M^3PC: Test-Time Model Predictive Control using Pretrained Masked Trajectory Model

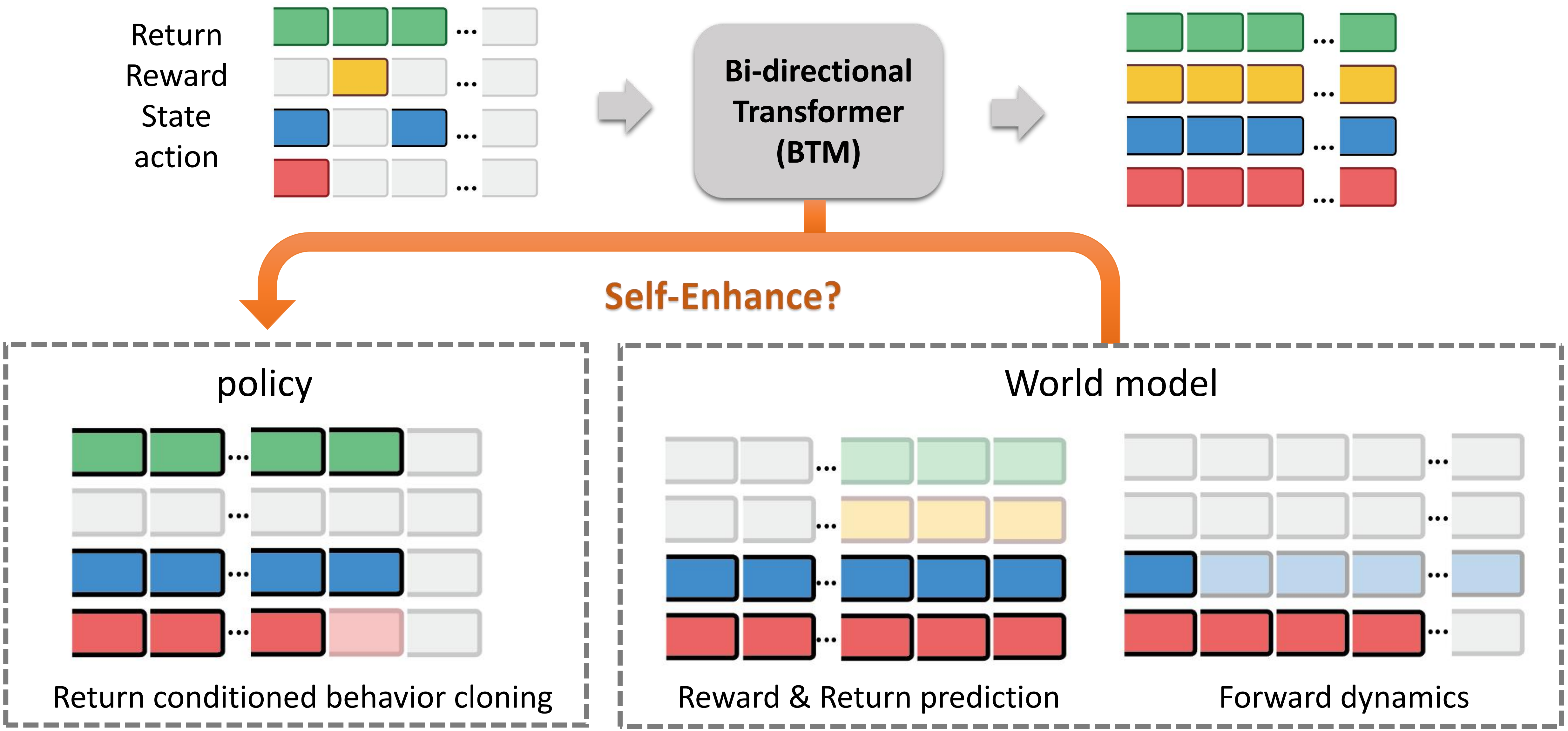
Presenter： Yao Mu

Masked Pre-trained Models

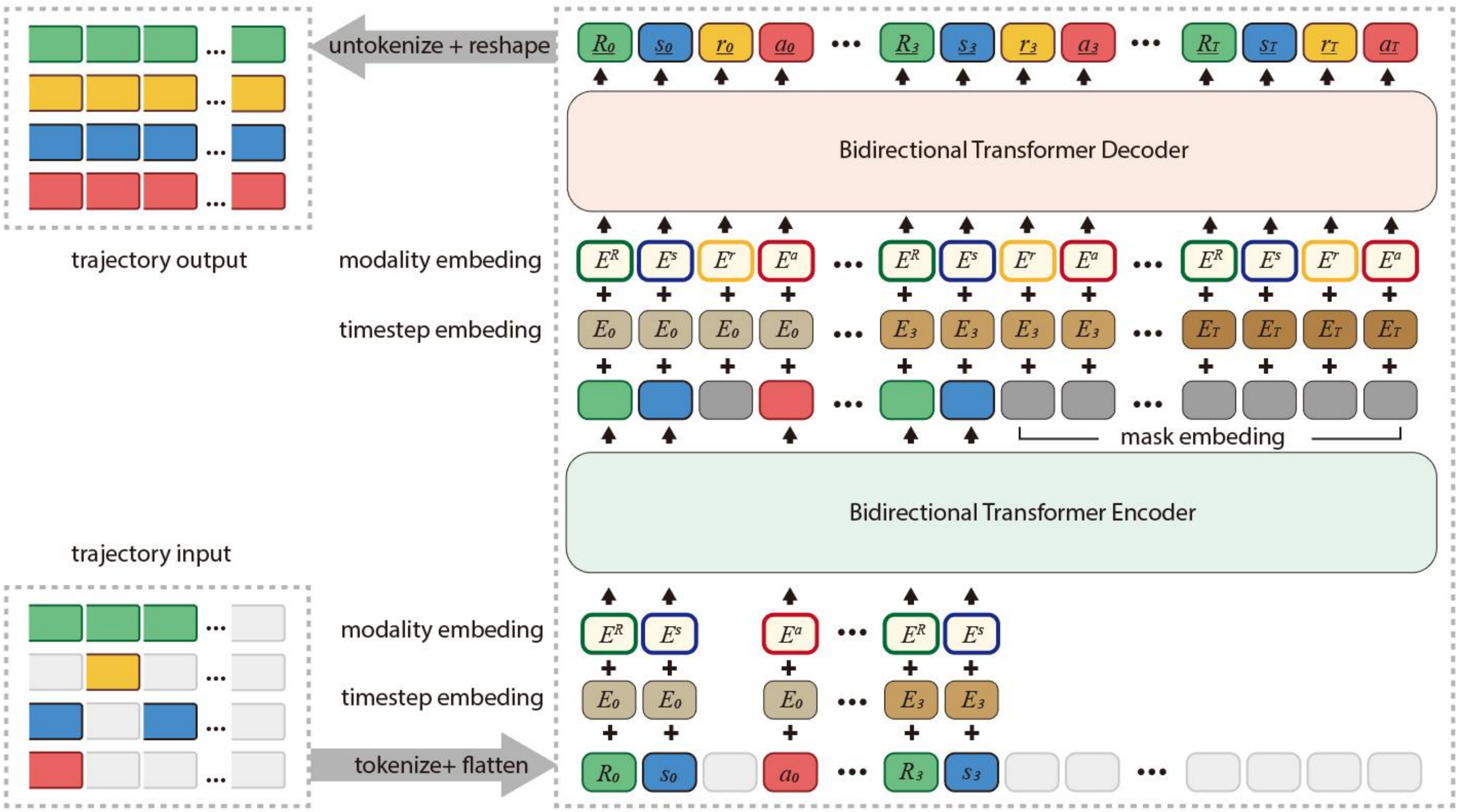
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *CVPR*. 2022.

