SMoSE: Sparse Mixture of Shallow Experts
for Interpretable Reinforcement Learning in Continuous Control Tasks

Mátyás Vincze[1,2] Laura Ferrarotti[2] Leonardo Lucio Custode[1]
Bruno Lepri[2] Giovanni Iacca[1]

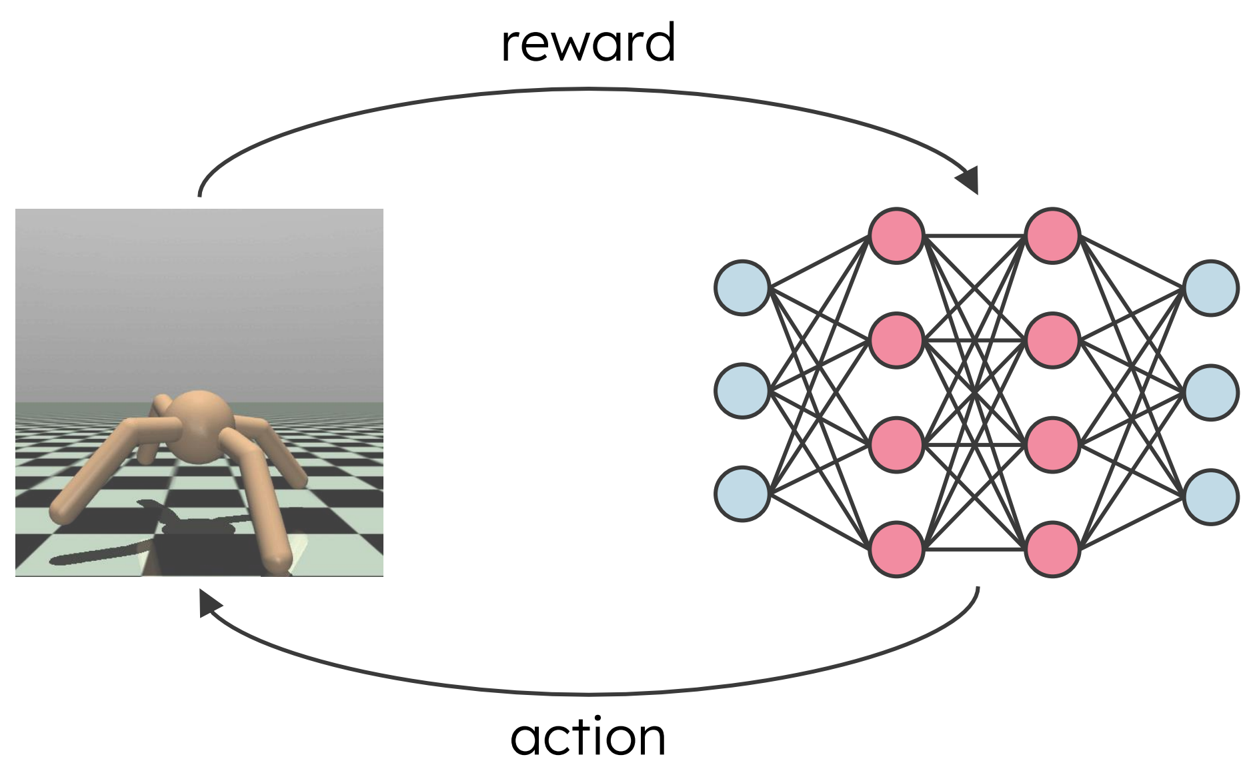
[1] UNIVERSITAS STUDIORUM ATHESINA
[2] FONDAZIONE BRUNO KESSLER

AAAI-25 / IAAI-25 / EAAI-25

Motivation

Unlock safe and efficient RL

SOTA approaches are not interpretable



- Scaling limits interpretability
- Closed-box models only allow explainability

Interpretable approaches do not work in continuous control

- Evolutionary solutions are sample-inefficient (10x environment interactions)
- Huge performance gap compared to SOTA

