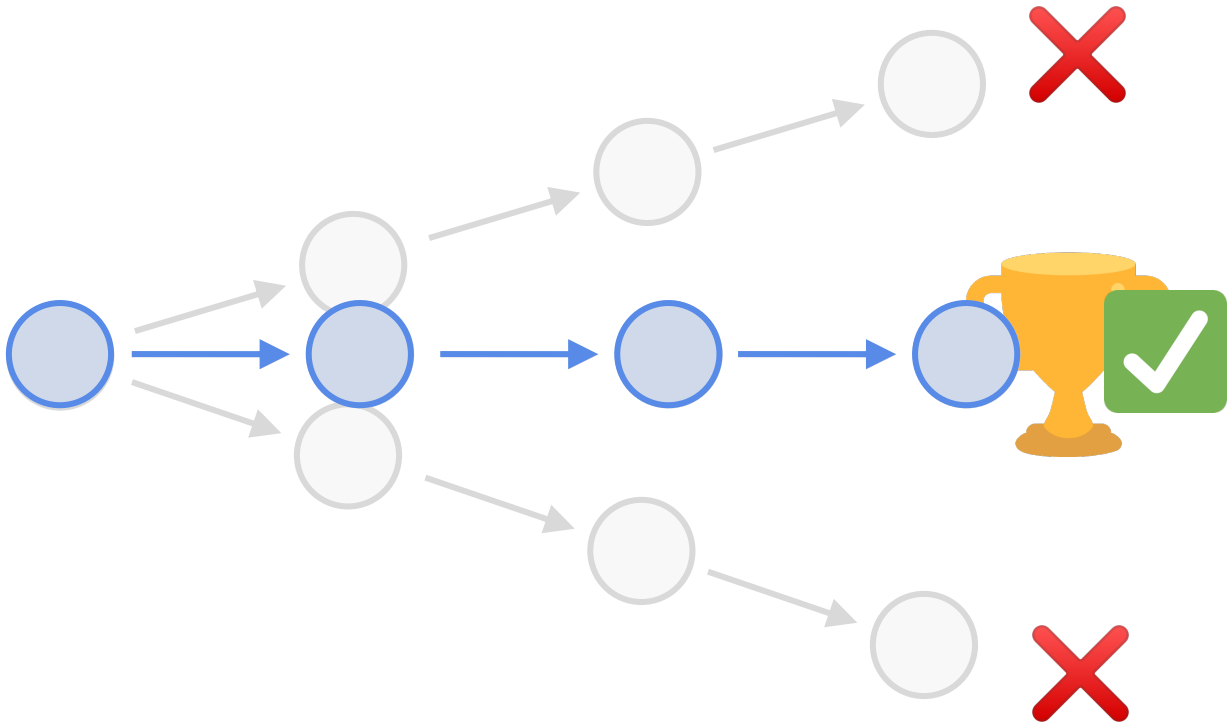
Scalable Offline Model-Based RL with Action Chunks

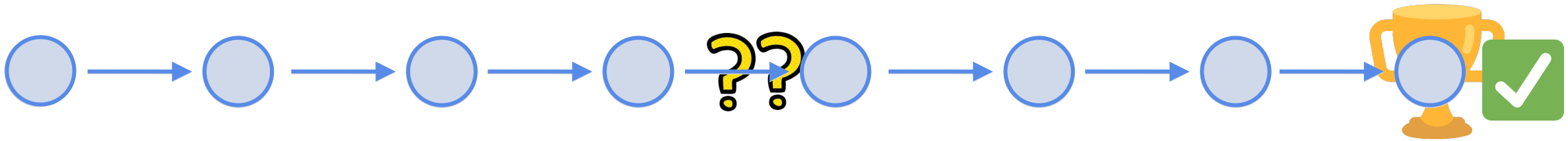
Kwanyoung Park, Seohong Park, Youngwoon Lee, Sergey Levine

In model-based RL, the agent imagines future trajectories of the policy, efficiently finding a best policy.