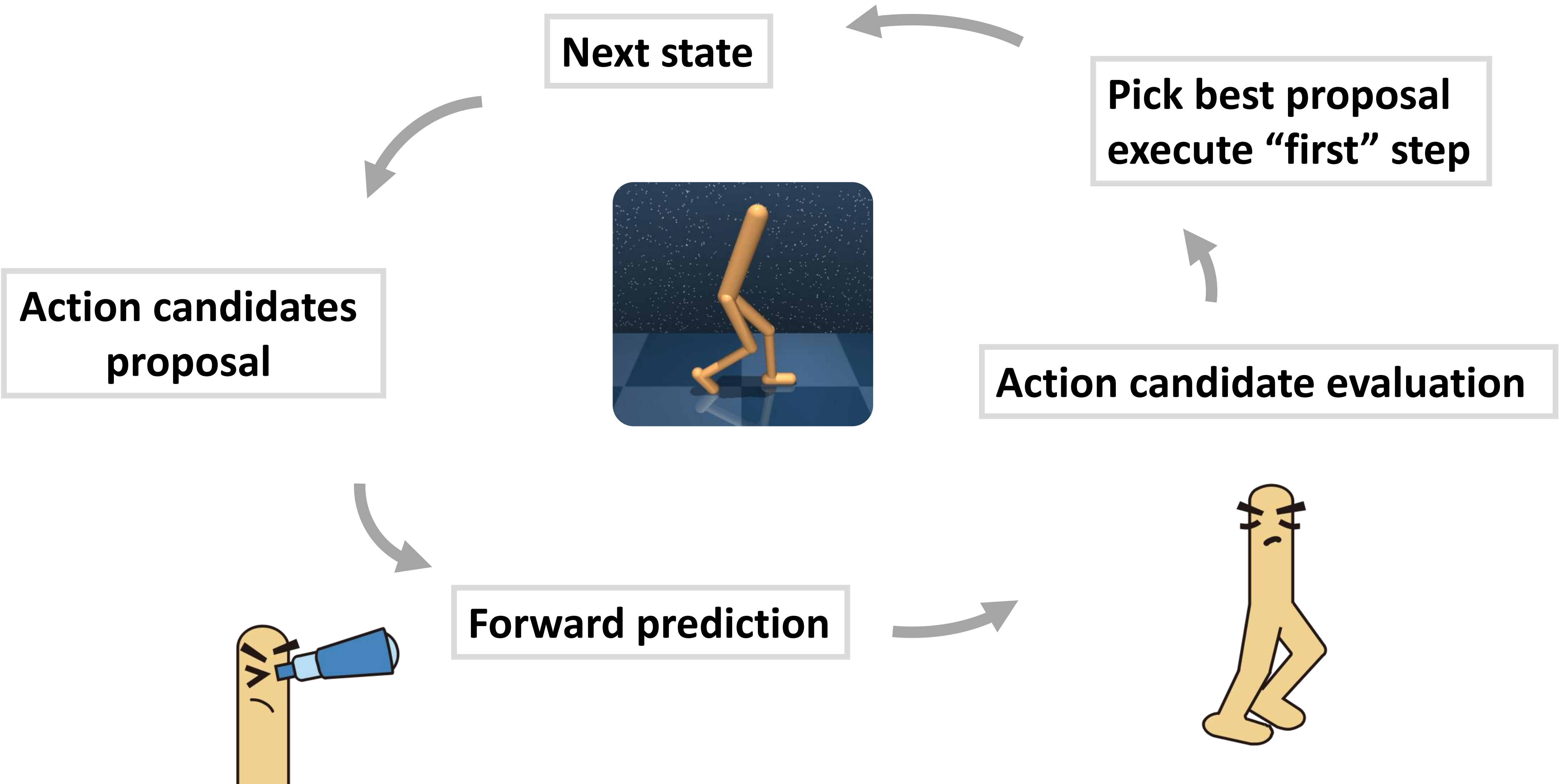
Masked Pre-trained Models for RL



M³PC: Test-time Model Predictive Control using **Pretrained Masked Trajectory Model**