Paper

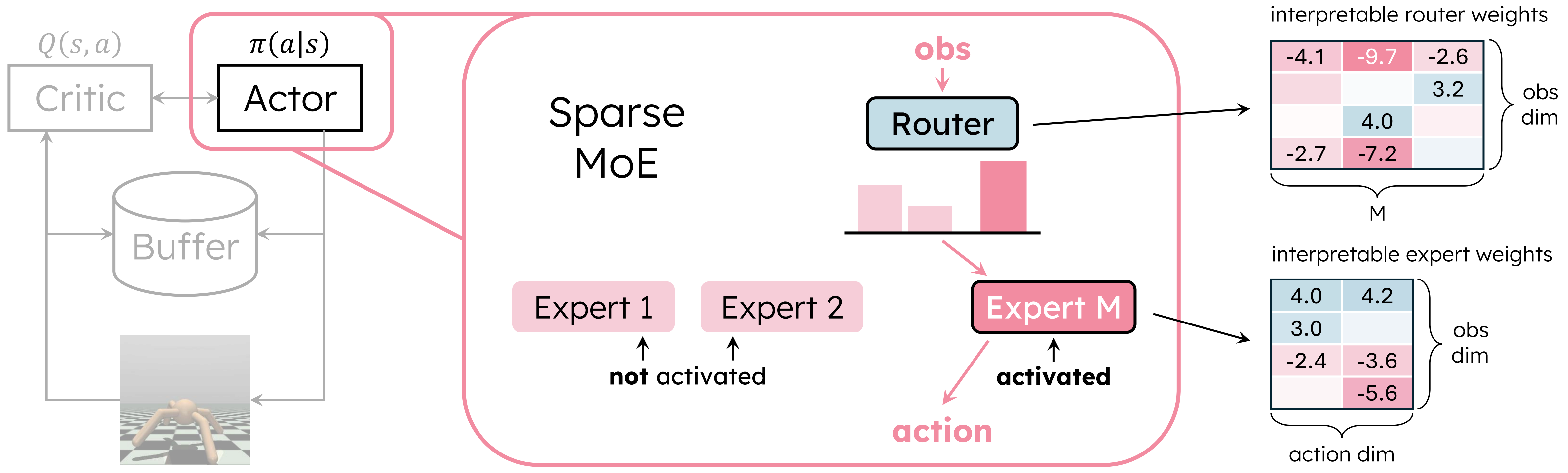
@vinczematyas_ matyas-vincze

Method

Sparse MoE actor, Linear experts, Post-training distillation

Architecture : Linear router, linear experts

Router partitions the state space while the experts specialize on simple skills



Training stabilization

- Load balancing with auxiliary loss

$$\mathrm{L_{aux}} = 0.1 * \left[\begin{array}{c} f_{\mathrm{imp}}(\mathrm{S}) = \frac{1}{2}\left(\frac{\mathrm{std}(\mathrm{Imp}(\mathrm{S}))}{\mathrm{mean}(\mathrm{Imp}(\mathrm{S})}\right)^2 \\ + \\ f_{\mathrm{load}}(\mathrm{S}) = \frac{1}{2}\left(\frac{\mathrm{std}(\mathrm{Load}(\mathrm{S}))}{\mathrm{mean}(\mathrm{Load}(\mathrm{S})}\right)^2 \end{array} \right]$$

