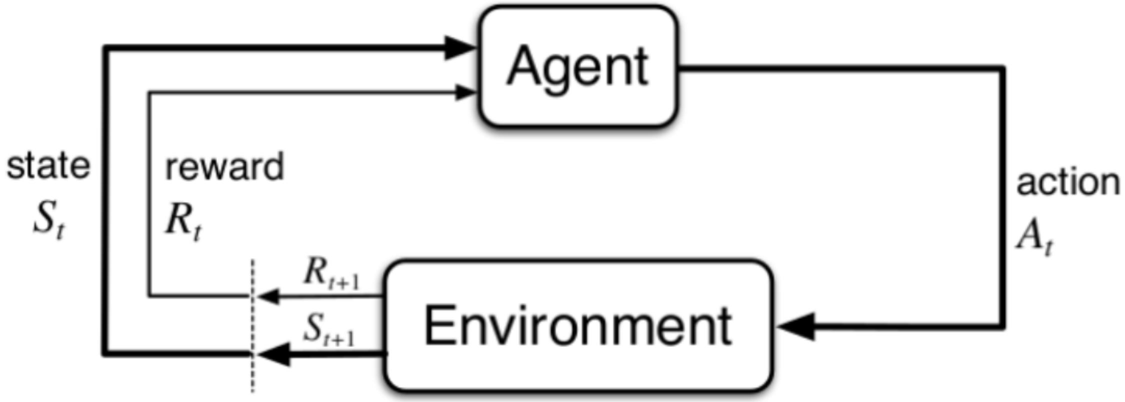
# Select before Act: Spatially Decoupled Action Repetition for Continuous Control

Buqing Nie[1], Yangqing Fu[1], Yue Gao[1,2]

[1] Shanghai Jiao Tong University [2] Shanghai Innovation Institute

# Motivation

- DRL methods achieve remarkable success in **continuous control tasks**
- DRL methods make decisions at **individual time steps** [1,2]
- No **temporal consistency** of action sequences
  - *inefficient* exploration [3]
  - *challenging* credit assignment, *poor* sample efficiency [4,5]

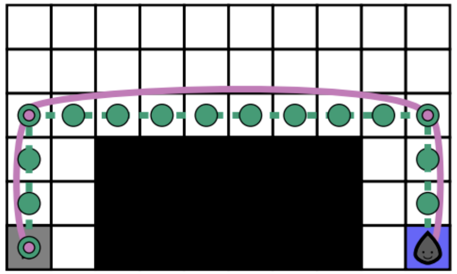



DRL decision process

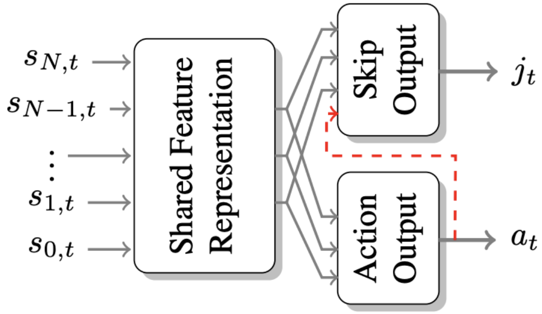

[1] Silver, David, et al. "Deterministic policy gradient algorithms.", ICML 2014
[2] Schulman, John, et al. "Proximal policy optimization algorithms.", ArXiv 2017.
[3] Dabney, Will, et al. "Temporally-extended $\epsilon$-greedy exploration." ICLR 2021.
[4] Biedenkapp, André, et al. "Temporl: Learning when to act.", ICML 2021.
[5] Haichao Zhang, et al. "Generative planning for temporally coordinated exploration in reinforcement learning", ICLR 2022.