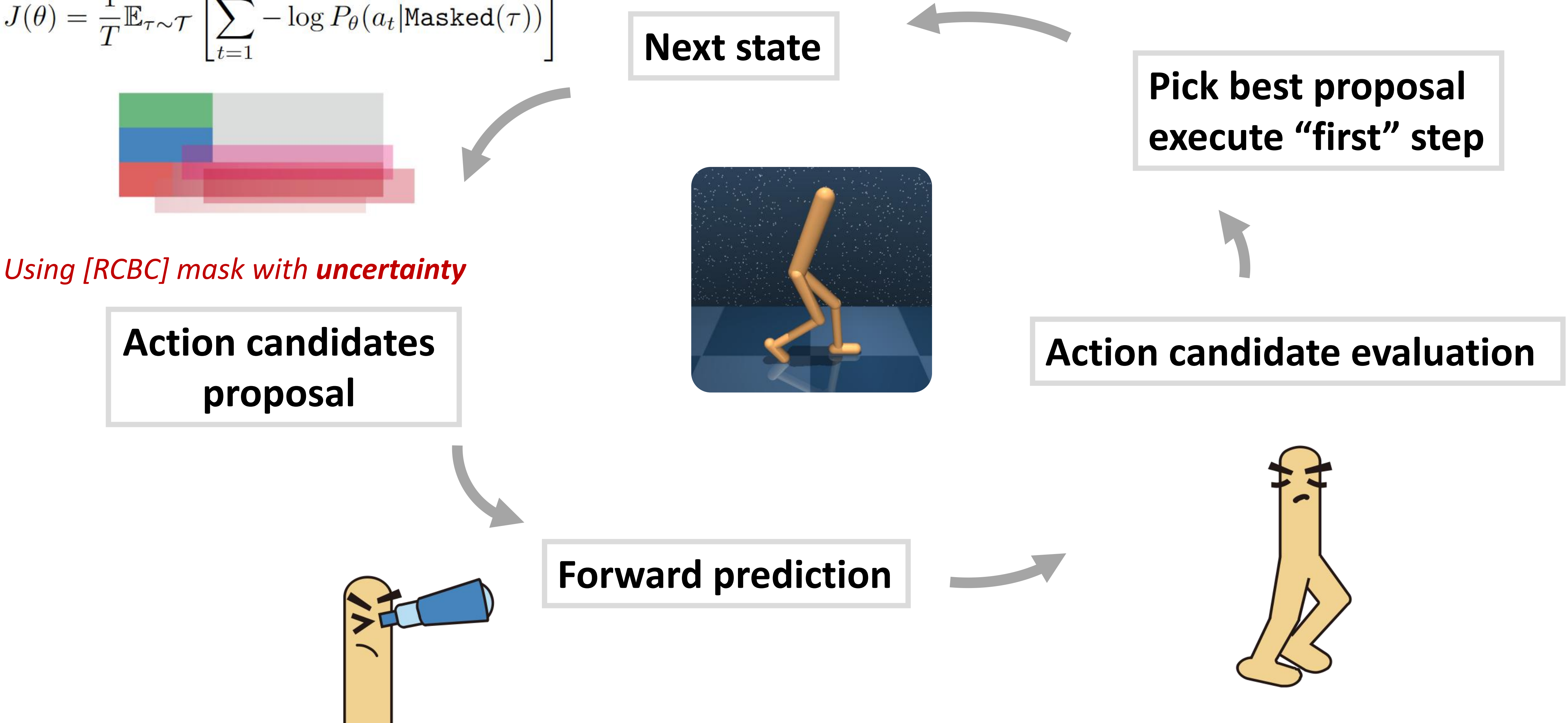


M^{3}PC: Test-time **Model Predictive Control** using Pretrained Masked Trajectory Model

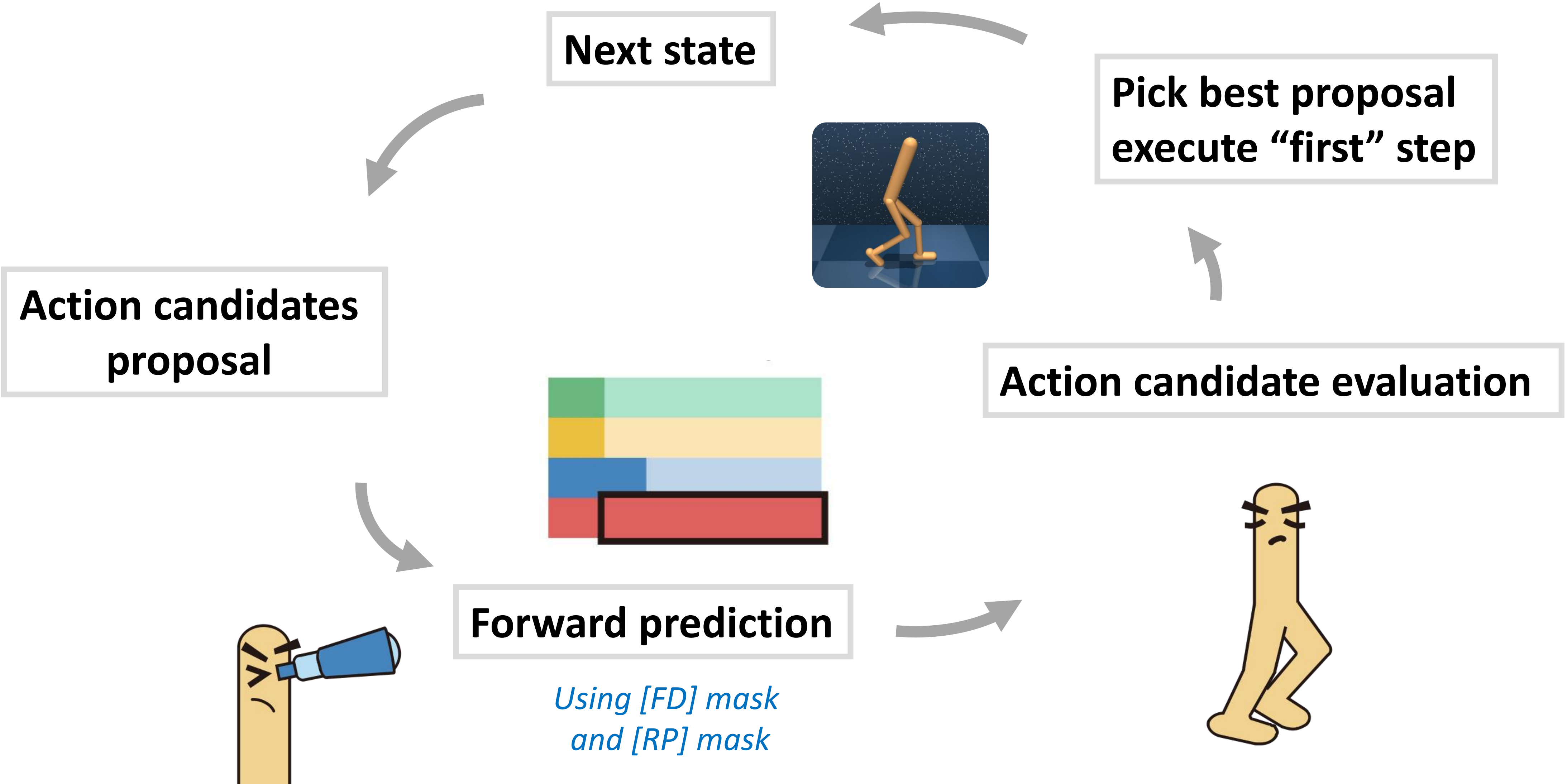


M³PC: Test-time **Model Predictive Control** using Pretrained Masked Trajectory Model

$$J(\theta) = \frac{1}{T}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\sum_{t=1}^{T} -\log P_\theta(a_t|\texttt{Masked}(\tau))\right]$$



M3PC: Test-time **Model Predictive Control** using Pretrained Masked Trajectory Model



M3PC: Test-time **Model Predictive Control** using Pretrained Masked Trajectory Model

Next state

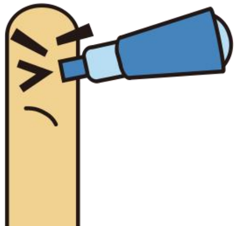
Pick best proposal execute "first" step

Action candidates proposal

Action candidate evaluation

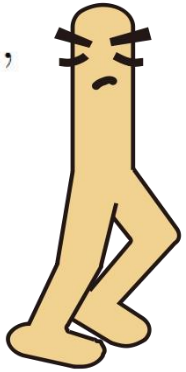
*Based on **Utility** calculated from TD(λ) return estimation*

