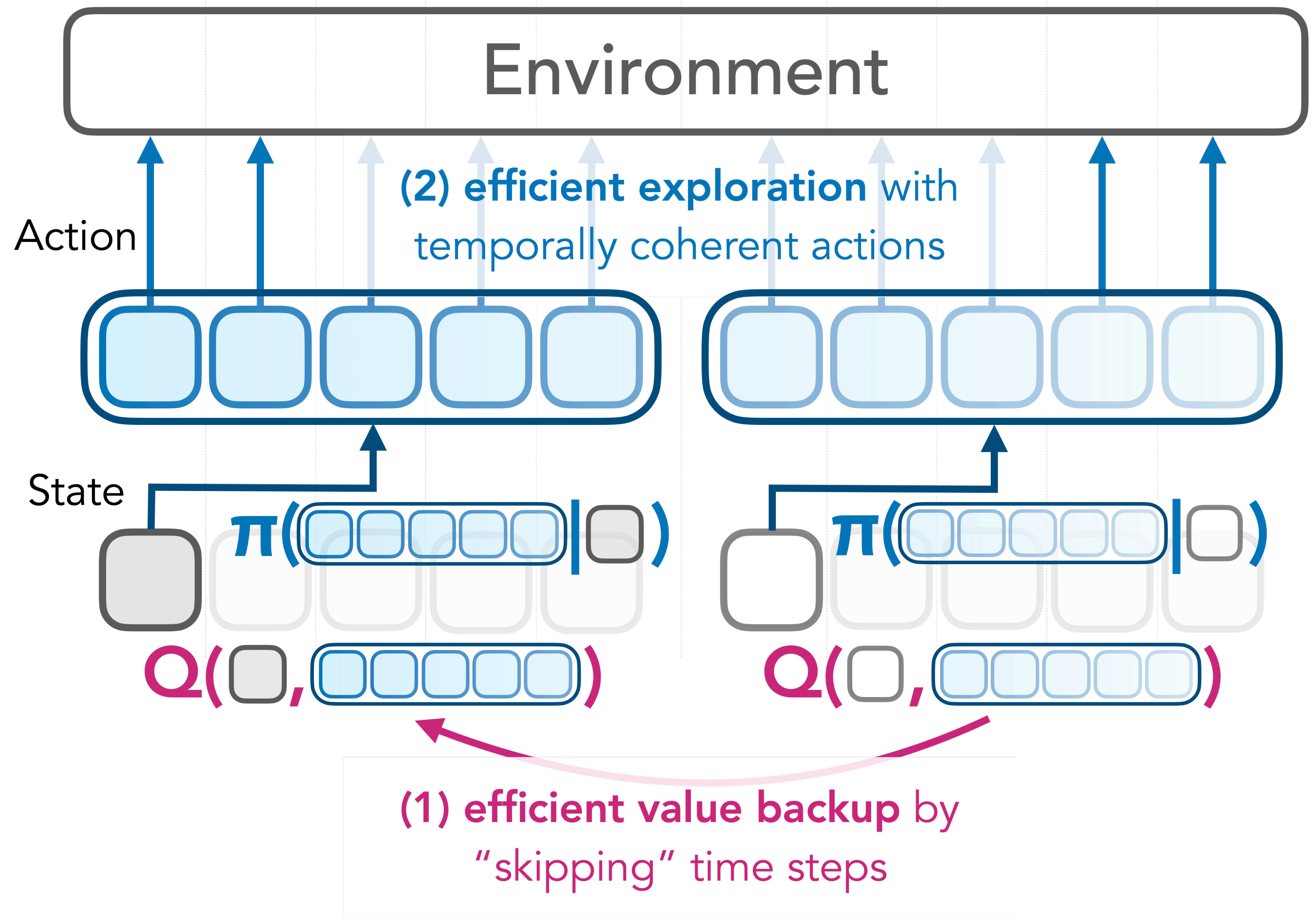
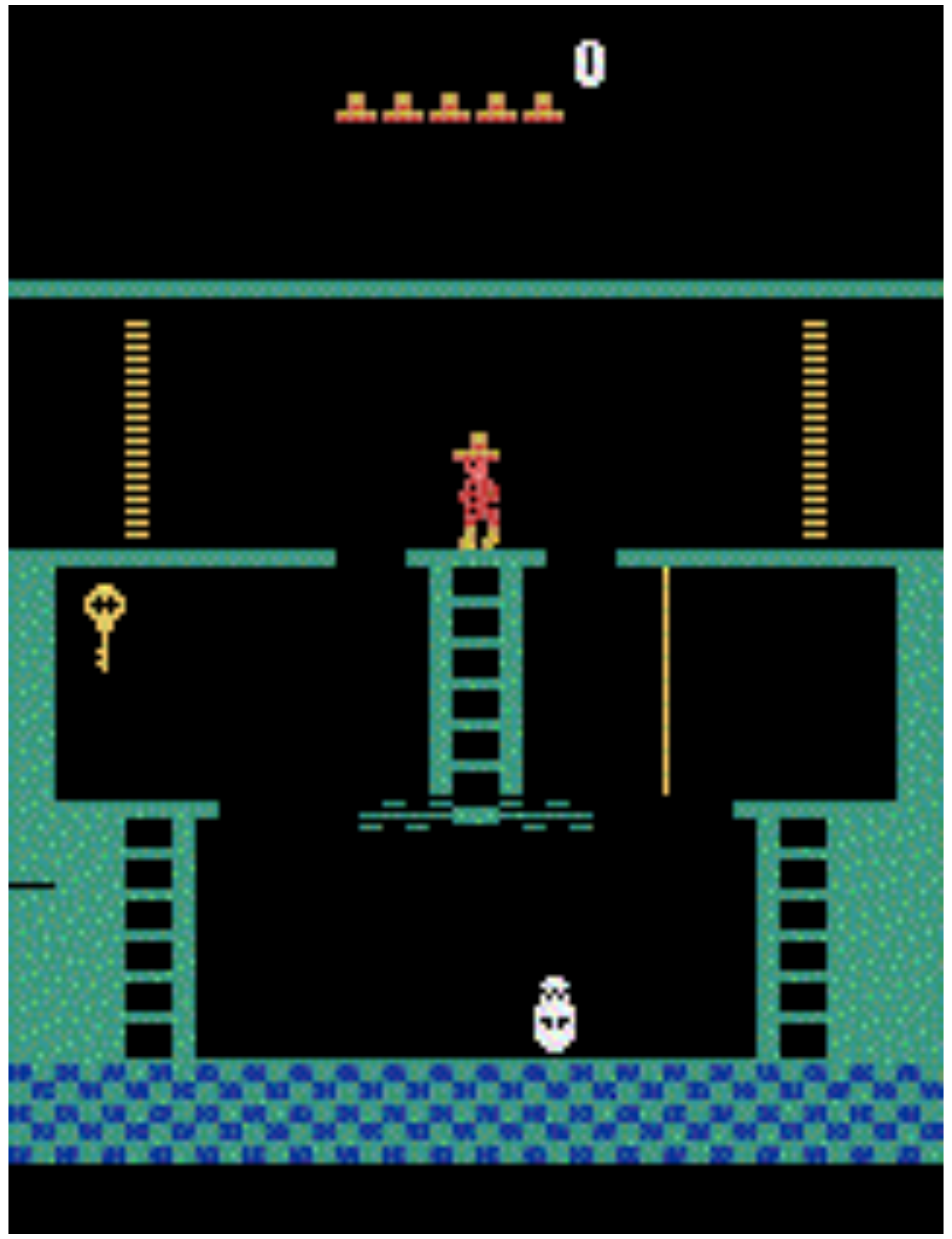
Reinforcement Learning with Action Chunking

