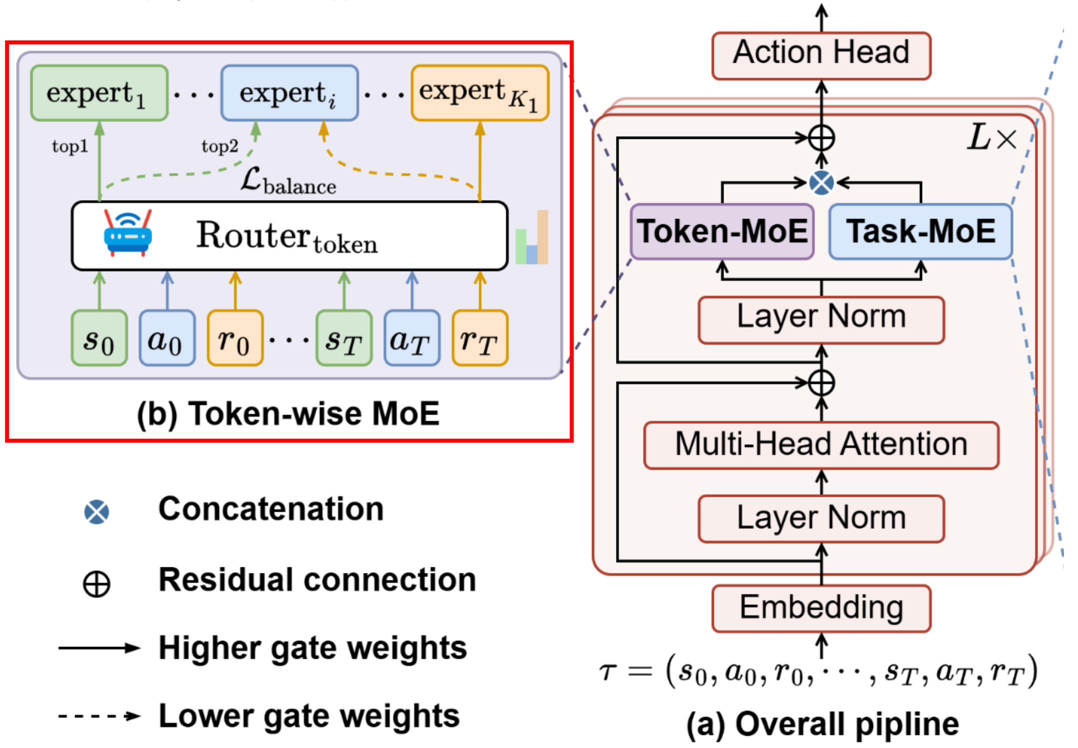
$$w_{\text{tok}}(i;h) = \text{softmax}\Big(\text{topk}\big(G_{\text{tok}}(i|h)\big)\Big)[i]$$

  - balance expert utilization with balance loss

$$\mathcal{L}_{\text{balance}} = w_{\text{imp}} \cdot CV\big(\text{Imp}(h)\big)^2 + w_{\text{load}} \cdot CV\big(\text{Load}(h)\big)^2$$