Forward prediction



$$U = (1-\lambda)\sum_{n=0}^{T-t-1} \lambda^n G_{t:t+n} + \lambda^{T-t} G_{t:T},$$

$$\text{where } G_{t:t+n} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n g_{t+n}$$



forward M^3PC for offline RL

Dataset	BC	TD3+BC	IQL	DT	TT	BTM	M^3PC-M	M^3PC-Q
hopper-m	53.5	60.4	63.8	65.1	61.1	64.3	$70.7_{\pm 6.2}$	$\mathbf{73.6}_{\pm 5.6}$
walker2d-m	63.2	82.7	79.9	67.6	79.0	72.5	$80.9_{\pm 2.5}$	$\mathbf{86.4}_{\pm 2.6}$
halfcheetah-m	42.4	48.1	47.4	42.2	46.9	43.0	$43.9_{\pm 3.9}$	$\mathbf{51.2}_{\pm 0.7}$
hopper-m-r	29.8	64.4	**92.1**	81.8	91.5	75.3	$80.4_{\pm 5.2}$	$78.3_{\pm 16.2}$
walker2d-m-r	21.8	85.6	73.7	82.1	82.6	76.6	$78.2_{\pm 10.2}$	$\mathbf{92.2}_{\pm 2.4}$
halfcheetah-m-r	35.7	44.8	44.1	**48.3**	41.9	41.1	$41.8_{\pm 0.5}$	$48.2_{\pm 0.4}$
Total	246.4	386.0	401.0	387.1	403.0	372.8	395.9	**429.8**

15.3% higher performance score
without any network weight change

forward M^3PC for online finetuning

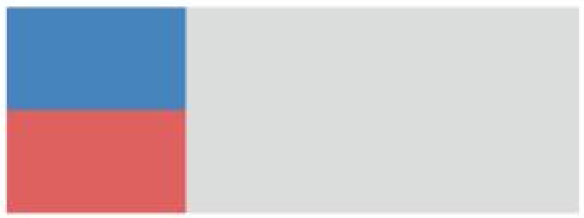
Dataset	IQL			ODT			M^3PC (Ours)		
	offline	online	δ	offline	online	δ	offline	online	δ
hopper-m	63.8	66.8	+3.0	67.0	**97.5**	+30.6	$73.6_{\pm 5.6}$	$93.9_{\pm 15.8}$	+20.3
walker2d-m	79.9	80.3	+0.4	72.2	76.8	+4.6	$86.4_{\pm 2.6}$	$\mathbf{91.9}_{\pm 7.8}$	+5.5
halfcheetah-m	47.4	47.4	+0.0	42.7	42.2	-0.6	$51.2_{\pm 0.7}$	$\mathbf{69.3}_{\pm 2.1}$	+18.1
hopper-m-r	92.1	96.2	+4.1	86.6	88.9	+2.3	$78.3_{\pm 16.2}$	$\mathbf{103.5}_{\pm 6.0}$	+25.2
walker2d-m-r	73.7	70.6	-3.1	68.9	76.9	+7.9	$92.2_{\pm 2.4}$	$\mathbf{105.2}_{\pm 1.0}$	+13.0
halfcheetah-m-r	44.1	44.1	+0.0	40.0	40.4	+0.4	$48.2_{\pm 0.4}$	$\mathbf{67.0}_{\pm 7.1}$	+18.8
Total	401.0	405.5	+4.5	377.4	422.7	+45.3	429.8	**530.8**	+101.0

31% higher final performance score than Online Decision Transformer

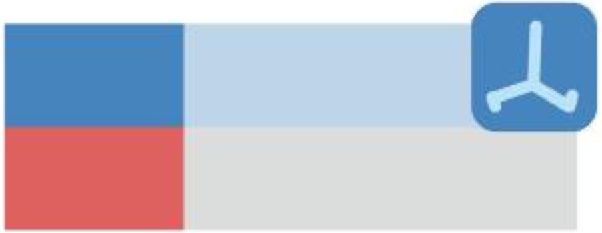
123% more substantial improvements than Online Decision Transformer

Zheng, Qinqing, Amy Zhang, and Aditya Grover. "Online decision transformer." *international conference on machine learning*. PMLR, 2022.