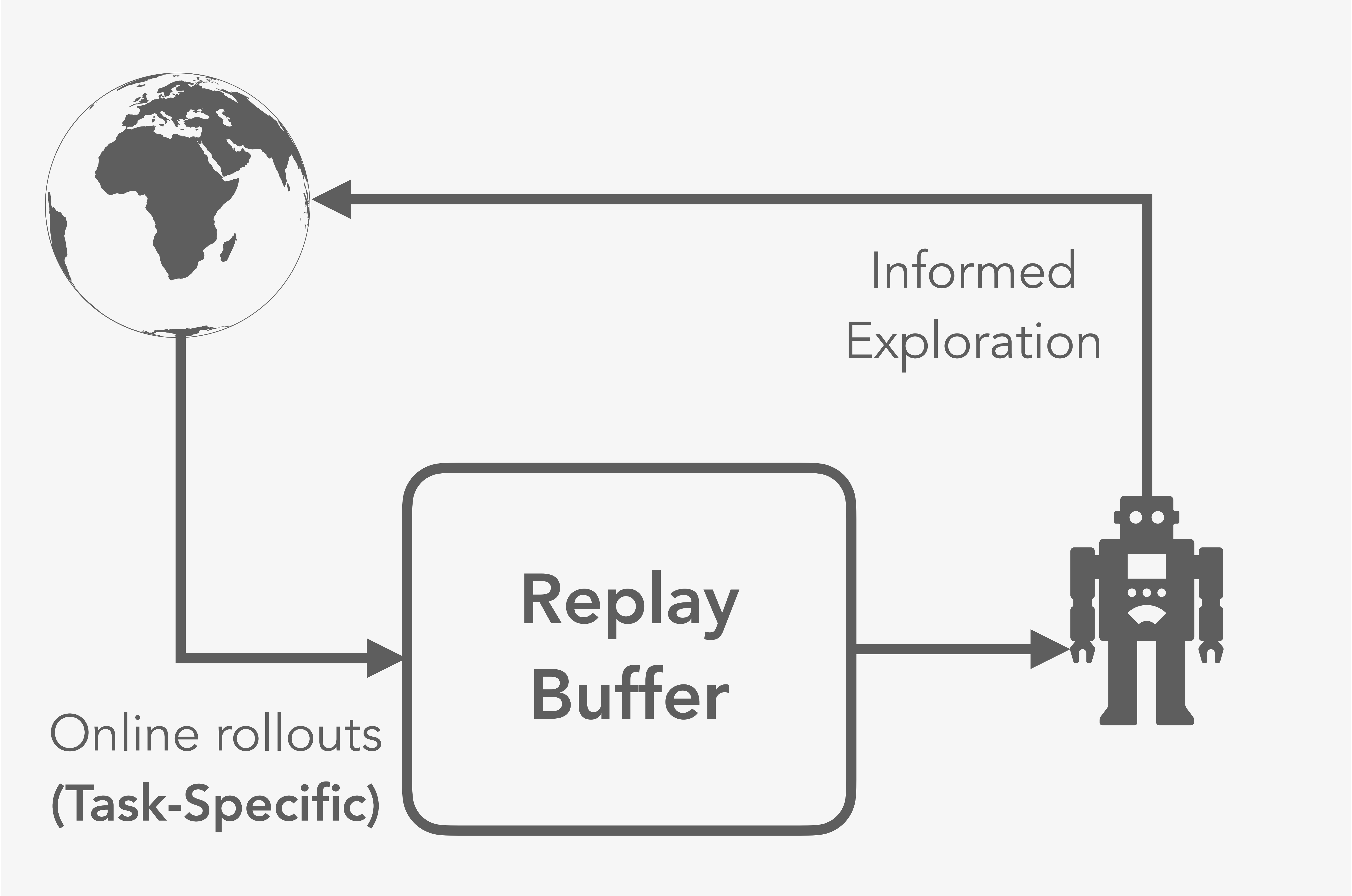

Environment
Action
(2) efficient exploration with temporally coherent actions
State
π(|)
π(|)
Q(,)
Q(,)
(1) efficient value backup by "skipping" time steps