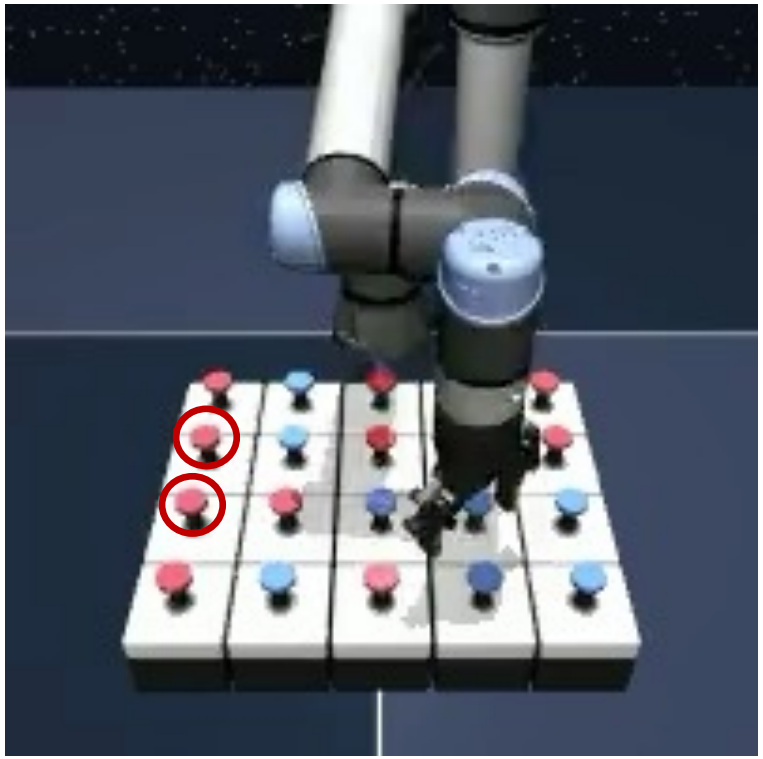


For long-horizon tasks, we need **long enough** model rollouts to propagate the learning signals.

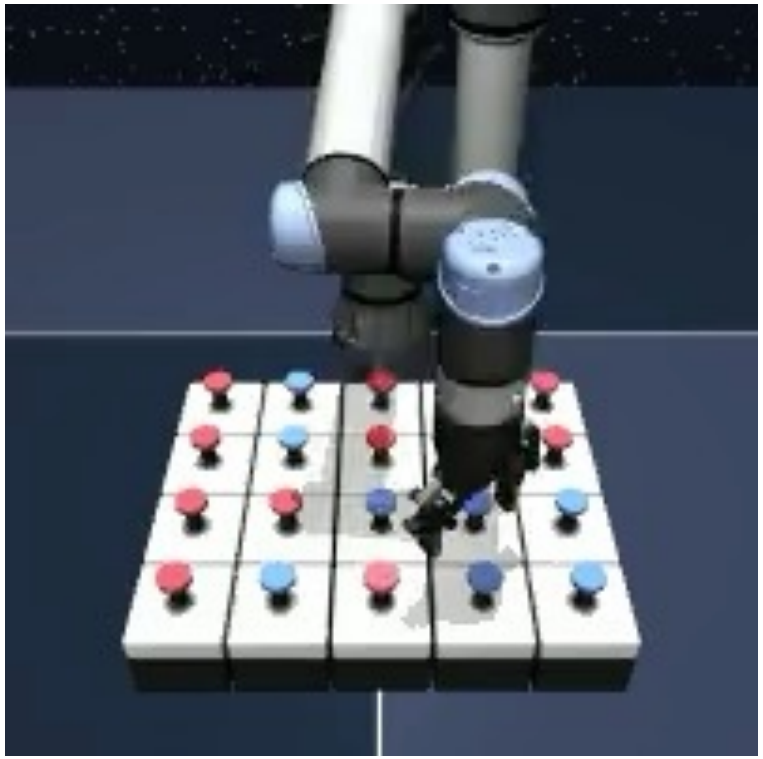
For long-horizon tasks, we need **long enough** model rollouts to propagate the learning signals.