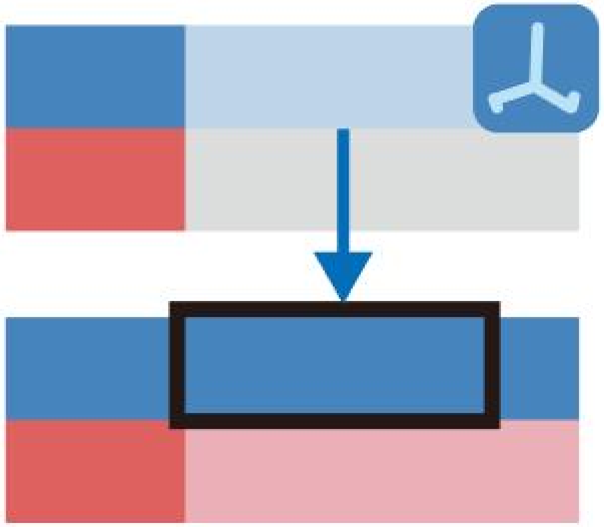
Backward M^{3}PC



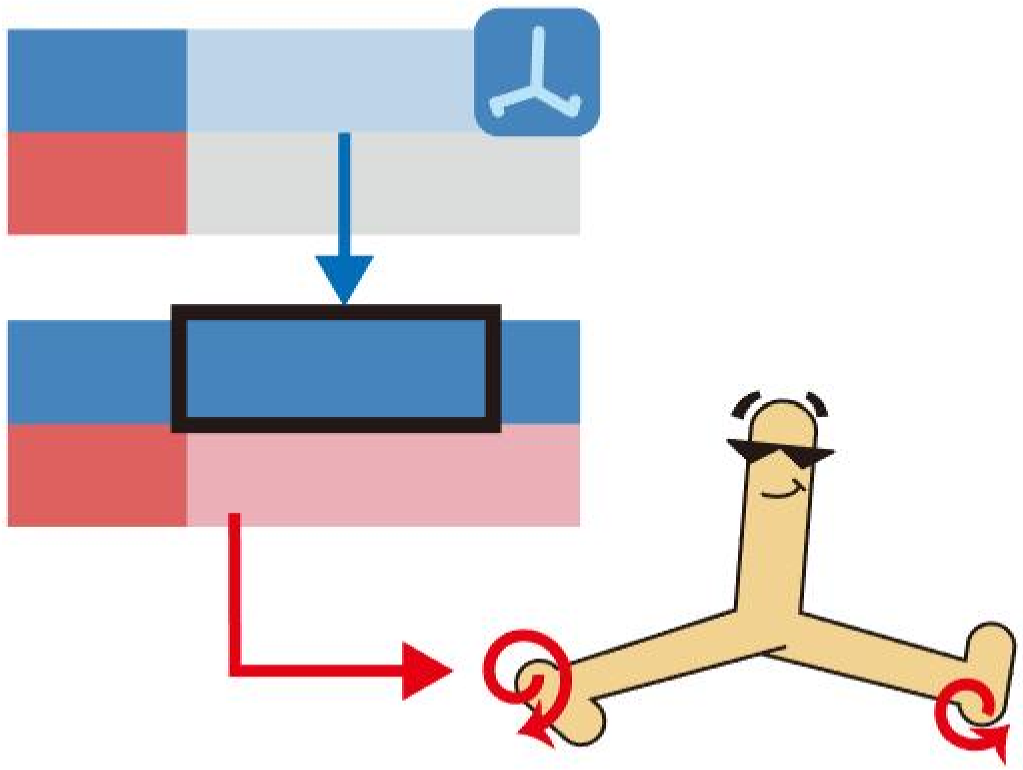
Backward M³PC



Backward M^{3}PC

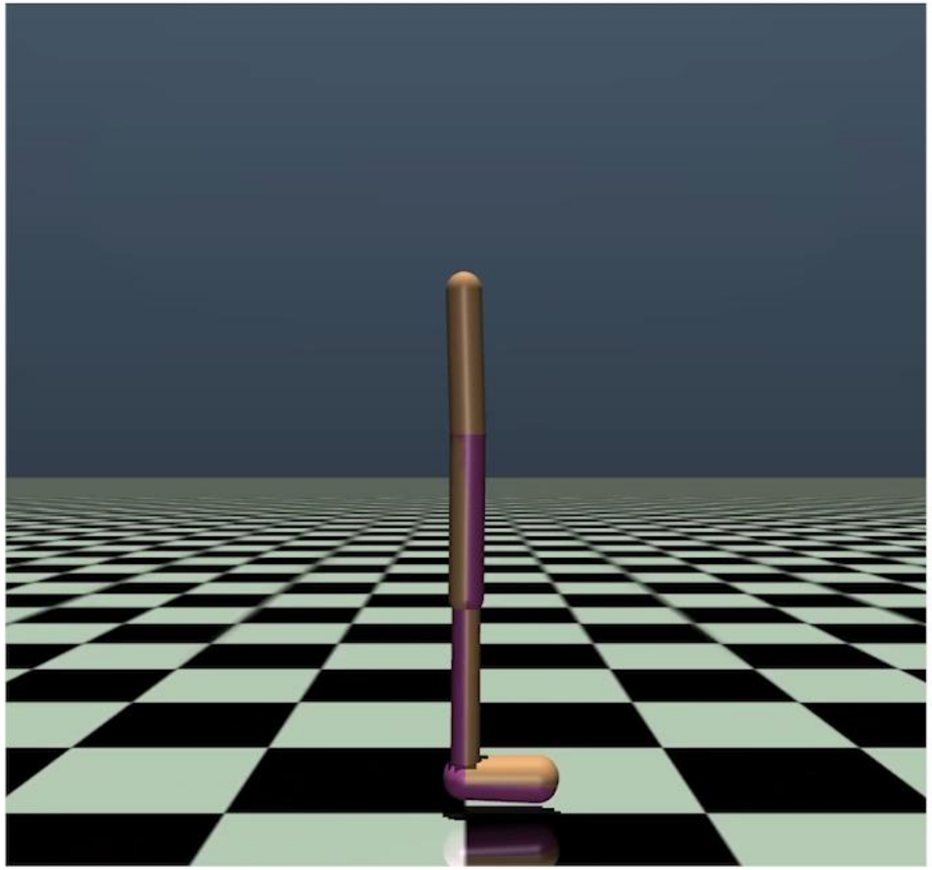


Backward M^{3}PC

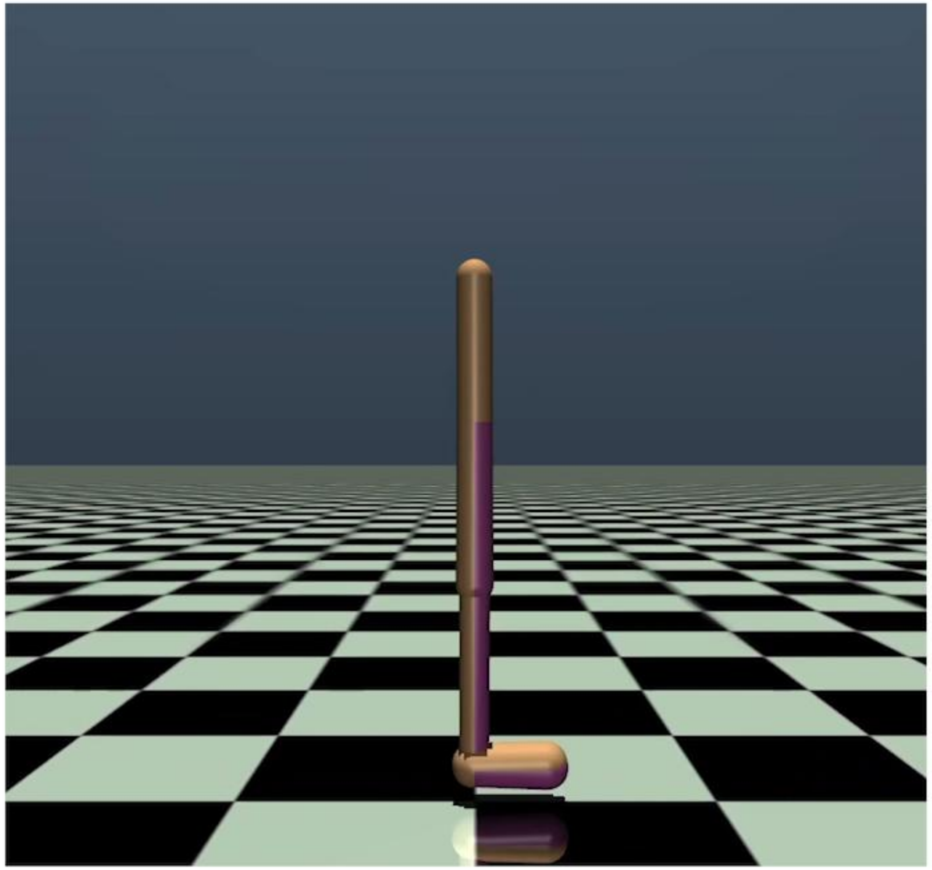


Video results