Qiyang (Colin) Li, Zhiyuan (Paul) Zhou, Sergey Levine

UC Berkeley

Exploration is hard

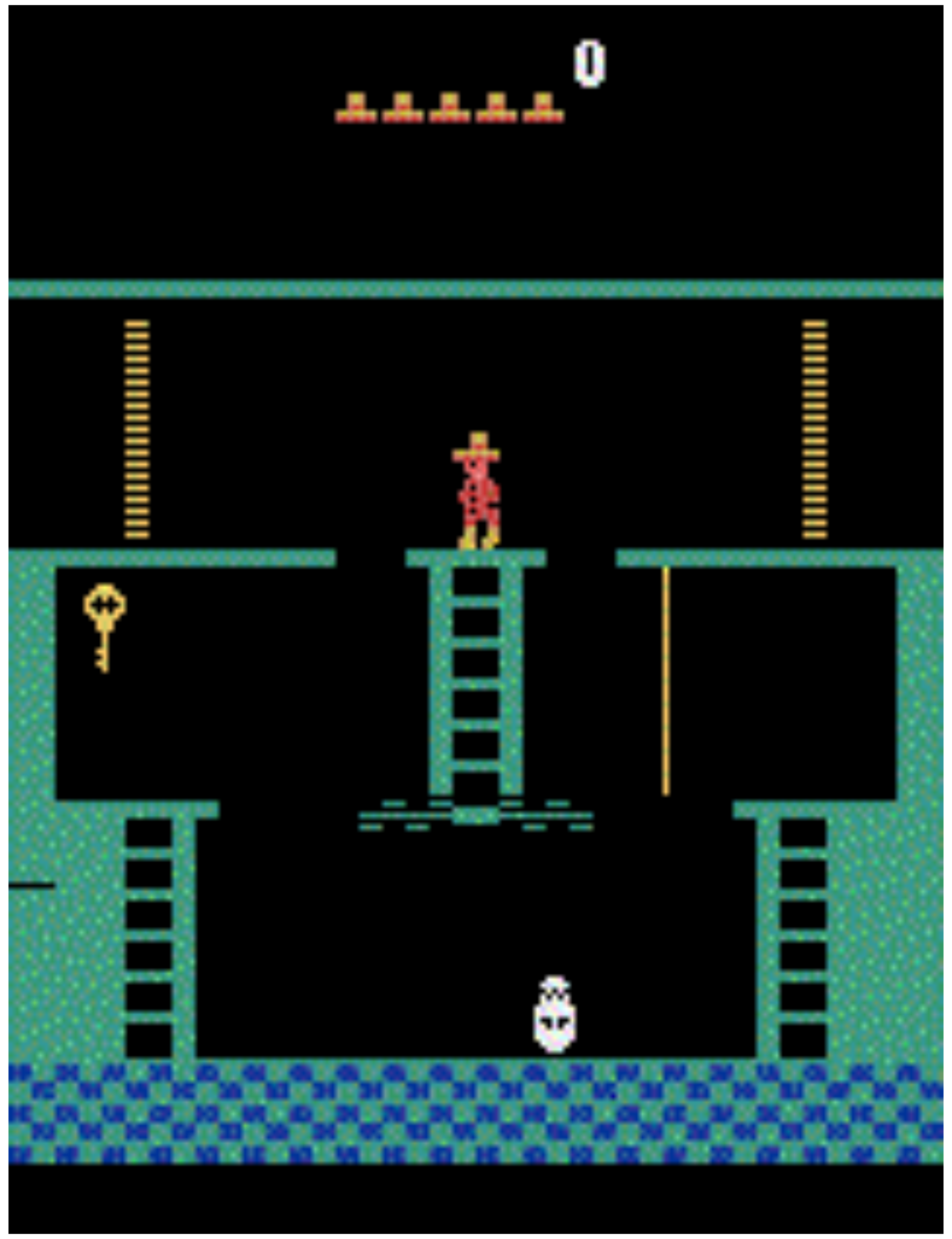


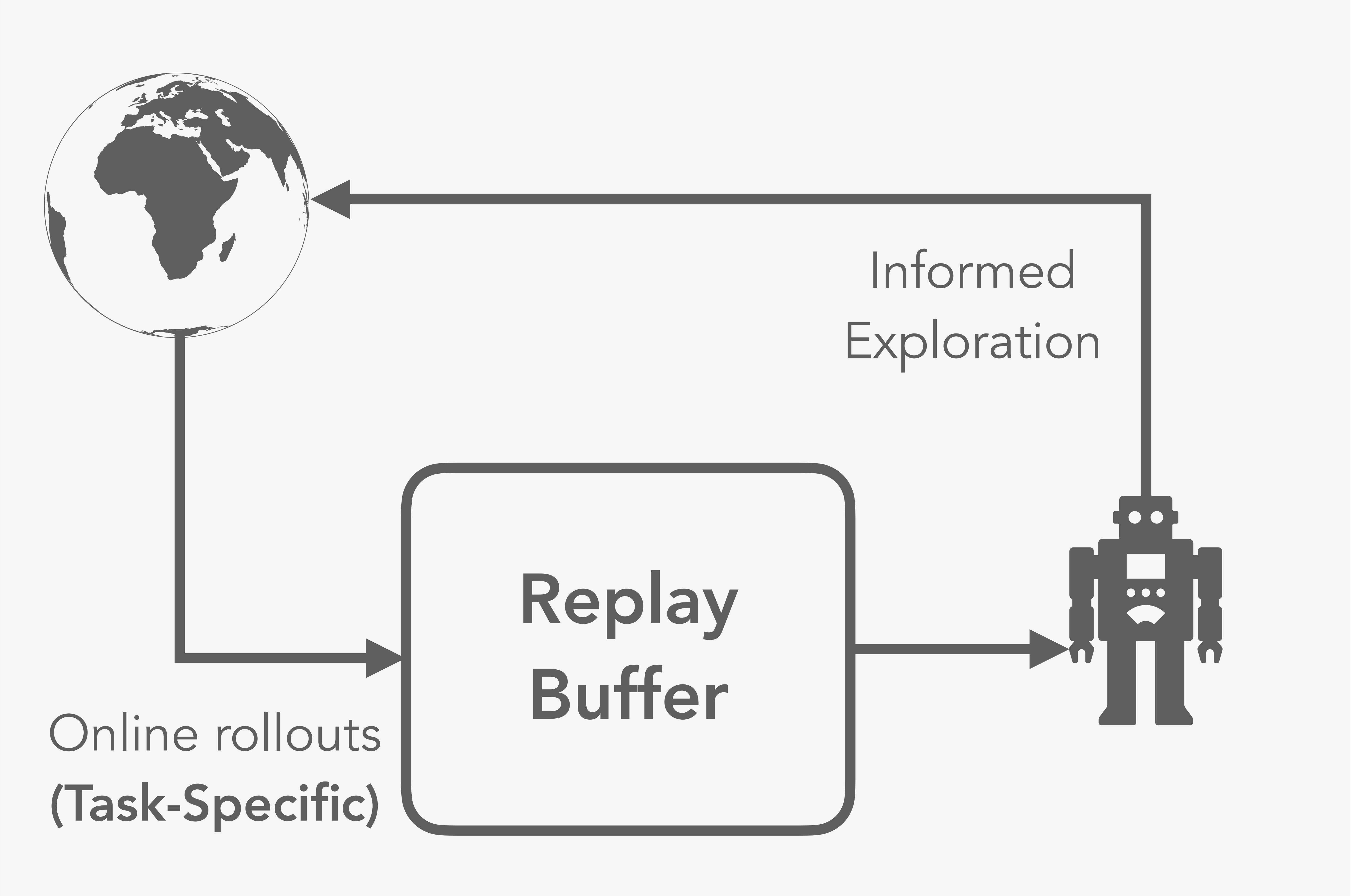


In the worst case, we must reach **every possible state**

www.gymlibrary.dev/environments/atari/montezuma_revenge/

Exploration is hard

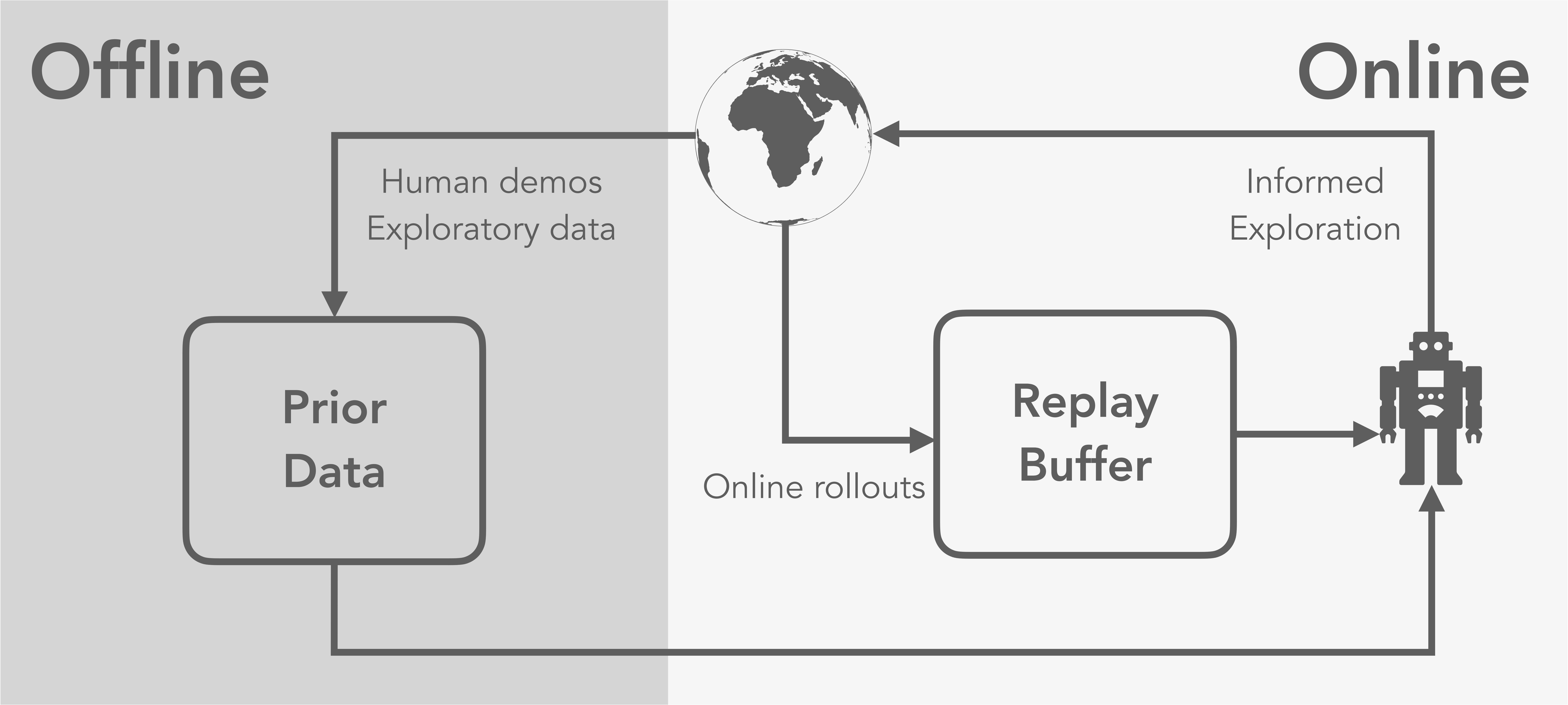