# Method

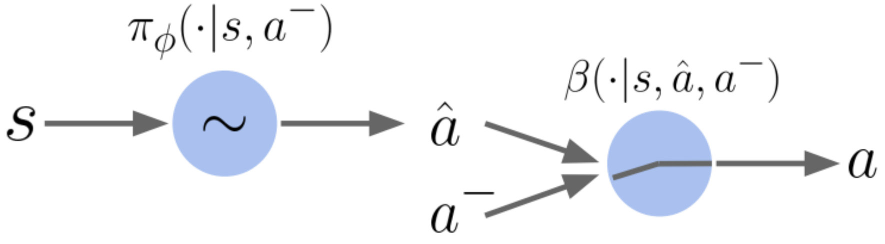
- Action Repetition-based RL
  - *Open-loop* methods: DAR[6] , FiGAR[7] , TempoRL[4] , UTE [8].
  - *Closed-loop* methods: PIC [9] , TAAC [10]
- **Higher action persistence**
  - Deeper exploration, higher learning efficiency



Action Repetition




TempoRL[4]




TAAC[5]

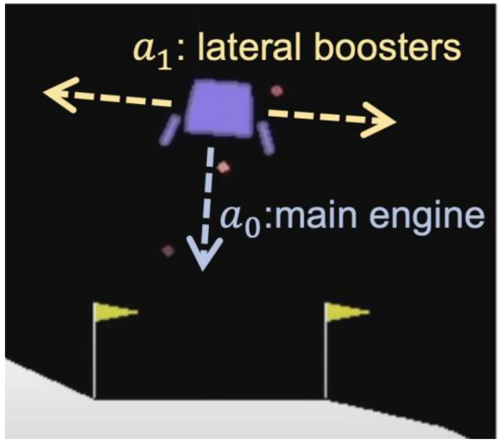

[6] Aravind Lakshminarayanan, et al. Dynamic action repetition for deep reinforcement learning. AAAI 2017
[7] Sahil Sharma, et al. Learning to repeat: Fine grained action repetition for deep reinforcement learning. ICLR 2017
[8] Joongkyu Lee, et al. Learning uncertainty-aware temporally-extended actions. AAAI 2024
[9] Chen Chen, et al. Addressing action oscillations through learning policy inertia. AAAI 2021
[10] Yu, Haonan, et al. "Taac: Temporally abstract actor-critic for continuous control." NeurIPS 2021.

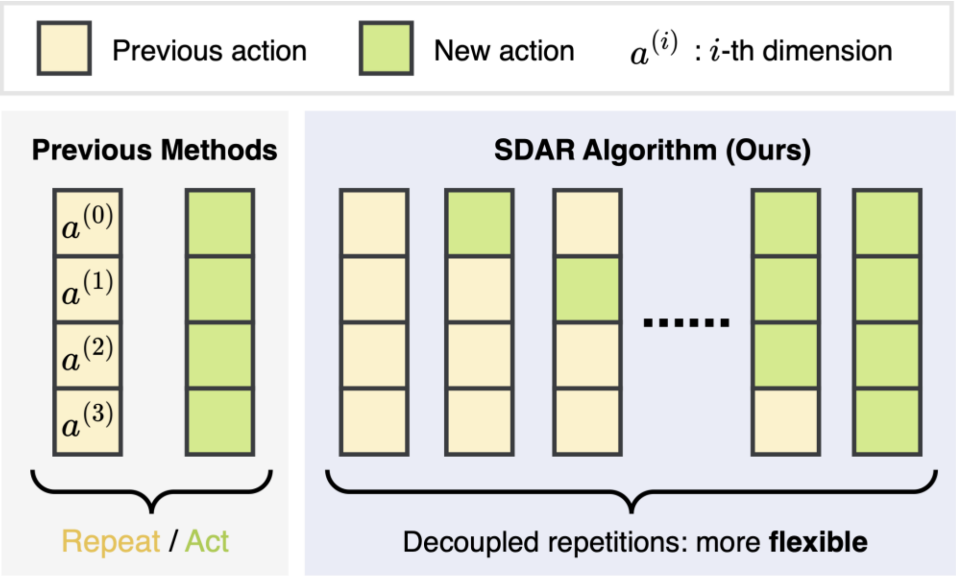
# Method: Spatially Decoupled Design

- Previous methods: **two choices** in each step
  - Repeat **all** engine actions
  - Act: make new decisions for **all** engines

- **Spatial features** are ignored
  - *Inflexible* repetition strategies
  - Reduce action *diversity*

- This work SDAR :
  - **Spatially decoupled** repetition strategy
  - Improve action **persistence & diversity**