D4RL: walker2D

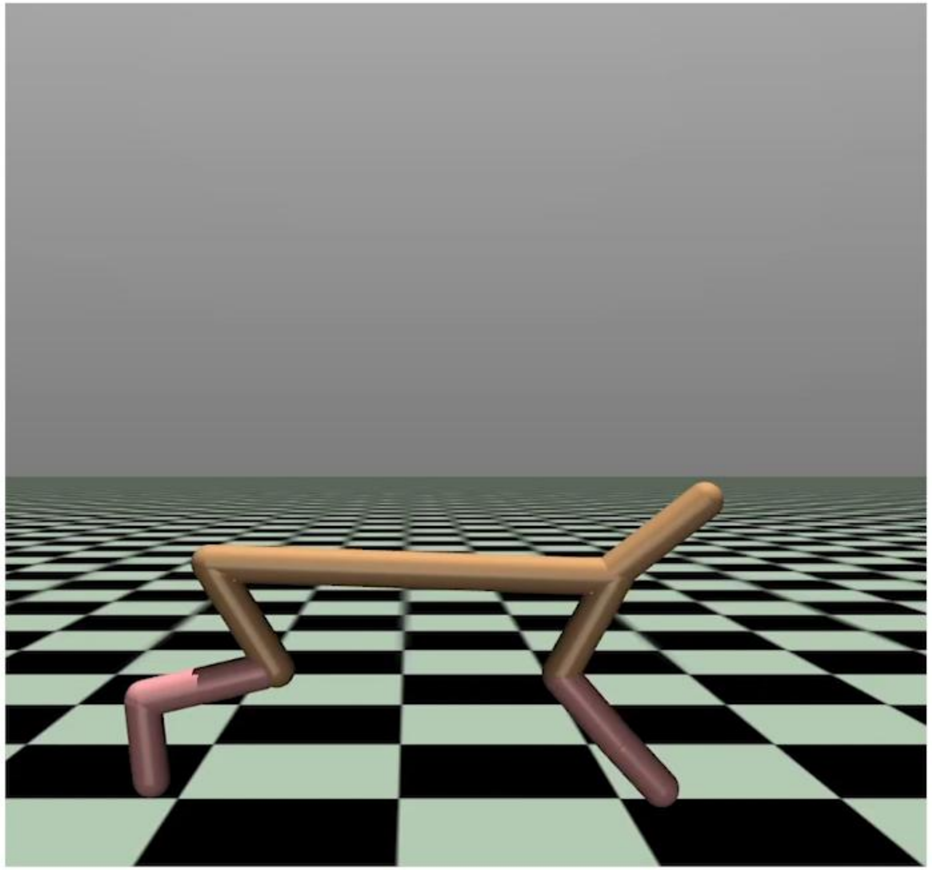


Original task

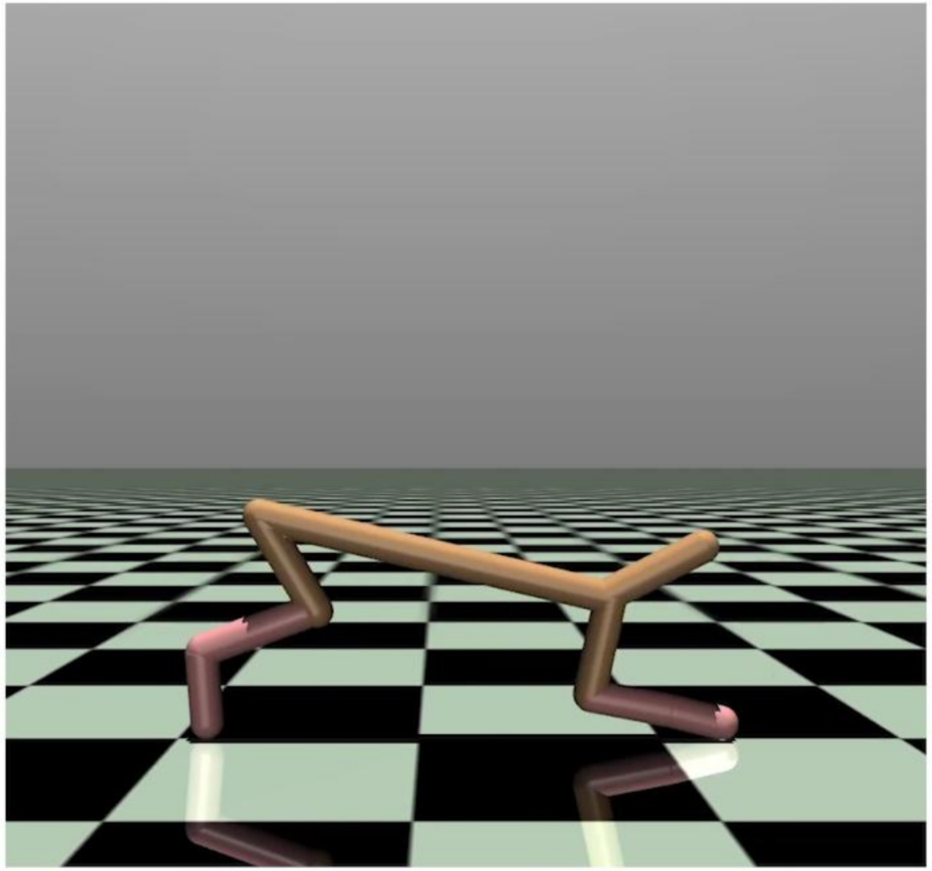


Goal-reaching task: split

D4RL: half-cheetah

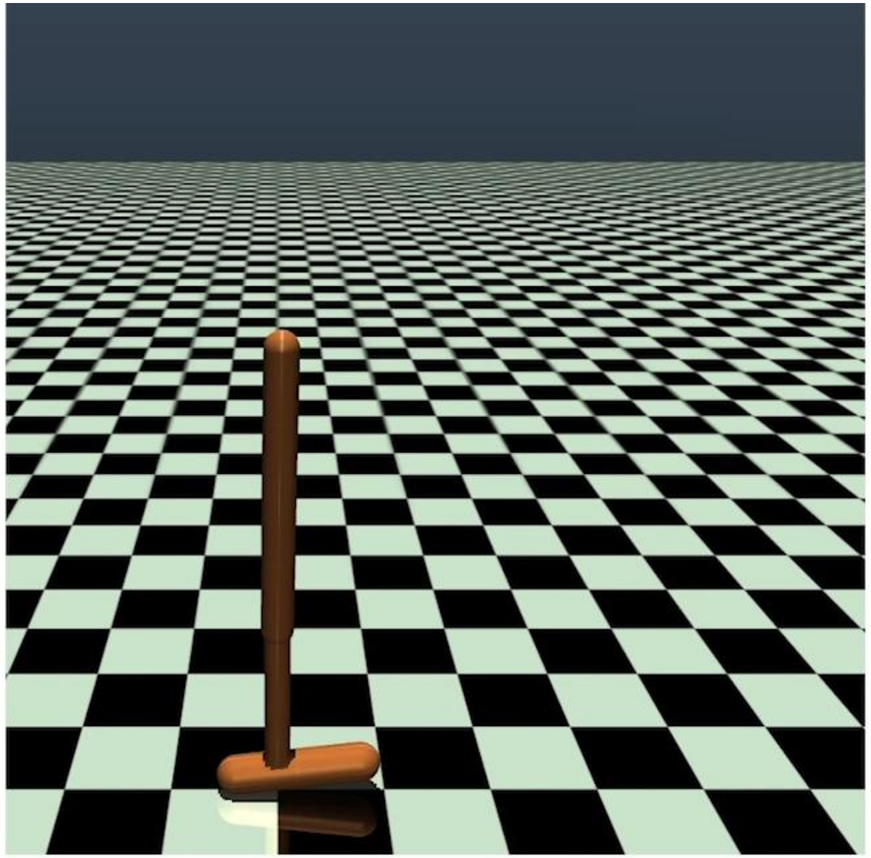


Original task

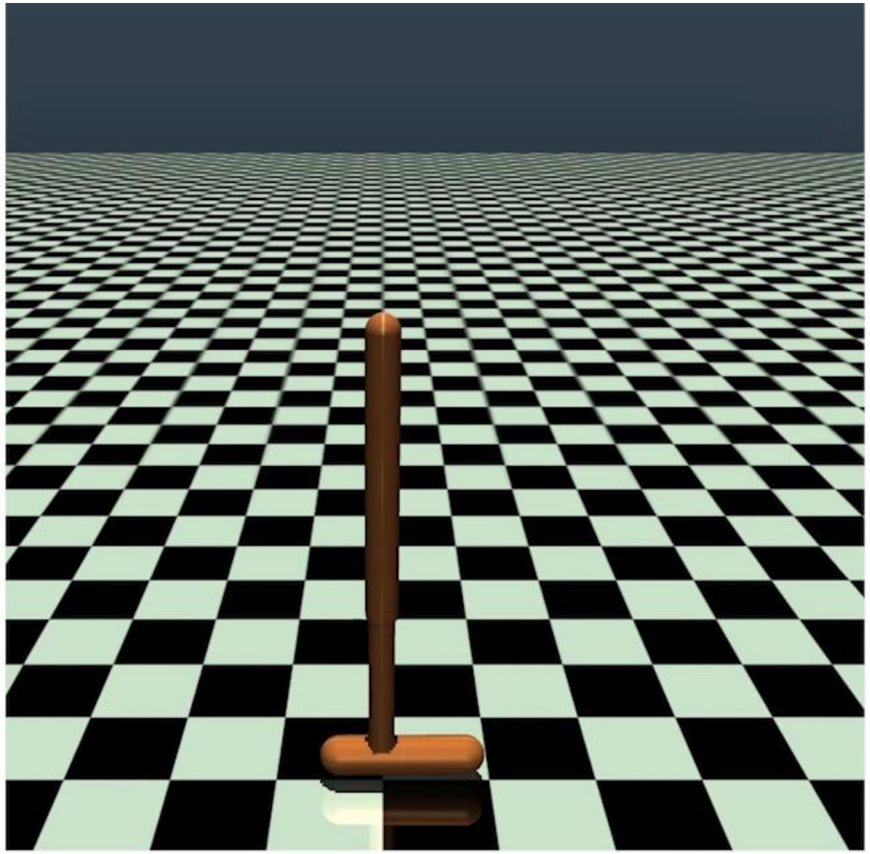