However, model prediction becomes **unstable** when we generate long model rollouts.



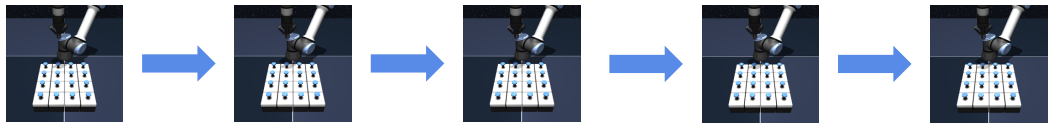
Model prediction



Real

Thus, prior works have been using rollout steps of 5 to 15,
which is **insufficient** for solving long-horizon tasks.

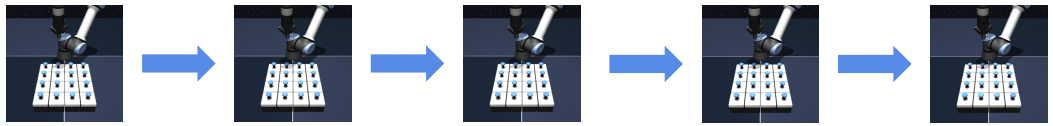
Standard MBRL (MOPO, MOBILE)



Long-horizon MBRL (LEQ, Dreamer)

Thus, prior works have been using rollout steps of 5 to 15,
which is **insufficient** for solving long-horizon tasks.

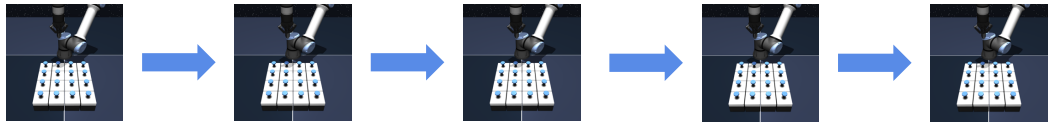
Standard MBRL (MOPO, MOBILE)



Long-horizon MBRL (LEQ, Dreamer)

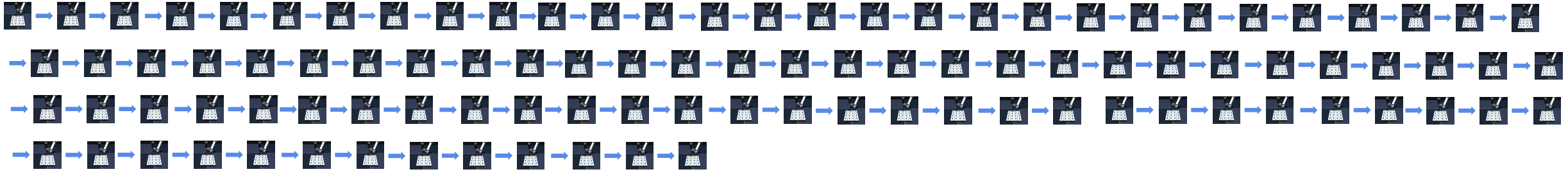
However, our method utilizes 100 steps of rollouts,
pushing the boundary of offline MBRL in long-horizon tasks.