- Forced expert-space exploration

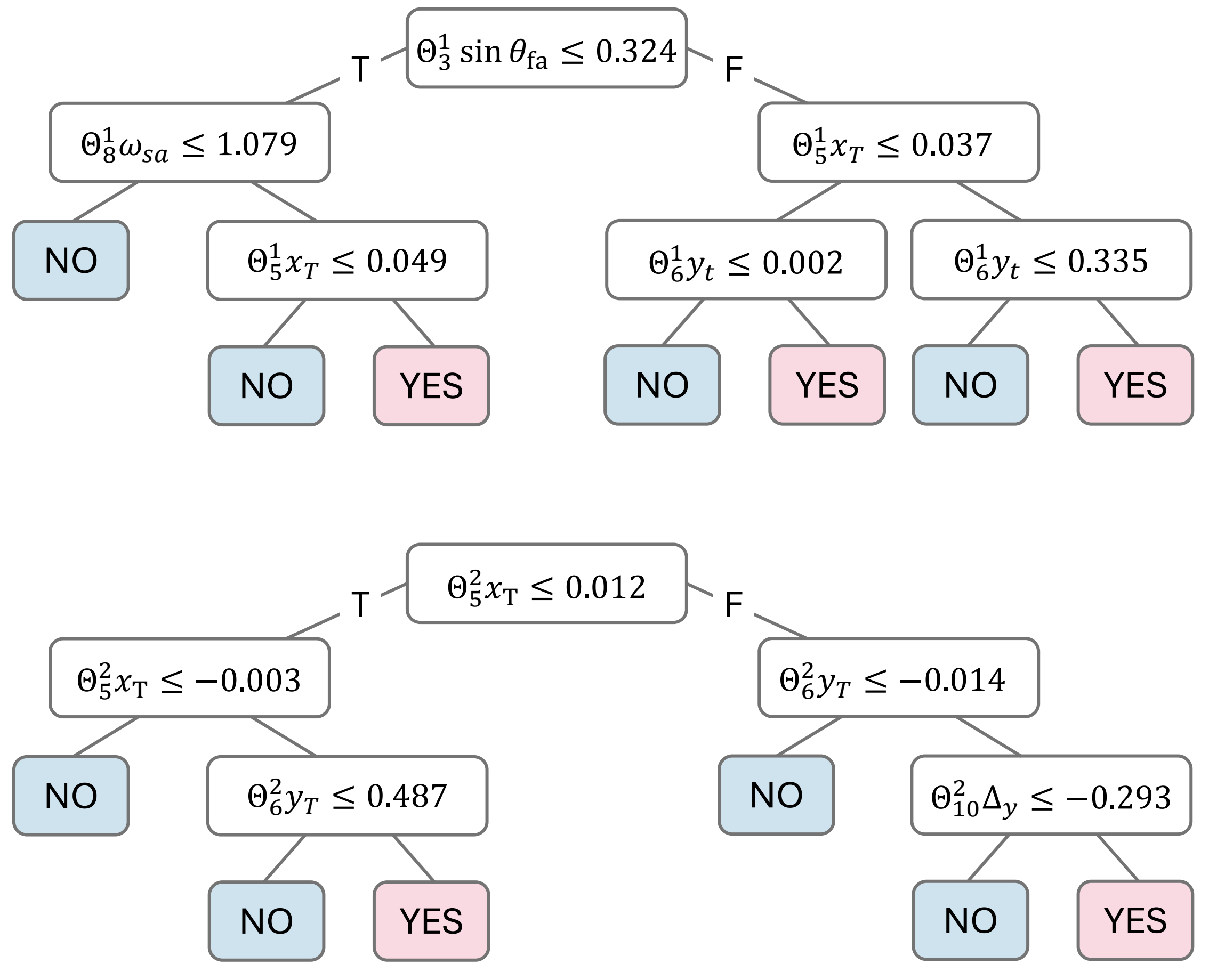
$$\varepsilon \sim \mathcal{N}(0, 1/M^2)$$

$$\mathrm{Load}_m(\mathrm{S}) = \sum_{s_k \in \mathrm{S}} \mathbb{P}(\varepsilon_{new} \geq \tau(s_k) - \pi_m(s_k))$$

$$\mathrm{Imp}_m(\mathrm{S}) = \sum_{s_k \in \mathrm{S}} \mathrm{softmax}(\pi_m(s_k | \theta_m, \sigma_m))$$

Router distillation

- Per-expert binary decision tree for “free”