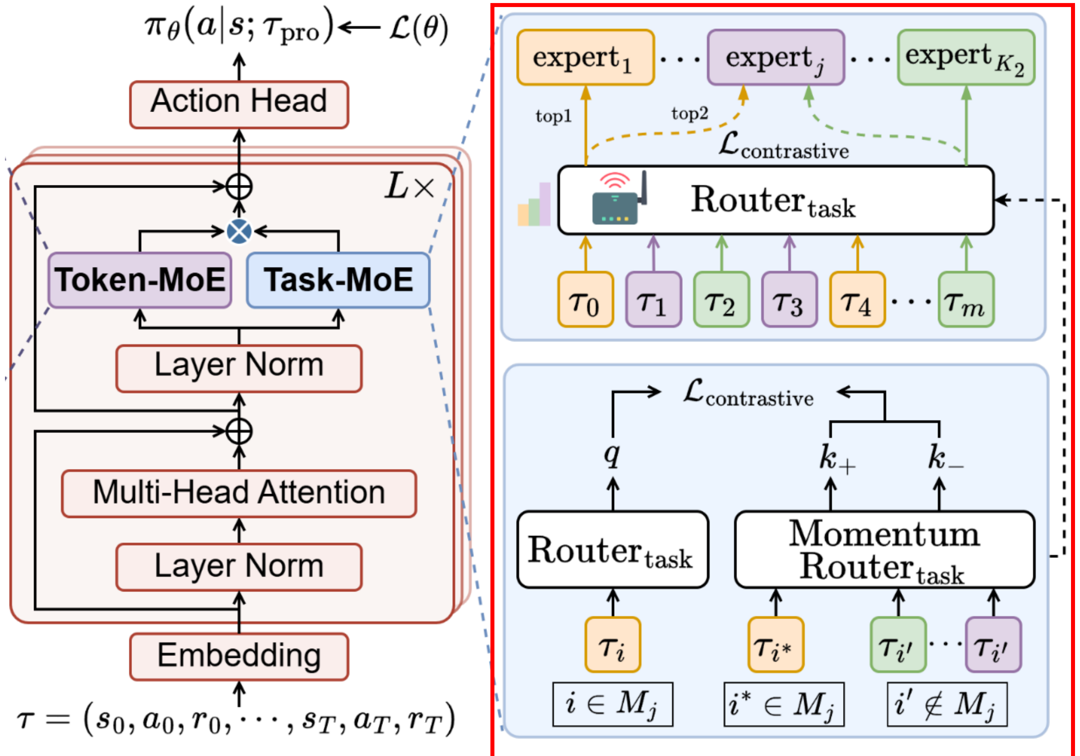
Token-wise MoE

T2MIR FrameWork



2. Task-wise MoE

Task-wise MoE



(a) Overall pipline

(c) Task-wise MoE

➢ **motivation: task diversity and heterogeneity**

- some tasks are similar and others differ significantly
- learning efficiency can be impeded by intrinsic gradient conflicts in scenarios with significant task variation
- router G_{task} learns to assign tokens to specialized experts at the task level

$$y_{\text{task}} = \sum_{i=1}^{K_2} w_{\text{task}}(i;\bar{h}) \cdot E_{\text{task}}(i|\bar{h})$$

$$w_{\text{task}}(i;\bar{h}) = \text{softmax}\left(\text{topk}\left(G_{\text{task}}(i|\bar{h})\right)\right)[i]$$

- view G_{task} as a task encoder, balance expert utilization with InfoNCE loss

$$\mathcal{L}_{\text{contrastive}} = \frac{\sum_{i^* \in M_j} \exp(z_i^\top W z_{i^*})}{\sum_{i^* \in M_j} \exp(z_i^\top W z_{i^*}) + \sum_{i' \notin M_j} \exp(z_i^\top W z_{i'})}$$