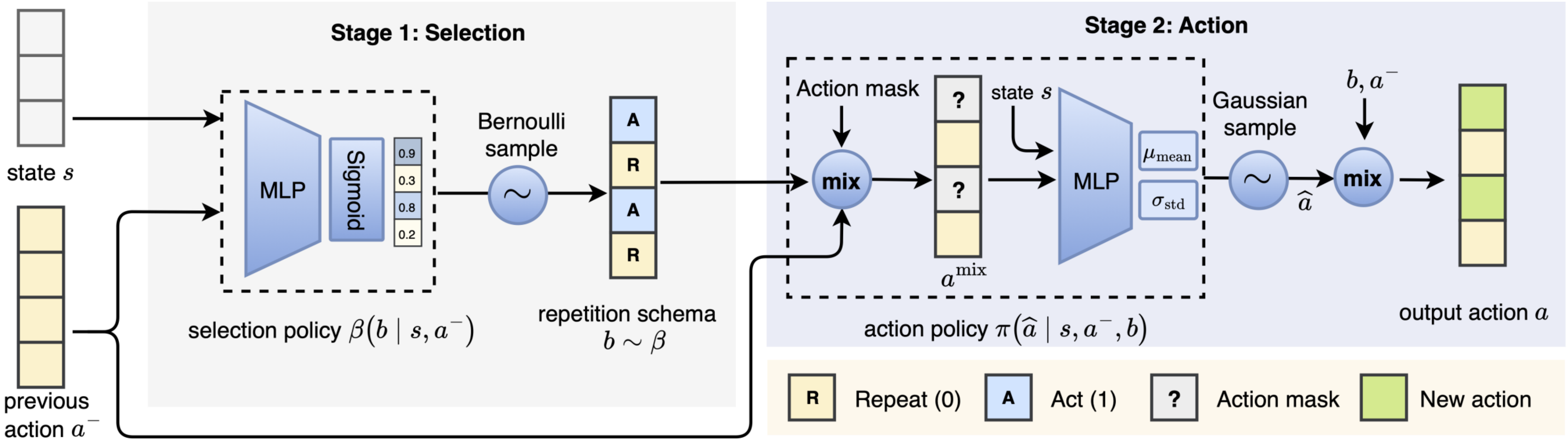


# Method: Two-stage Policy

**(1) Selection:** *act-or-repeat* decision for each action dimension

**(2) Action:** generate new actions

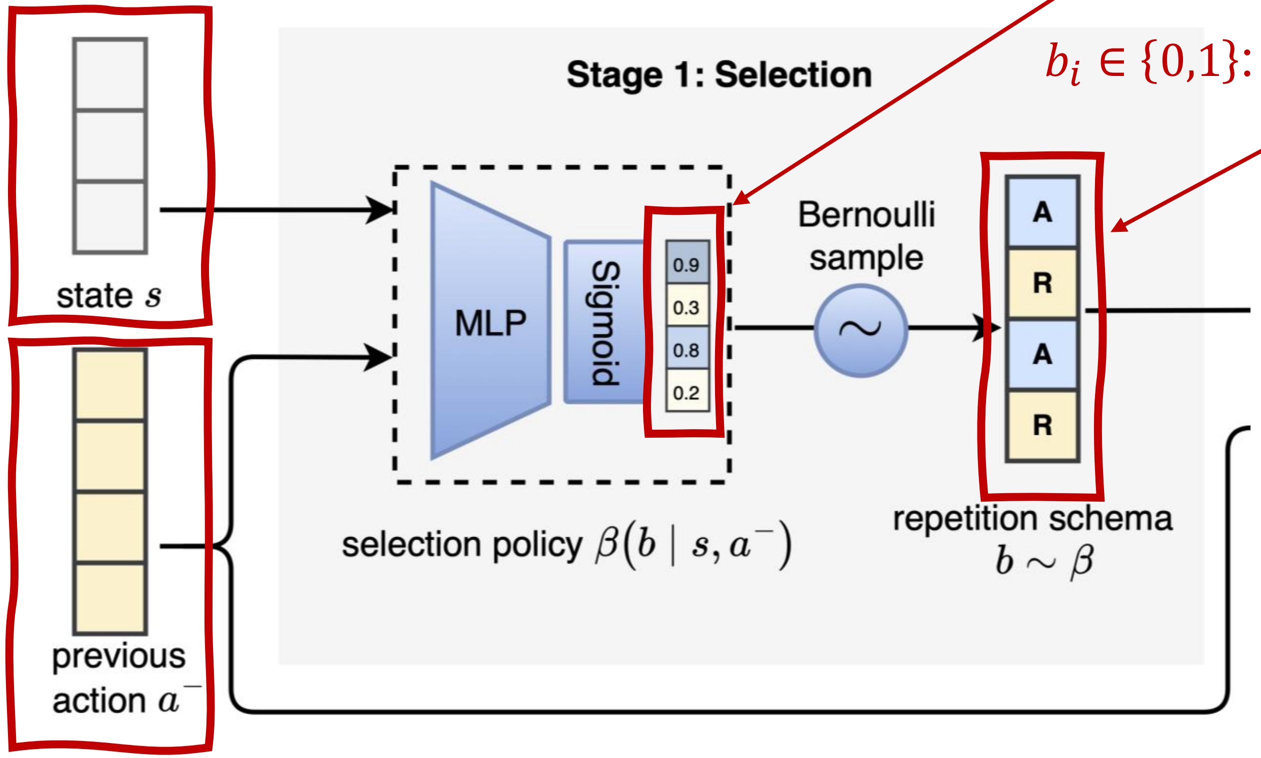


# Method: Two-stage Policy

- Selection policy $\beta(b|s,a^-) \in [0,1]^{|A|}$

$\beta_i \in [0,1]$: probability of choosing *act*