In the worst case, we must reach **every possible state**

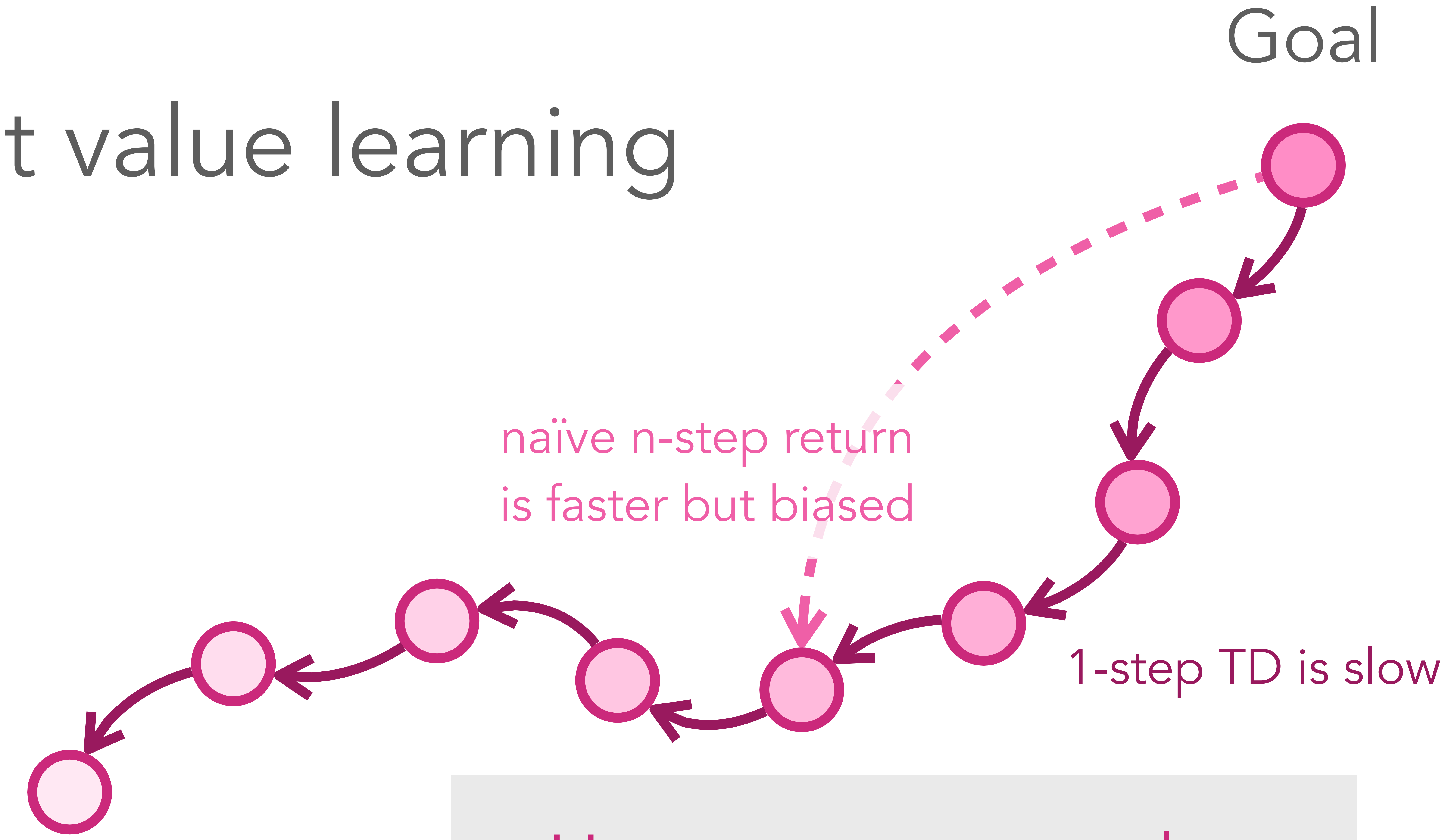
www.gymlibrary.dev/environments/atari/montezuma_revenge/

Offline-to-online RL



Challenges

(1) Inefficient value learning



Goal

naïve n-step return is faster but biased

1-step TD is slow

How can we speedup TD backup?