Goal-reaching task: flip

D4RL: hopper

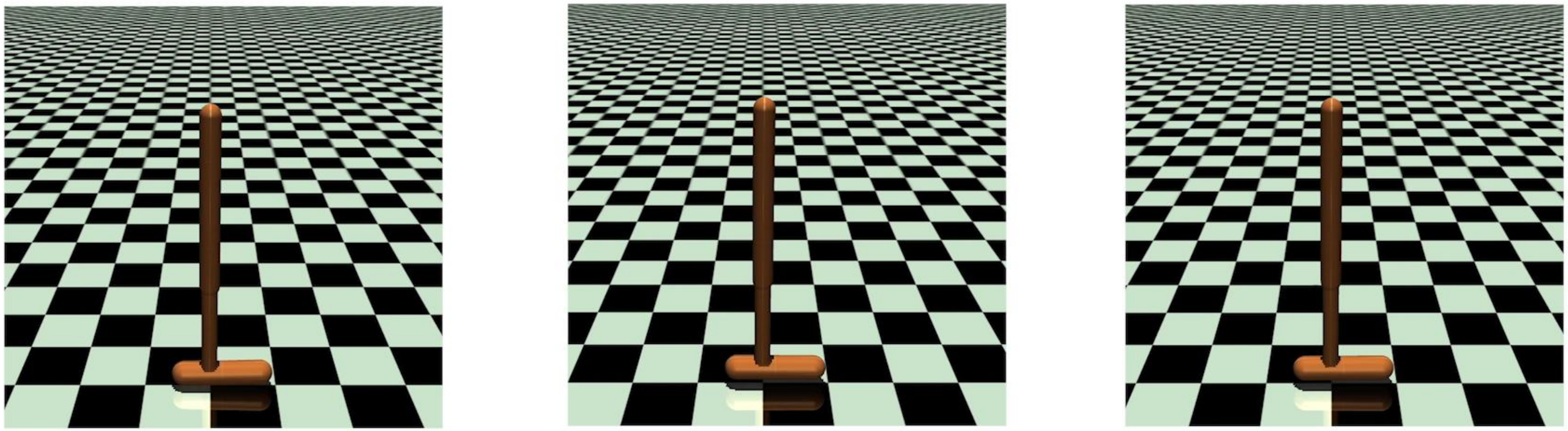


Original task

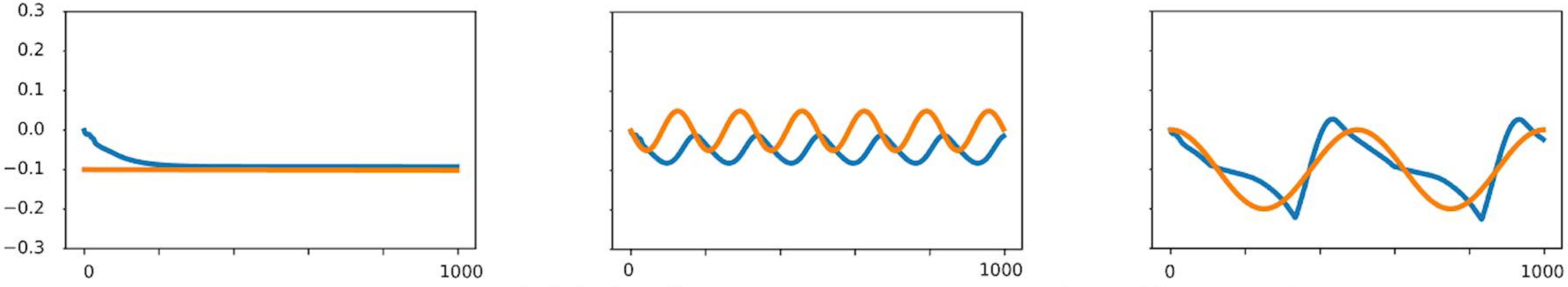


Goal-reaching task: wiggle

D4RL: hopper

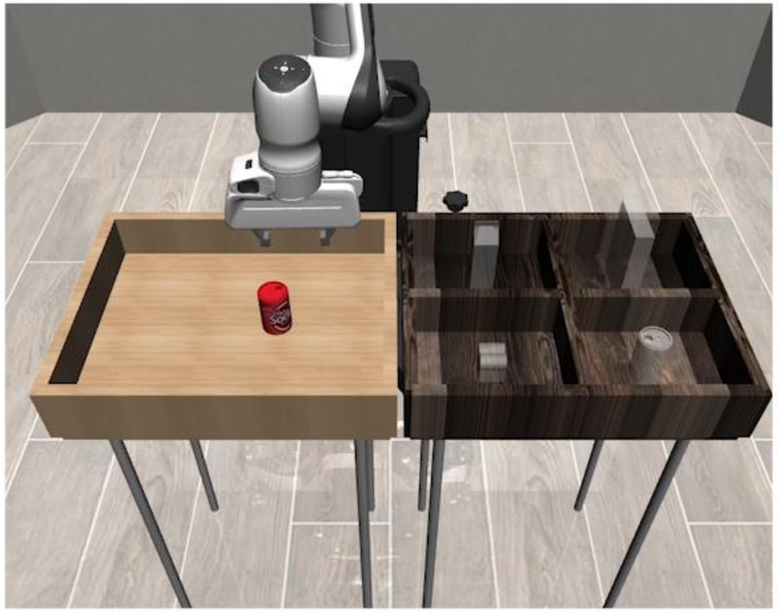


frequency: 0 **frequency: 6** **frequency: 2**