$b_i \in \{0,1\}$: act or repeat of $i$-th dim.



# Method: Two-stage Policy

- Action policy $\pi(\hat{a}|s, a^{-}, b) \in A$ generate new actions

- Make masked action:

$$a^{\text{mix}} = \text{Mix}(b, a^{-}, \xi) = (1-b) \odot a^{-} + b \odot \xi,$$

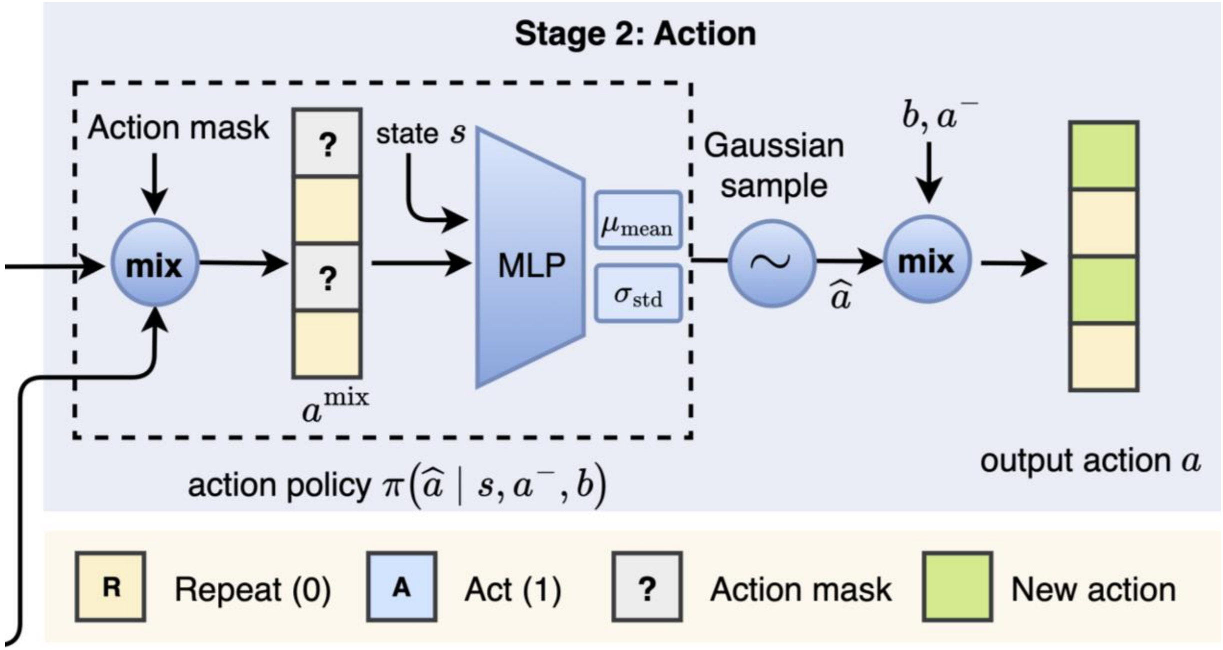
- Make new decisions $\hat{a}$:

$$\mu_{\text{mean}}, \sigma_{\text{std}} = \text{MLP}(s, a^{\text{mix}})$$

$$\hat{a} = \mu_{\text{mean}} + \sigma_{\text{std}} \cdot n, \quad n \sim \mathcal{N}(0, \mathcal{I})$$

- Synthesize new action $a$

$$a = \text{Mix}(b, a^{-}, \hat{a}) = (1-b) \odot a^{-} + b \odot \hat{a}$$



# Method: How to Train SDAR

- Policy Evaluation

$$\min \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[Q(s,a) - \mathcal{T}Q(s,a)\right]^2, \text{ with } \mathcal{T}Q(s,a) = R(s,a) + \gamma\mathbb{E}_{P,\beta,\pi}\left[Q(s',a')\right]$$

- Policy Improvement

$$J(\theta^\beta,\theta^\pi) = \underbrace{\mathbb{E}_{(s,a^-)\sim\mathcal{D}}\mathbb{E}_{b\sim\beta,\hat{a}\sim\pi}}_{\text{make decisions on samples}}\Bigg[\underbrace{Q(s,a)}_{\max Q \text{ values}}\underbrace{-\alpha_\beta\log\beta\left(b|s,a^-\right) - \alpha_\pi\log\pi\left(\hat{a}|s,a^-,b\right)}_{\text{entropy-based exploration}}\Bigg]$$

- Update Selection policy $\beta$:

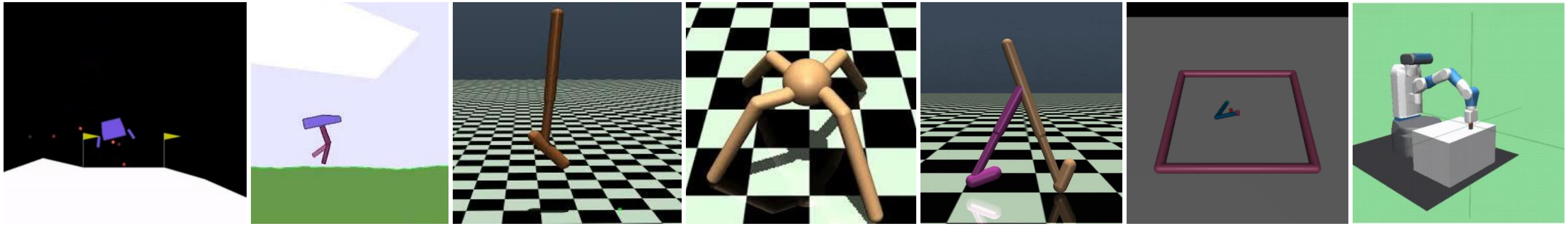
$$\max_{\theta^\beta}\mathbb{E}_{(s,a^-)\sim\mathcal{D}}\sum_{b\in\mathcal{B}}\beta(b|s,a^-)\mathbb{E}_{\hat{a}\sim\pi}\left[Q(s,a) - \alpha_\beta\log\beta\left(b|s,a^-\right) - \alpha_\pi\log\pi\left(\hat{a}|s,a^-,b\right)\right]$$

- Update $\beta$ with importance sampling:

$$\max_{\theta^\beta}\mathbb{E}_{\mathcal{D}}\mathbb{E}_{b\sim\beta_{\text{old}},\hat{a}\sim\pi}\left[Q(s,a) - \alpha_\beta\log\beta_{\text{old}}\left(b|s,a^-\right) - \alpha_\pi\log\pi\left(\hat{a}|s,a^-,b\right)\right]\cdot\frac{\beta\left(b|s,a^-\right)}{\beta_{\text{old}}\left(b|s,a^-\right)}$$

# Experiment: **Efficiency**

- Environments: Classic control / Locomotion / Manipulation



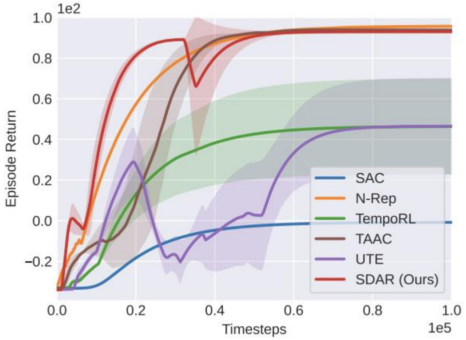
Etc.