Standard MBRL (MOPO, MOBILE)



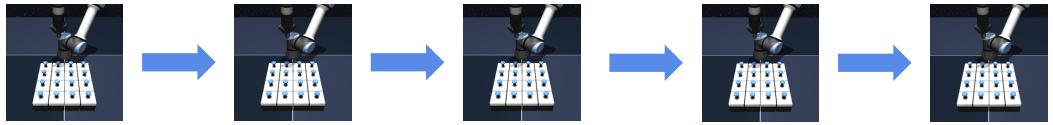
Long-horizon MBRL (LEQ, Dreamer)

Ours (MAC)



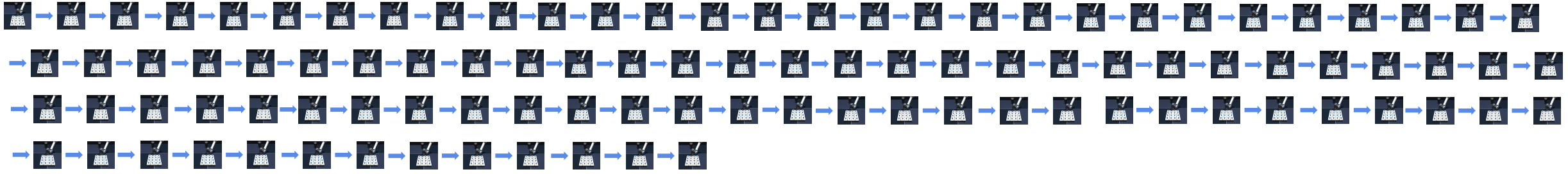
However, our method utilizes 100 steps of rollouts,
pushing the boundary of offline MBRL in long-horizon tasks.

Standard MBRL (MOPO, MOBILE)