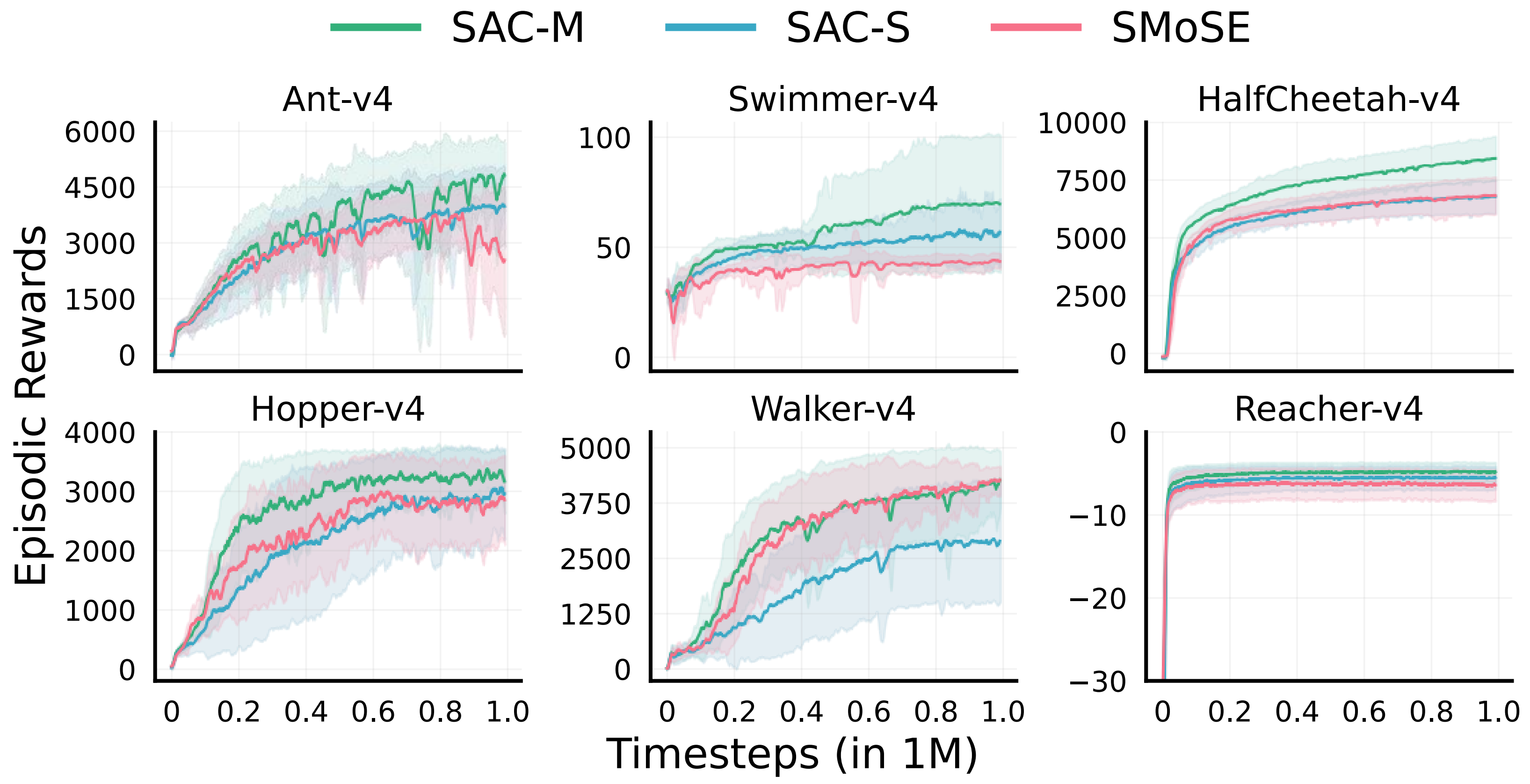

$\Theta_3^1 \sin\theta_{\mathrm{fa}} \leq 0.324$
T
F
$\Theta_8^1 \omega_{sa} \leq 1.079$
$\Theta_5^1 x_T \leq 0.037$
NO
$\Theta_5^1 x_T \leq 0.049$
$\Theta_6^1 y_t \leq 0.002$
$\Theta_6^1 y_t \leq 0.335$
NO
YES
NO
YES
NO
YES
$\Theta_5^2 x_{\mathrm{T}} \leq 0.012$
T
F
$\Theta_5^2 x_{\mathrm{T}} \leq -0.003$
$\Theta_6^2 y_T \leq -0.014$
NO
$\Theta_6^2 y_T \leq 0.487$
NO
$\Theta_{10}^2 \Delta_y \leq -0.293$
NO
YES
NO
YES

Results

Strong performance on Mujoco tasks

- **Fully interpretable**
- **2x performance** compared to SOTA interpretable approaches
- **99% less parameters** than SAC-L
- **Closes the gap** with close-box solutions

Training speed



Episodic Reward on Mujoco

	Walker2d	Hopper	Ant	HalfCheetah	Reacher	Swimmer
SAC-L	**4358.06**	2636.49	**5255.46**	**11809.87**	**-3.75**	68.59
SAC-M	4020.51	**3224.25**	4894.18	8992.22	-4.02	71.94
SAC-S	2967.14	3076.09	4162.97	7214.3	-4.82	59.42
PPO	3362.16	2311.9	2327.12	2308.29	-6.57	93.26
CGP	1090.00	1150.00	1130.00	6375.00	-68.50	**280.00**
LGP	1080.00	1120.00	1210.00	6388.50	-58.50	278.50
Metric-40	775.00	2005.00	2210.50	2210.50	x	x
Ours	**4224.29**	**2816.08**	**3245.43**	**7310.17**	**-5.49**	**45.4**