- AUC scores:
  - SDAR achieves higher sample efficiency

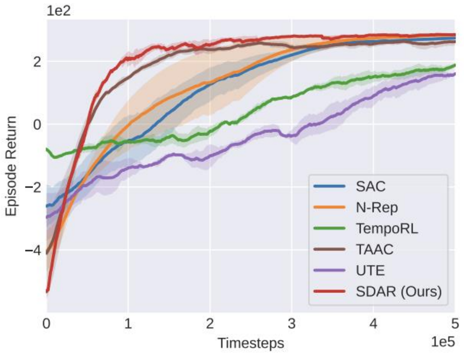
| Env. Category | Normalized AUC Score | | | | | |
|---|---|---|---|---|---|---|
| | SAC | N-Rep | TempoRL | UTE | TAAC | SDAR |
| Classic Control | 0.60±0.13 | 0.89±0.01 | 0.66±0.02 | 0.55±0.03 | 0.92±3E-3 | **1.0±0.0** |
| Locomotion | 0.78±2E-3 | 0.71±7E-3 | 0.35±0.02 | 0.43±0.01 | 0.80±0.02 | **1.0±0.0** |
| Manipulation | 0.91±4E-5 | 0.90±8E-4 | 0.77±0.02 | 0.79±0.02 | 0.95±6E-3 | **1.0±0.0** |
| **Average** | 0.76±0.02 | 0.83±0.01 | 0.59±0.05 | 0.59±0.03 | 0.90±6E-3 | **1.0±0.0** |

# Experiment: Efficiency

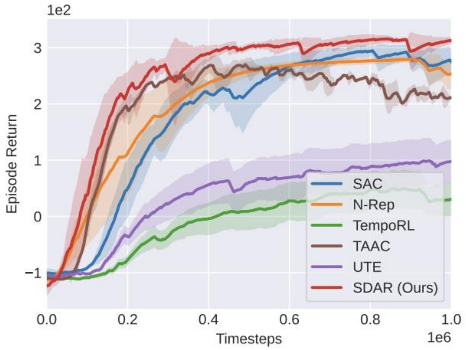
- Learning curves:
  - SDAR (red) achieves higher sample efficiency