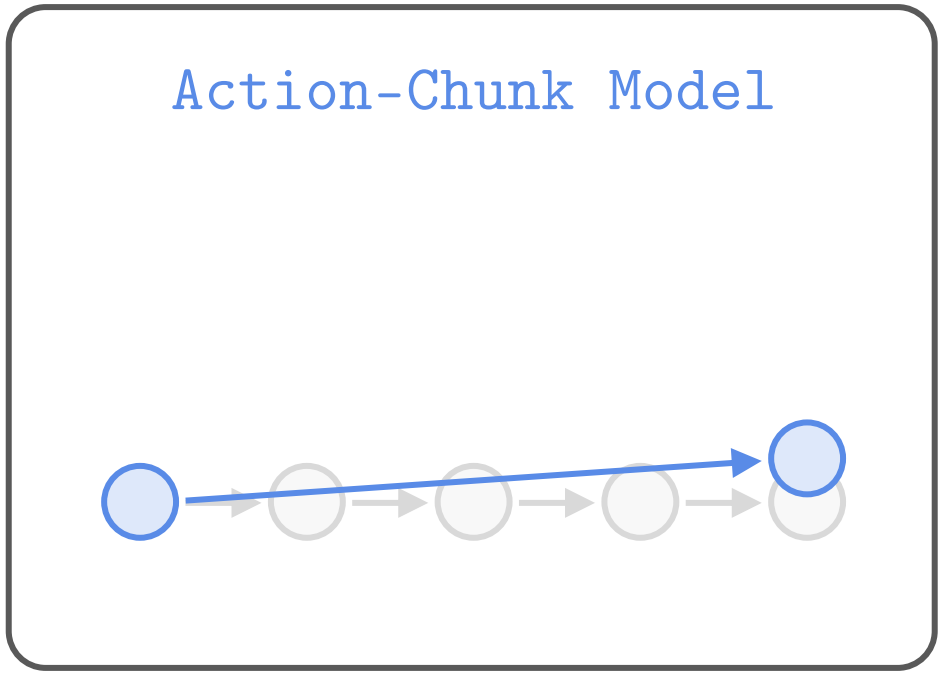


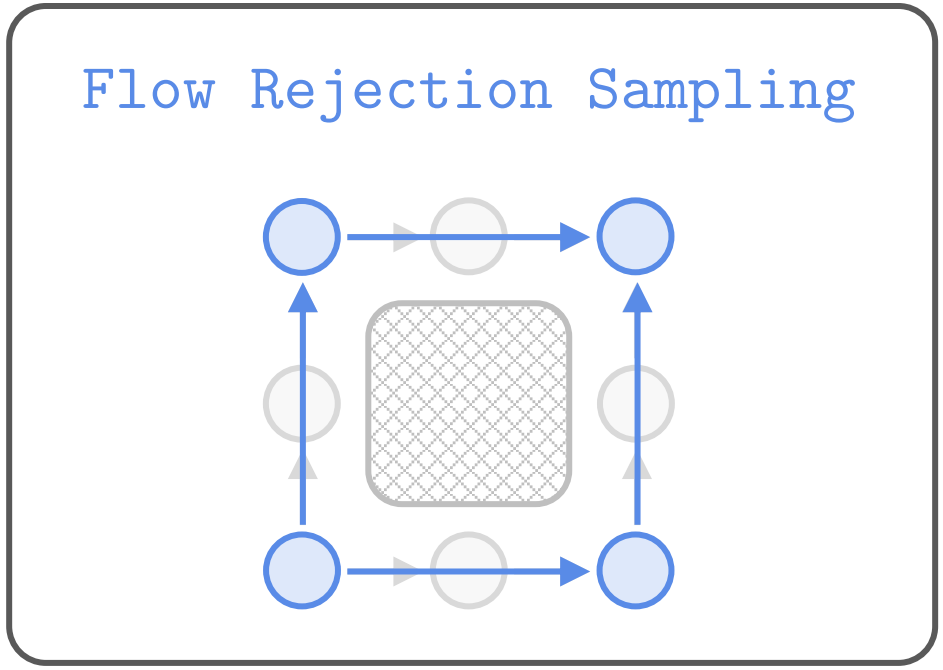
Long-horizon MBRL (LEQ, Dreamer)

Ours (MAC)