3. Experiments

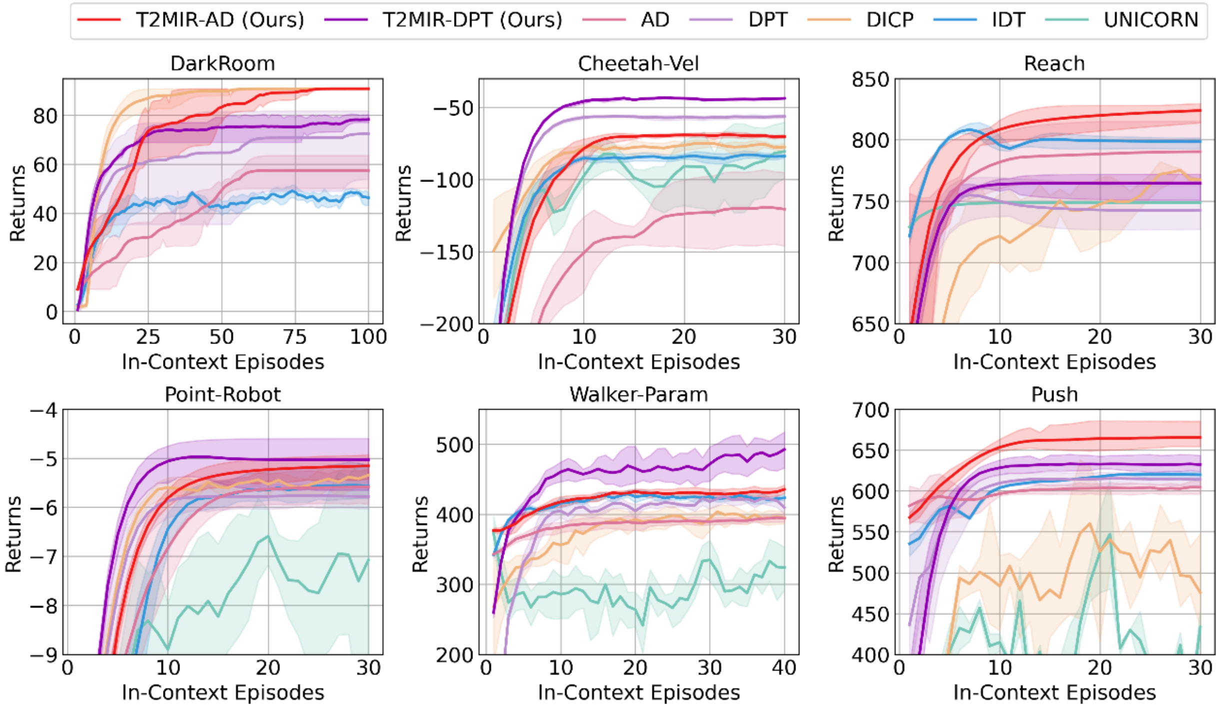


Figure 3: Test return curves of two T2MIR implementations against baselines using Mixed datasets