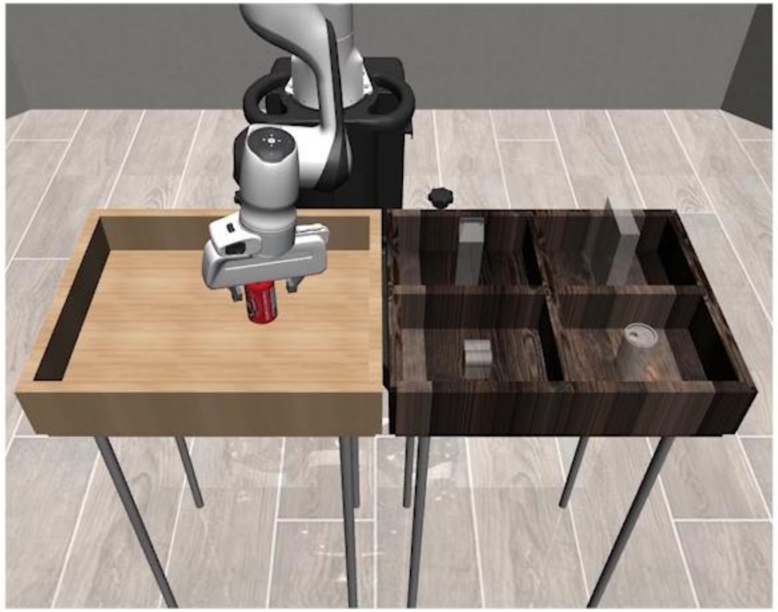


RoboMimic: Can-Pick

Goal-reaching task (unseen)

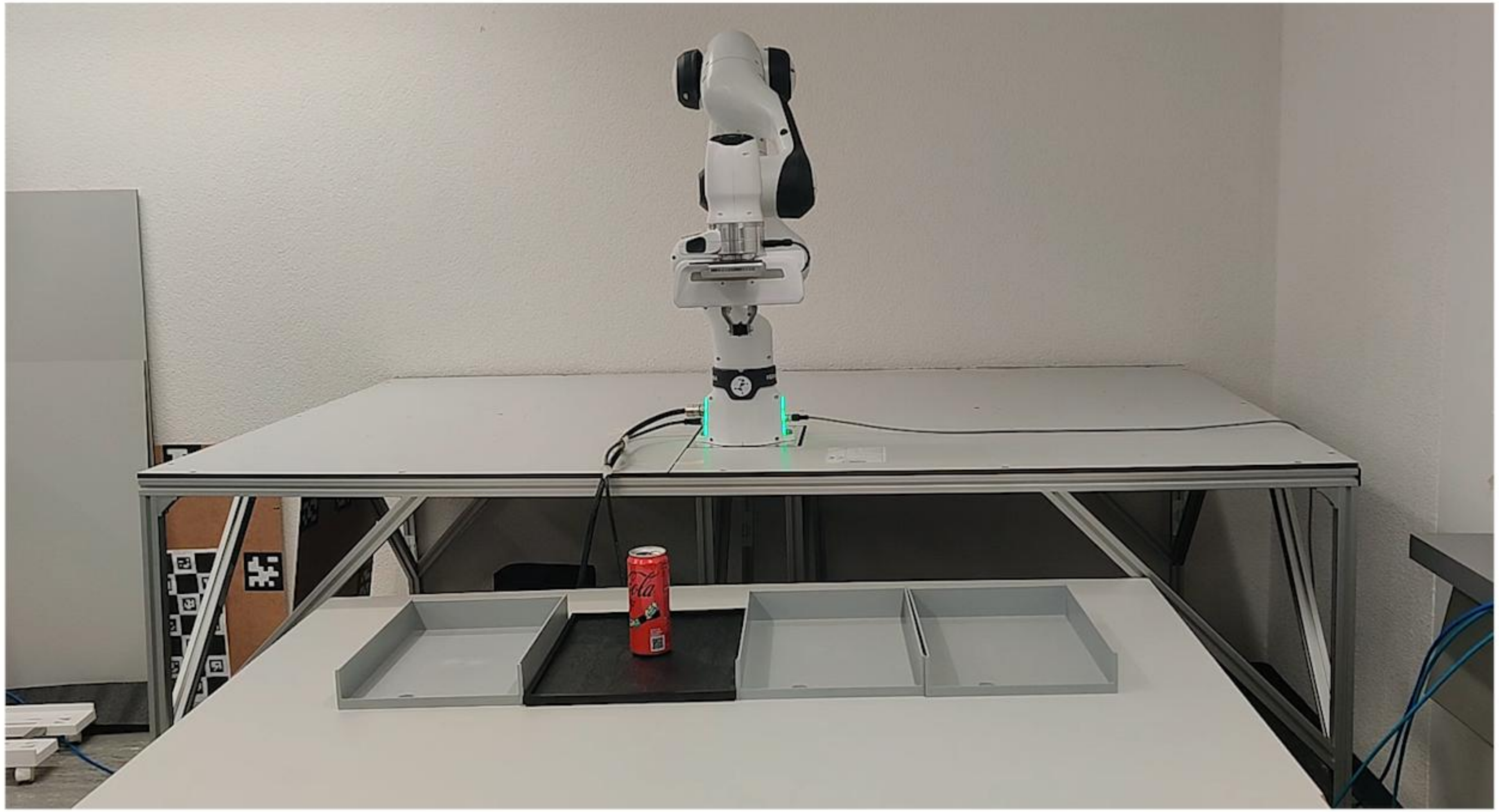


Goal-reaching task (seen)



Original task

Real World: Can-Pick



Goal-reaching task (unseen)

Original task (seen)