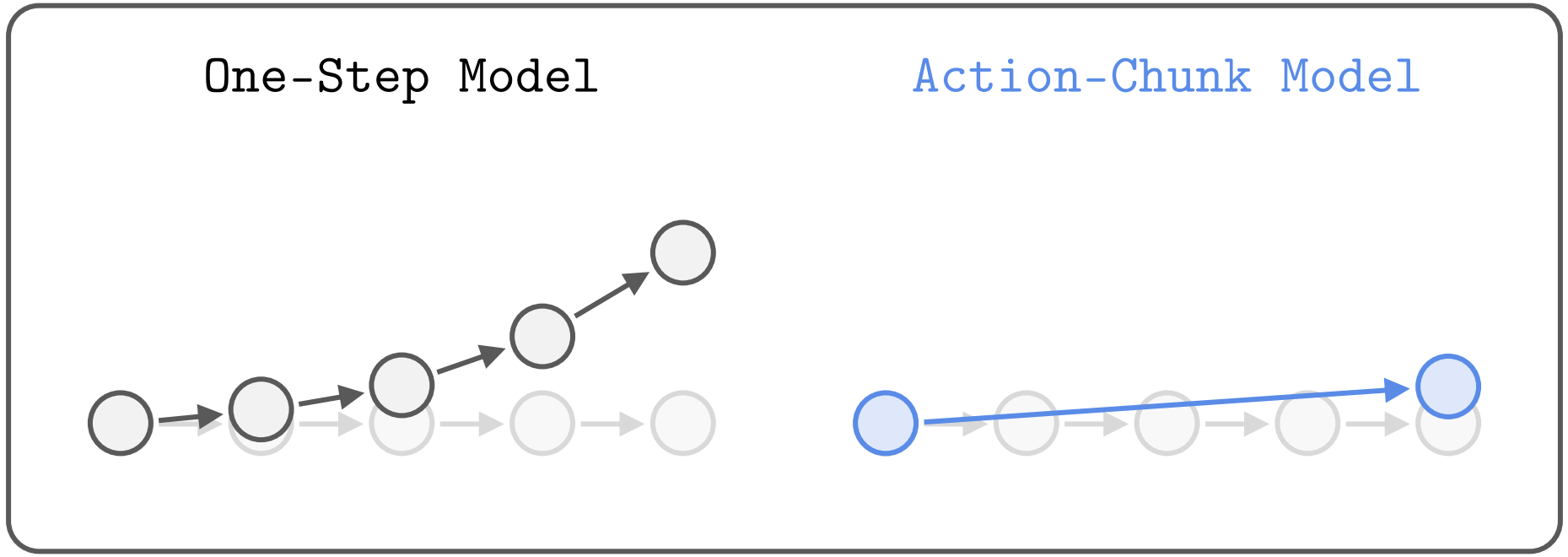


Our core recipes consists of
1) Action chunking, 2) Flow rejection sampling.

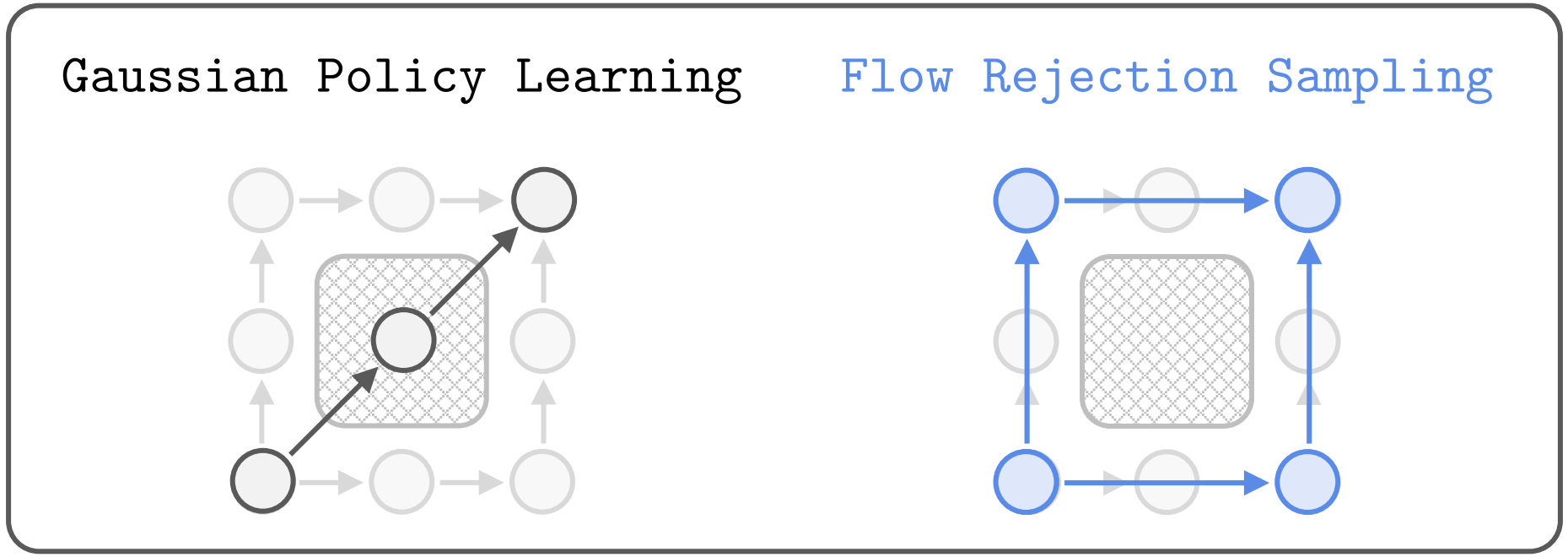




Action chunking reduces the number of model execution, mitigating the compounding error problem.



Rejection sampling from flow BC policy enables modeling multi-modal action distributions, preventing OOD actions.