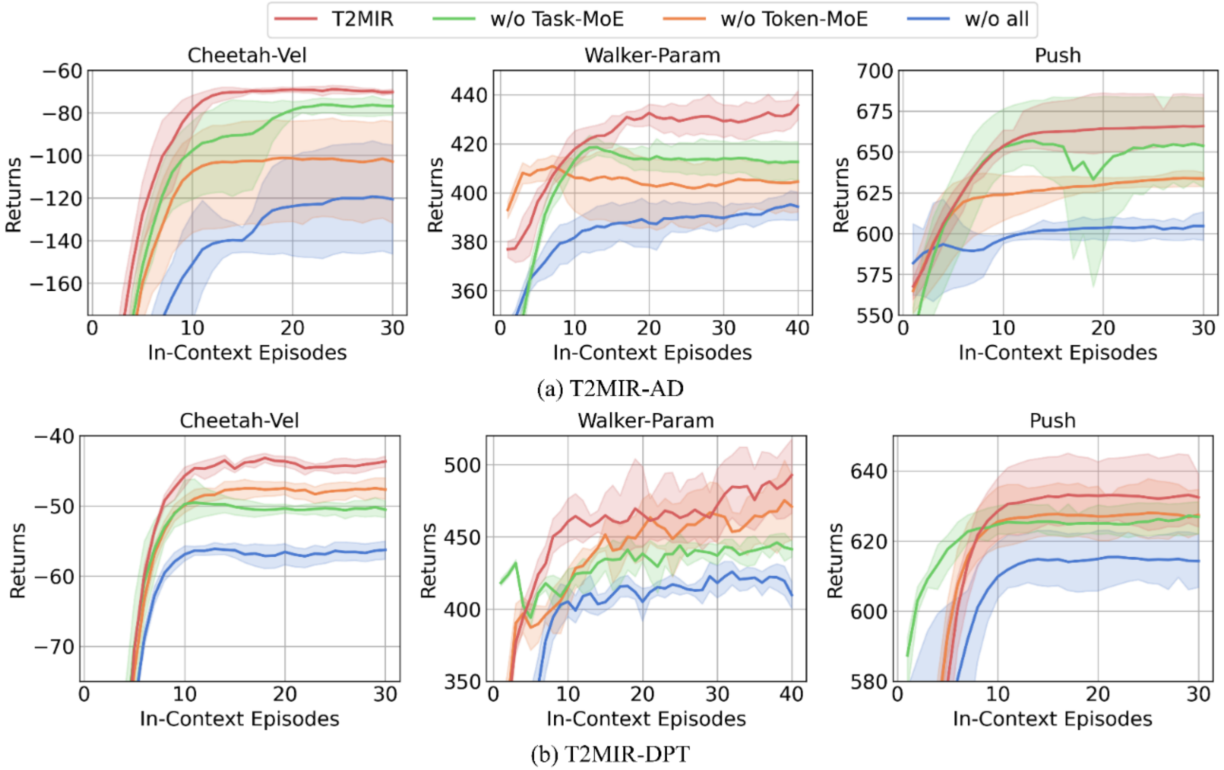


Figure 4: Ablation results of both T2MIR-AD and T2MIR-DPT architectures using Mixed datasets