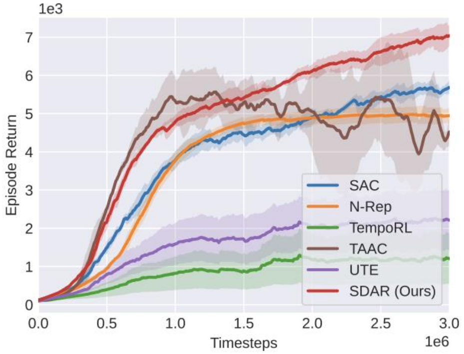

(a) *MountainCar*



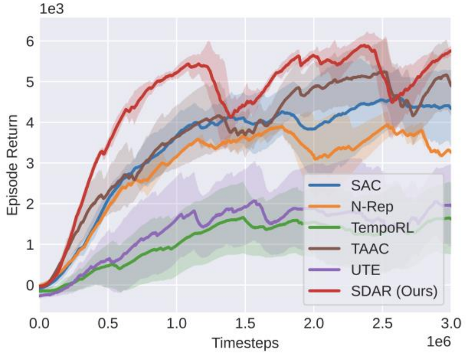

(b) *LunarLander*



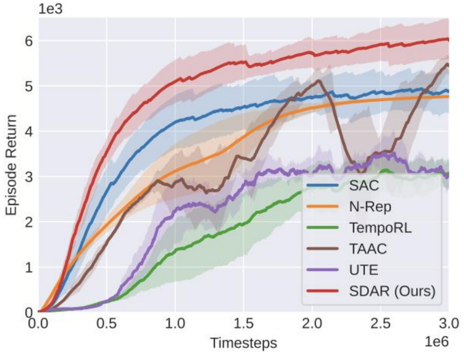

(c) *BipedalWalker*



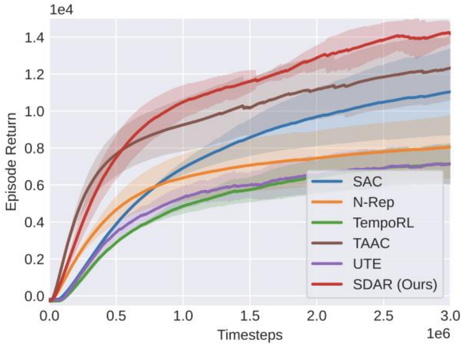

(d) *Humanoid*



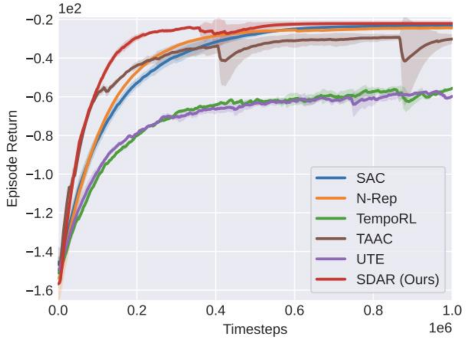

(e) *Ant*




(f) *Walker2d*




(g) *HalfCheetah*




(h) *Pusher*

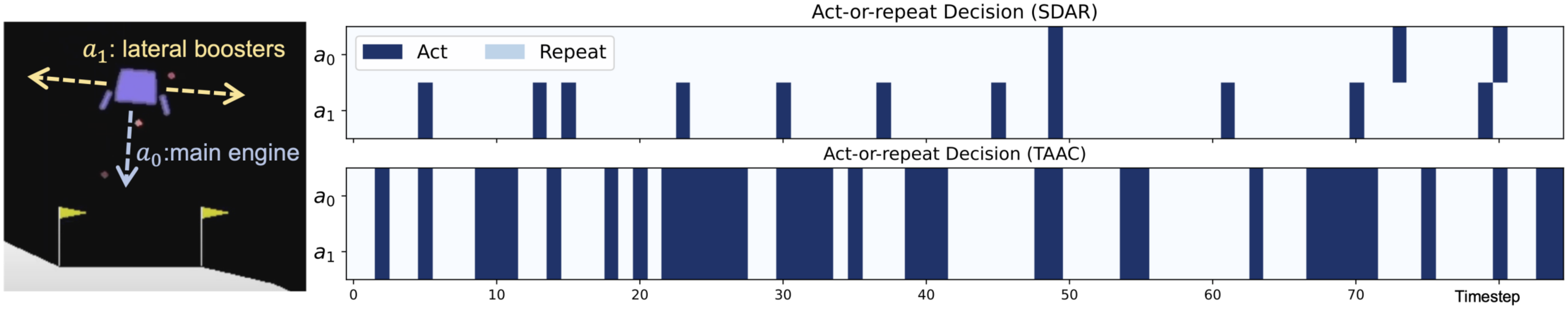
# Experiment: Persistence

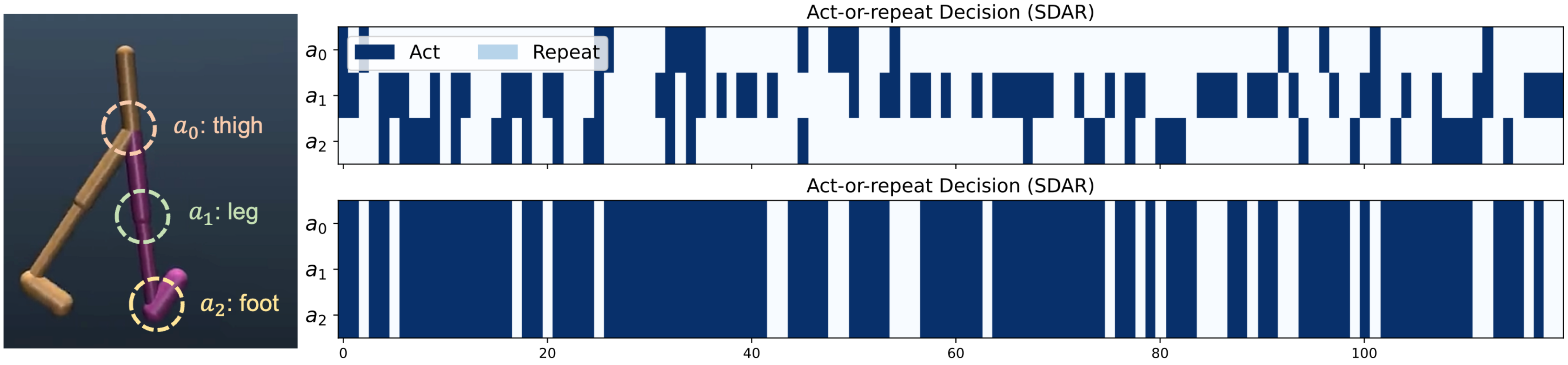
- SDAR > previous closed-loop methods > open-loop methods
- SDAR: higher return, higher persistence, lower fluctuation

| Tasks | Episode Return (Mean ± Standard Error) / Action Persistence Rate (APR) / Action Fluctuation Rate (AFR) | | | | | |
|---|---|---|---|---|---|---|
| | SAC | N-Rep | TempoRL | UTE | TAAC | SDAR (Ours) |
| LunarLander | 275.9±6.83 | 280.8±1.21 | 281.5±7.22 | **282.8±8.74** | 261.4±18.9 | 282.2±5.84 |
| | 1.00 / 0.09 | 4.00 / 0.08 | 1.43 / 0.18 | **1.38 / 0.36** | 3.05 / 0.11 | 11.18 / 0.10 |
| Walker2d | 5305±367 | 4724±163 | 2866±897 | 2986±836 | 5660±394 | **6028±406** |
| | 1.00 / 0.15 | 4.00 / 0.09 | 5.74 / 0.26 | 7.81 / 0.24 | 1.30 / 0.22 | **2.96 / 0.12** |
| HalfChee. | 13122±2877 | 8378±1753 | 8065±1799 | 7917±293 | 11148±3921 | **15131±1279** |