Specifically, MAC consists of 3 steps:

Step 1: Action-chunked flow BC policy and world model

Flow BC policy: $\pi_\beta(a_{t:t+h} \mid s_t)$

World model: $p(\hat{r}_{t:t+h}, \hat{s}_{t+h} \mid s_t, a_{t:t+h})$

Step 2: Policy extraction with rejection sampling

$$\hat{a}_{t:t+h} = \max_i Q\left(s_t, a^{(i)}_{t:t+h}\right), \text{ where } a^{(i)}_{t:t+h} \sim \pi_\beta(\cdot \mid s_t)$$

Step 3: Q-learning with model-generated rollouts

$$Q(s_t, a_t) \leftarrow \gamma^{hn} Q\left(\hat{s}_{t+nh}, \hat{a}_{t+(n-1)h:t+nh}\right) + \sum_{i=0}^{n} \gamma^{hi} \hat{r}_{t+ih:t+(i+1)h}$$

(On-policy n-step returns with action chunks, n=10, h=10)

On 100M-scale datasets for long-horizon tasks, our method **drastically improves** the performance of offline MBRL approaches.

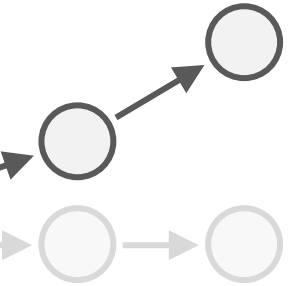
Environment	MOPO	MOBILE	LEQ	FMPC	MAC
humanoidmaze-medium-navigate-oraclerep-v0	27 ±5	23 ±3	0 ±0	18 ±5	36 ±2
humanoidmaze-giant-navigate-oraclerep-v0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0
cube-double-play-oraclerep-v0	25 ±12	15 ±3	0 ±0	37 ±13	100 ±1
cube-octuple-play-oraclerep-v0	0 ±0	0 ±0	0 ±0	0 ±0	30 ±6
puzzle-3x3-play-oraclerep-v0	19 ±2	15 ±5	1 ±1	12 ±6	100 ±0
puzzle-4x5-play-oraclerep-v0	0 ±0	0 ±0	1 ±3	0 ±0	99 ±3

Our method also **drastically improves** the performance of offline MBRL in standard-scale (1M) datasets.

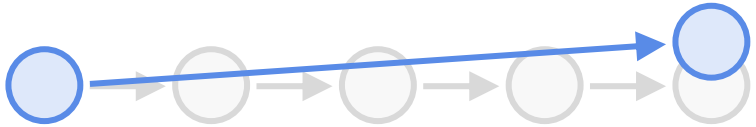
Environment	MOPO	MOBILE	LEQ	FMPC	MAC
cube-single-play-v0 (5 tasks)	12 ± 4	81 ± 8	0 ± 0	9 ± 5	99 ± 2
cube-double-play-v0 (5 tasks)	1 ± 1	1 ± 2	0 ± 0	3 ± 2	53 ± 4
scene-play-v0 (5 tasks)	6 ± 8	8 ± 4	0 ± 0	4 ± 4	97 ± 4
puzzle-3x3-play-v0 (5 tasks)	20 ± 0	12 ± 9	10 ± 7	1 ± 1	20 ± 0
puzzle-4x4-play-v0 (5 tasks)	0 ± 0	0 ± 0	0 ± 0	0 ± 0	78 ± 13

If you are interested in our project, please check the paper!

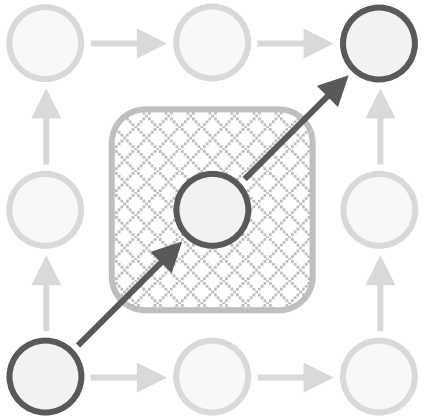
Model



Action-Chunk Model



Gaussian Policy Learning



Flow R