3. Analysis

NANJING UNIVERSITY

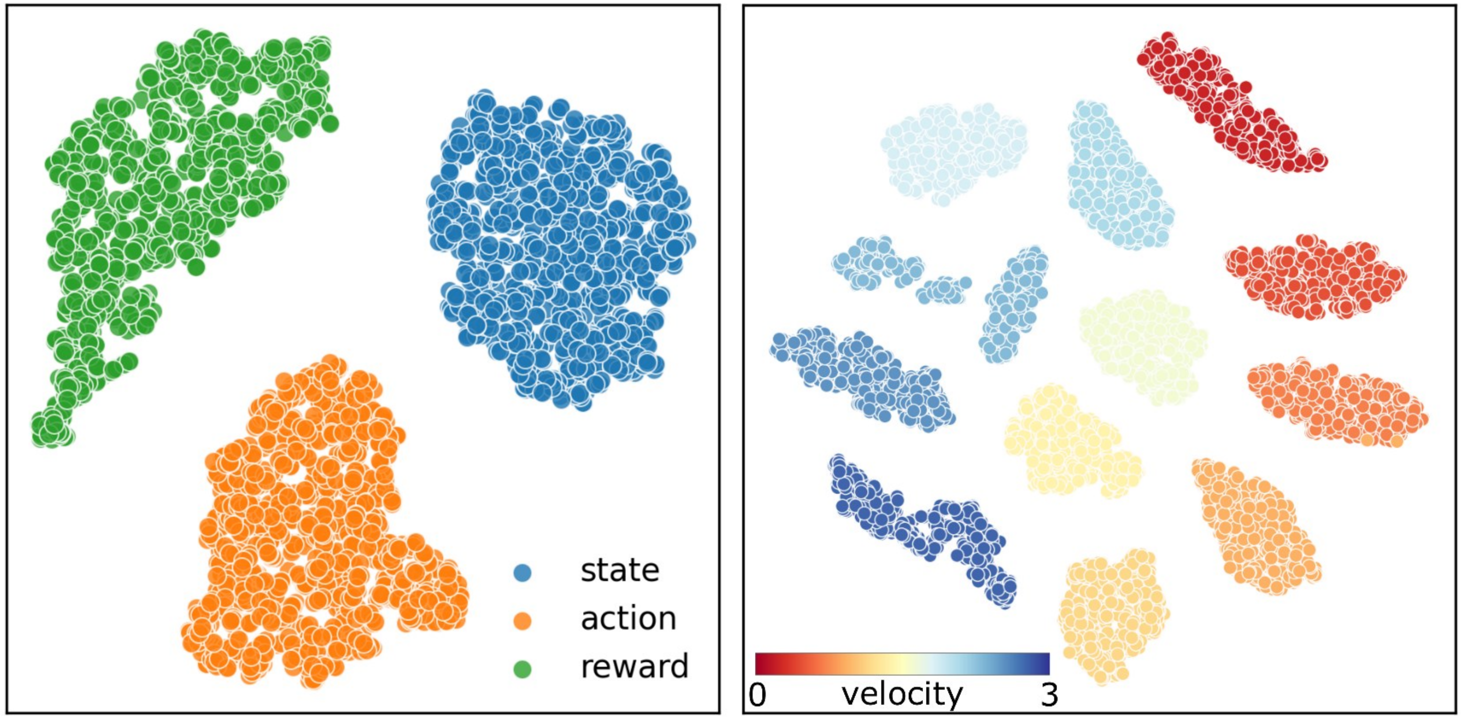


Figure 1: t-SNE visualization of expert assignments on Cheetah-Vel where tasks differ in target velocities. Left: token-wise MoE enables different experts to process tokens with distinct semantics. Right: task-wise MoE effectively manages a broad task distribution, where the difference between expert assignments is positively related to the difference between tasks.

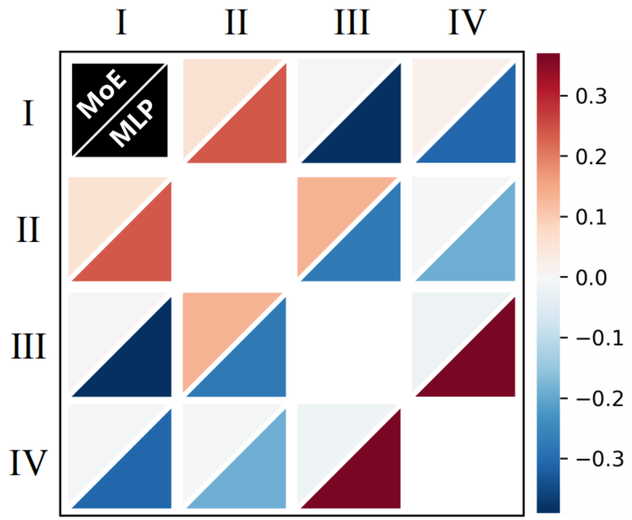


Figure 8: Cosine similarity of gradients between Point-Robot tasks in four quadrants (I-IV), comparing T2MIR-AD (MoE) with AD (MLP).

4. Conclusions

NANJING UNIVERSITY

- **Innovative Framework**: This study introduces T2MIR, a novel framework that integrates the mixture-of-experts (MoE) architecture into transformer-based decision models for in-context reinforcement learning (ICRL).

- **Key Contributions**: The proposed Token-wise MoE effectively handles multi-modal inputs by capturing distinct semantics of input tokens. The Task-wise MoE manages a broad task distribution and reduces gradient conflicts through specialized experts and contrastive learning-enhanced task routing.

- **Significant Advantages**: T2MIR significantly boosts in-context learning capacity. It demonstrates superior performance over various baselines across multiple benchmarks, proving its effectiveness in advancing ICRL.

- **Future Prospects**: An urgent improvement is to evaluate on more complex environments such as XLand-MiniGrid with huge datasets, unlocking the scaling properties of MoE in ICRL domains. Another step is to deploy our method to vision-language-action (VLA) tasks that naturally involve more complex input multi-modality and task diversity.

Thank you!

Wenhao Wu (吴文浩)

Department of Control Science and Intelligent Engineering
Nanjing University

Email: wenhaowu@smail.nju.edu.cn

Code: https://github.com/NJU-RL/T2MIR

Paper: https://arxiv.org/abs/2506.05426