| | 1.00 / 0.68 | 2.00 / 0.47 | 2.10 / 0.66 | 2.69 / 0.58 | 1.02 / 0.61 | **1.22 / 0.62** |
| Humanoid | 6184±717 | 5074±310 | 1022±397 | 2595±334 | 7308±244 | **7483±288** |
| | 1.00 / 0.28 | 4.00 / 0.09 | 5.27 / 0.15 | 5.69 / 0.18 | 1.21 / 0.25 | **1.67 / 0.19** |
| Pusher | -22.5±1.40 | **-21.2±1.08** | -48.2±2.05 | -41.3±5.56 | -30.5±1.85 | -21.3±1.26 |
| | 1.00 / 0.022 | **4.00 / 0.019** | 1.15 / 0.032 | 1.01 / 0.018 | 1.75 / 0.031 | 1.69 / 0.015 |
| **Average** | 0.89±0.07 | 0.81±0.19 | 0.59±0.33 | 0.64±0.27 | 0.91±0.10 | **1.00±0.001** |
| | 1.00 / 0.245 | 3.60 / 0.150 | 3.12 / 0.257 | 3.71 / 0.276 | 1.66 / 0.244 | **3.75 / 0.208** |

# Experiment: Visualization

- SDAR is more flexible than previous method (TAAC)



# Conclusion

- Spatially decoupled action repetition
  - More *flexible* repetition strategy
  - Higher *persistence* & *diversity*
  - Higher *efficiency* and final *performance*
- First work to consider **spatial features** in temporal abstraction / action repetition
- Future works:
  - How to learn synergy more efficiently
  - How to share/transfer $\beta$ policy in different tasks
  - Incorporate into other RL methods

# Thank You !

Buqing Nie[1], Yangqing Fu[1], Yue Gao[1,2]

[1] Shanghai Jiao Tong University [2] Shanghai Innovation Institute