Challenges

(2) Unstructured Exploration

☆ Goal

□ states visited by the exploration policy

Challenges

(2) Unstructured Exploration

Common, but not ideal:

☆ Goal

□ **states visited by the exploration policy**

Challenges

(2) Unstructured Exploration

Common, but not ideal:

Take the best policy and add temporally independent action noises

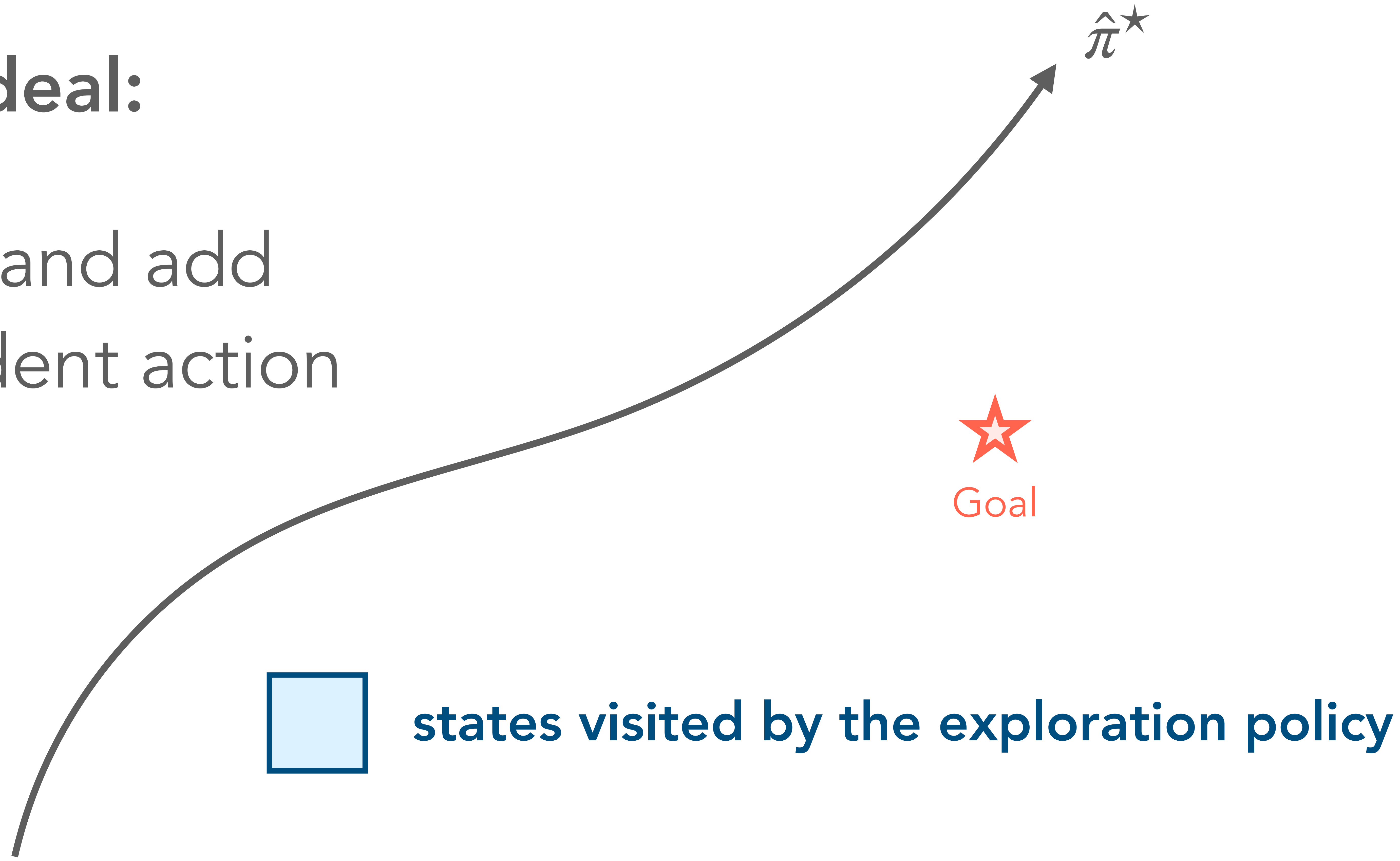
□ **states visited by the exploration policy**

Challenges

(2) Unstructured Exploration

Common, but not ideal:

Take the best policy and add temporally independent action noises

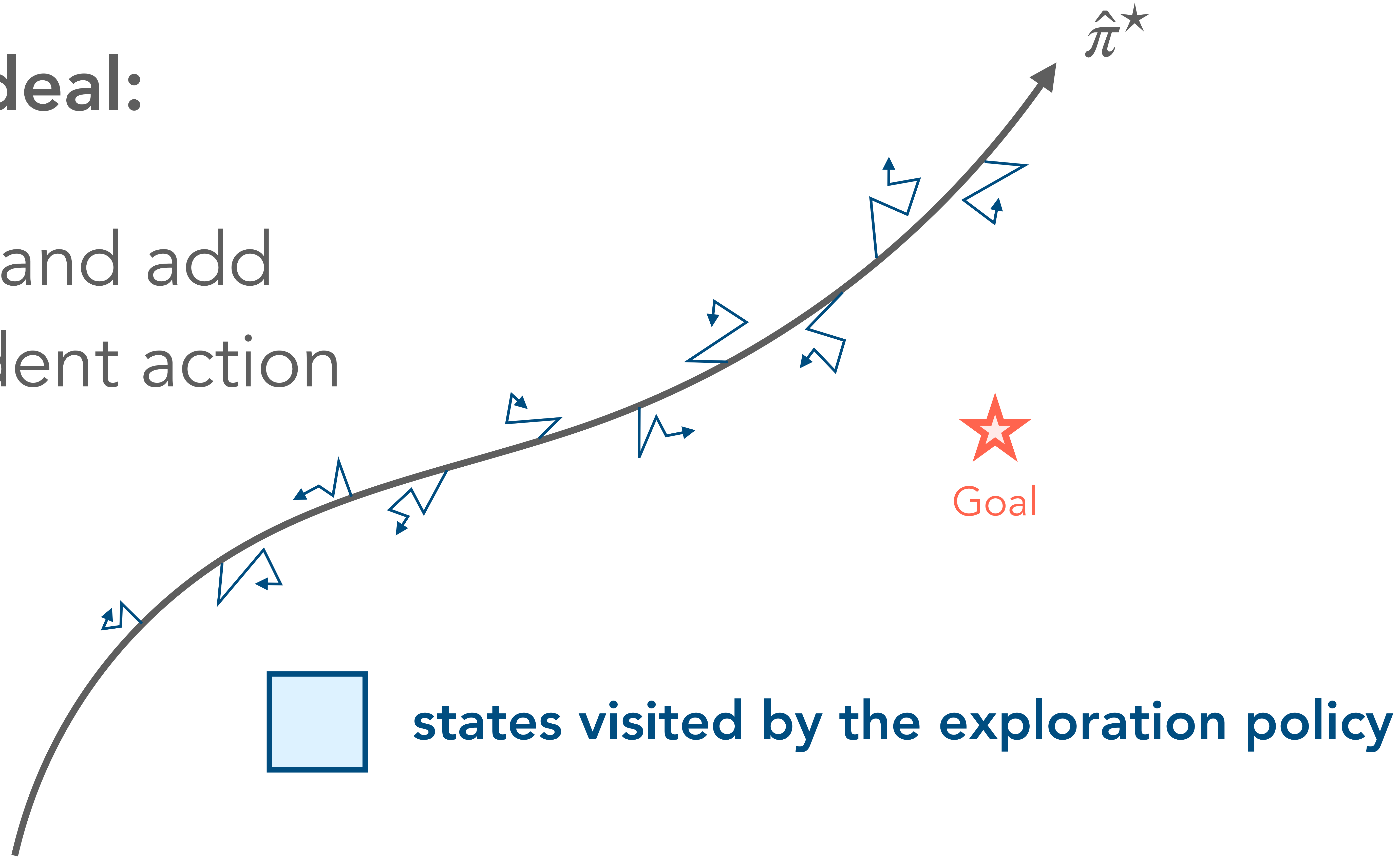


Challenges

(2) Unstructured Exploration

Common, but not ideal:

Take the best policy and add temporally independent action noises

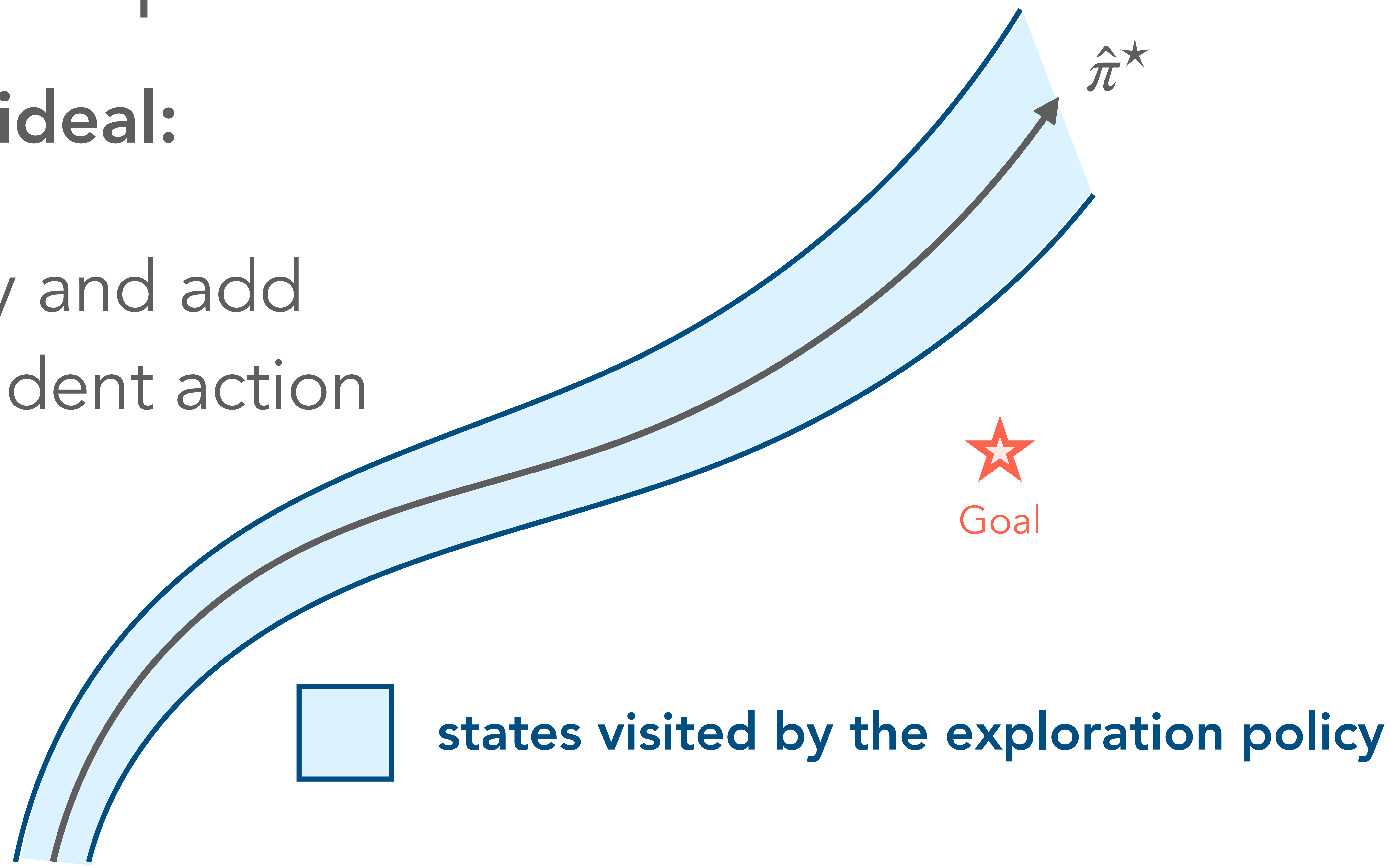


Challenges

(2) Unstructured Exploration

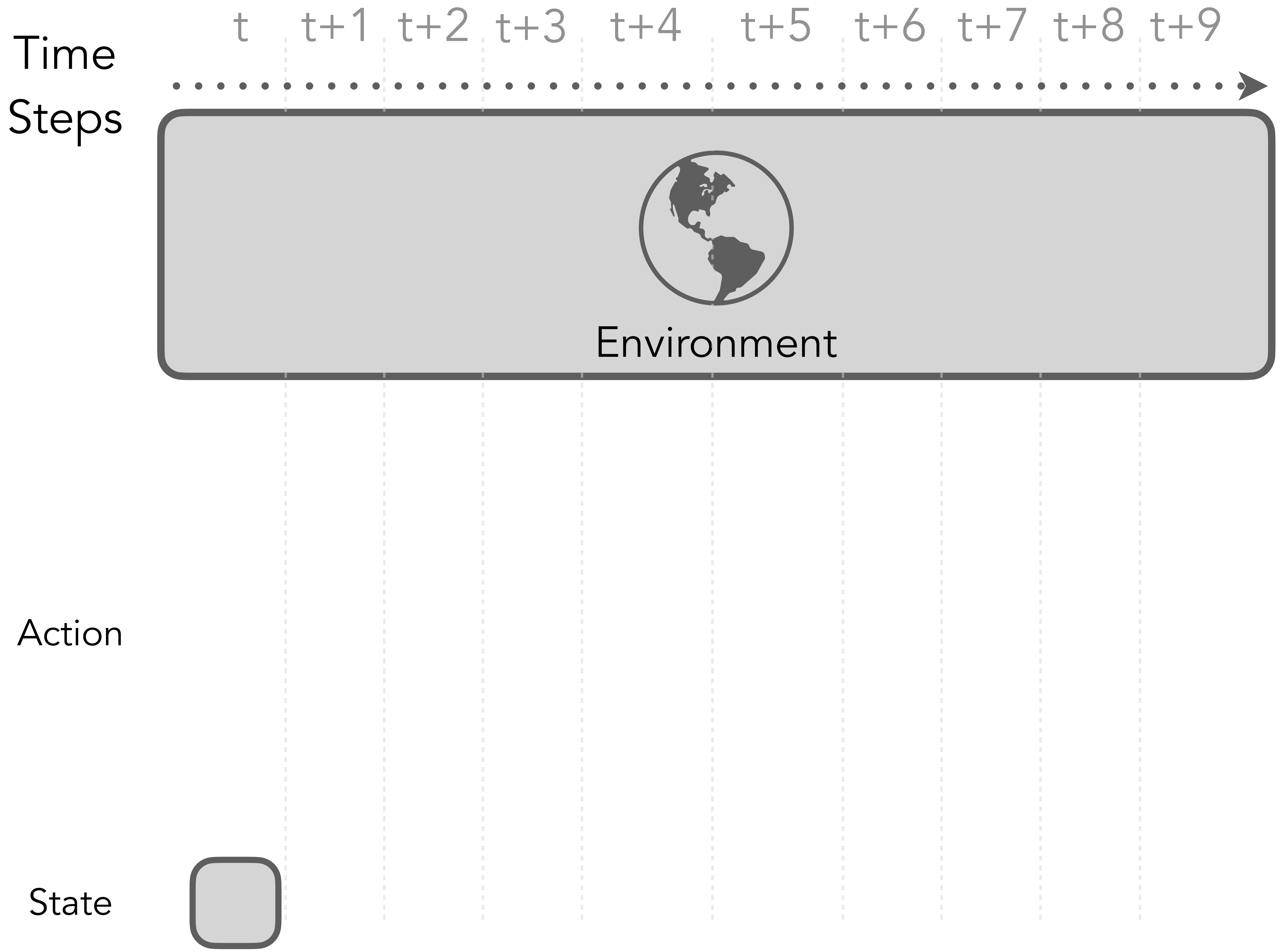
Common, but not ideal:

Take the best policy and add temporally independent action noises



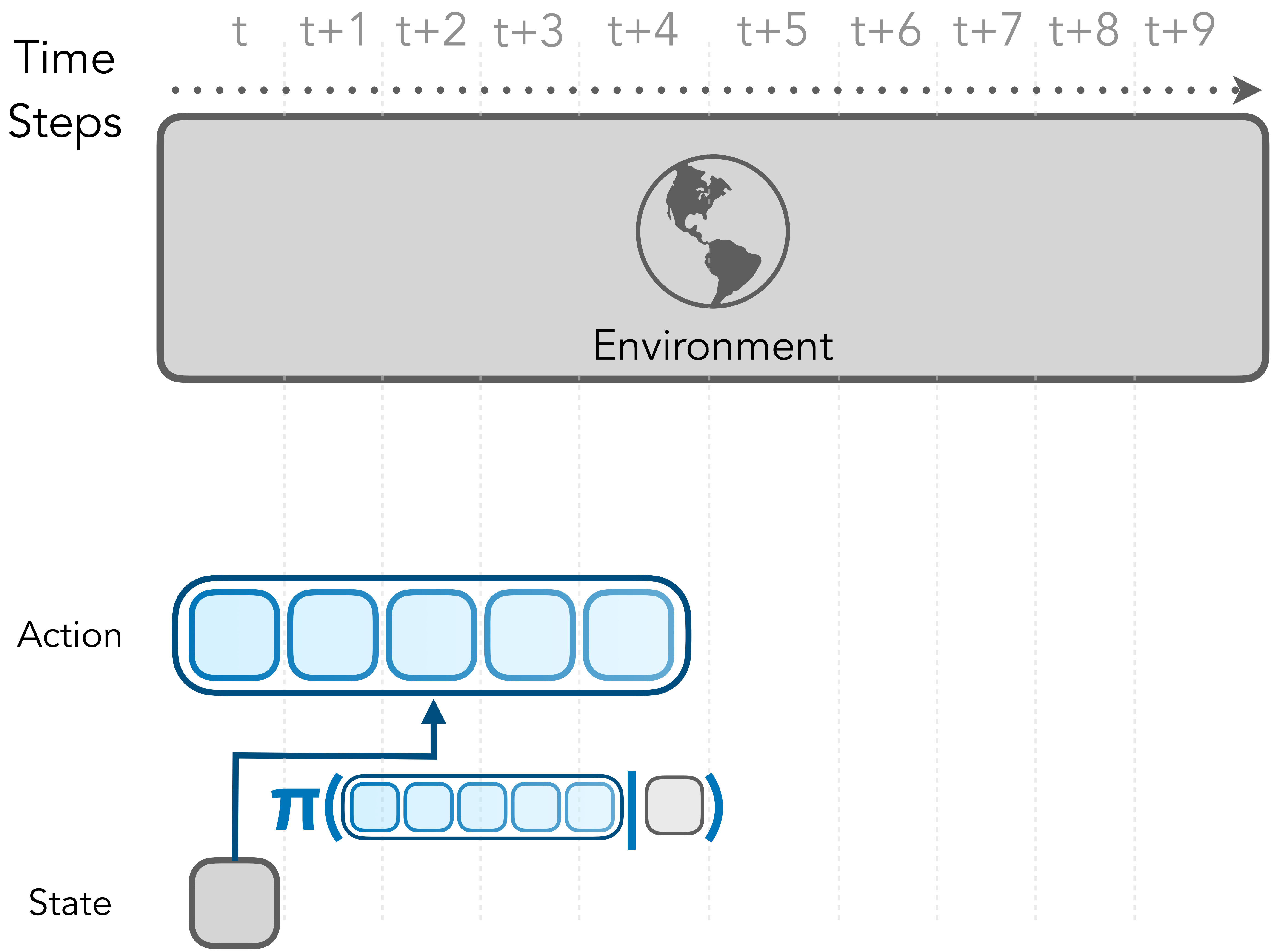
Q-learning with Action Chunking

an action chunk: a sequence of actions



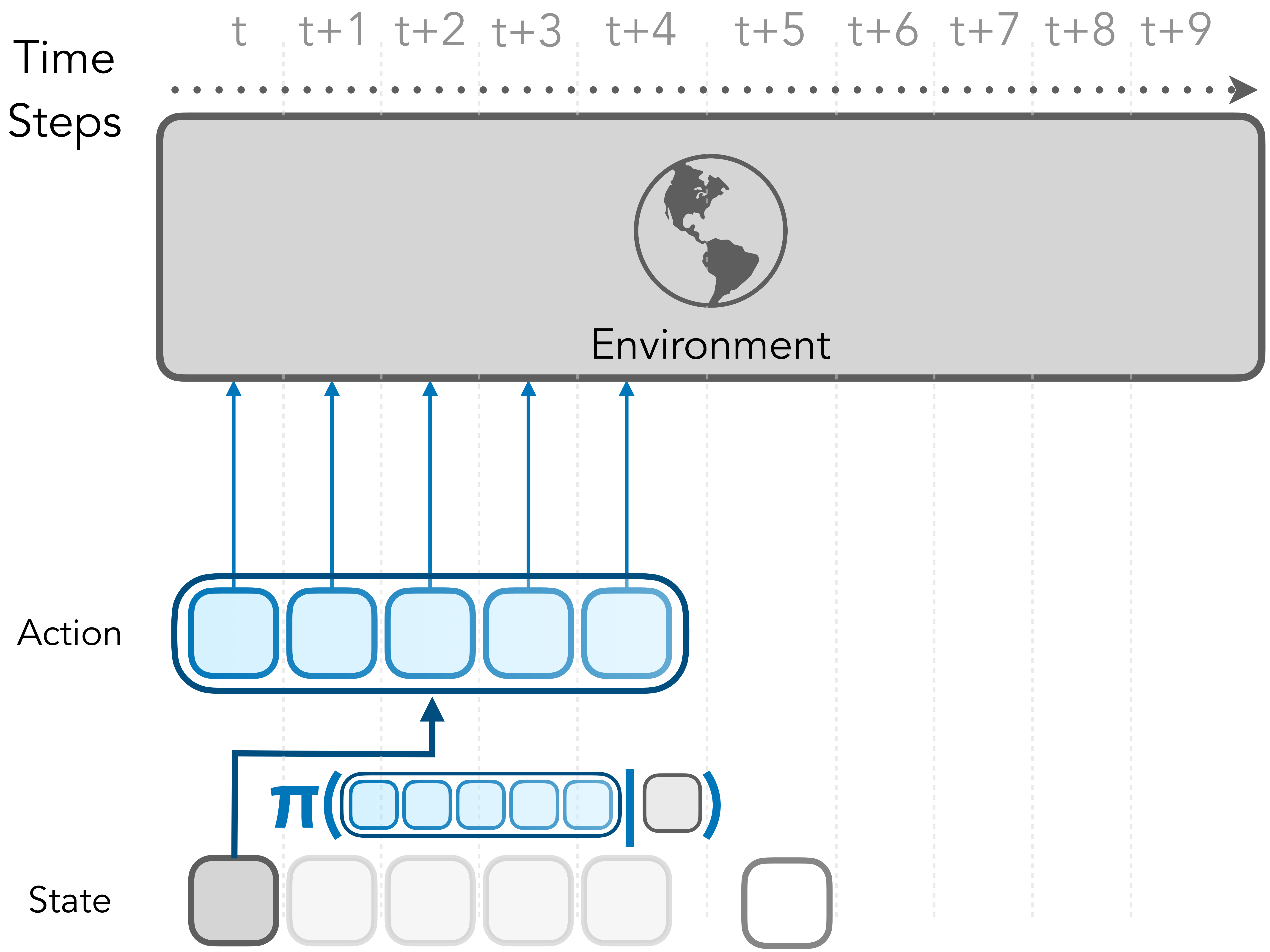
Q-learning with Action Chunking

an action chunk: a sequence of actions



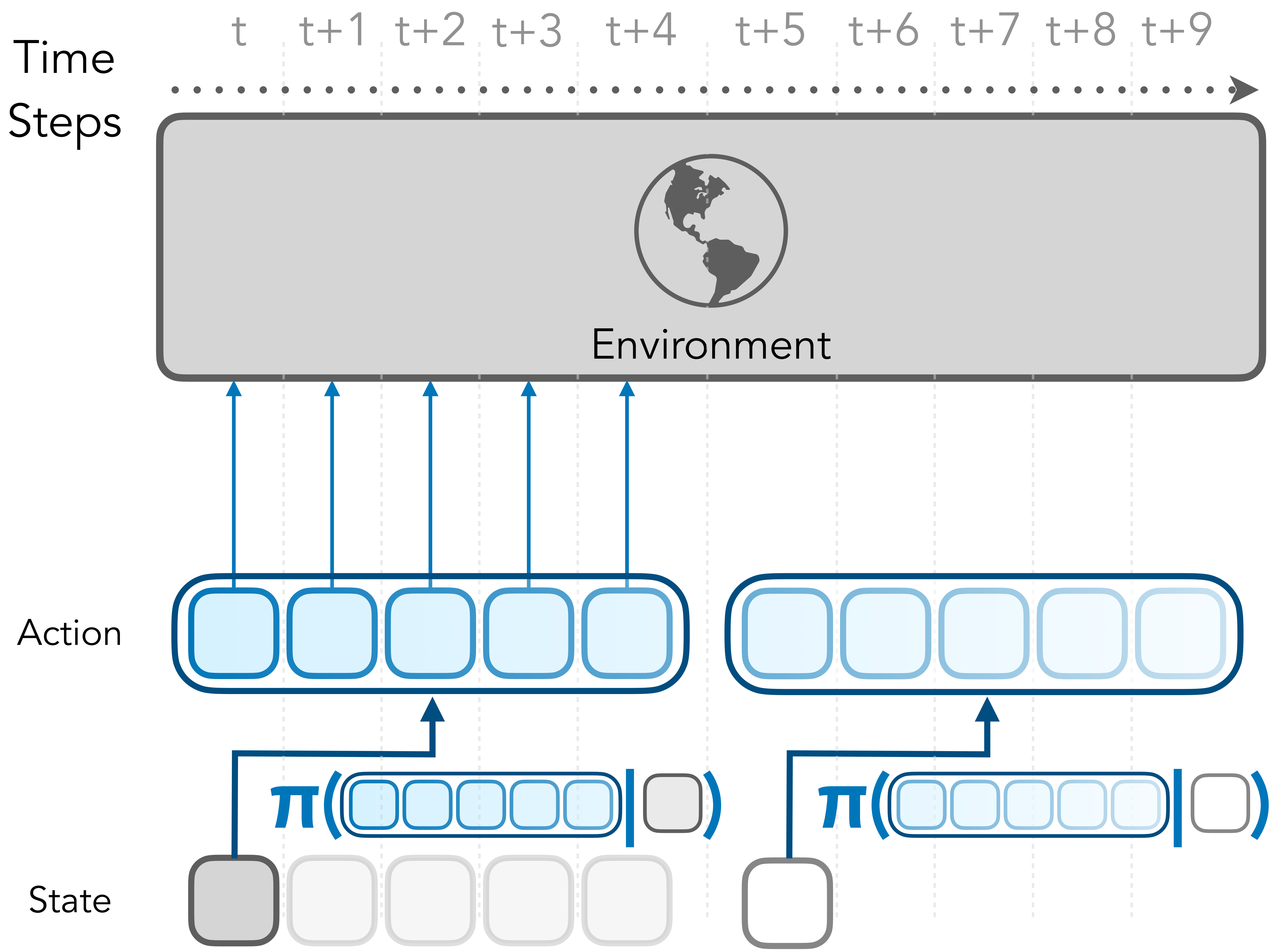
Q-learning with Action Chunking

an action chunk: a sequence of actions



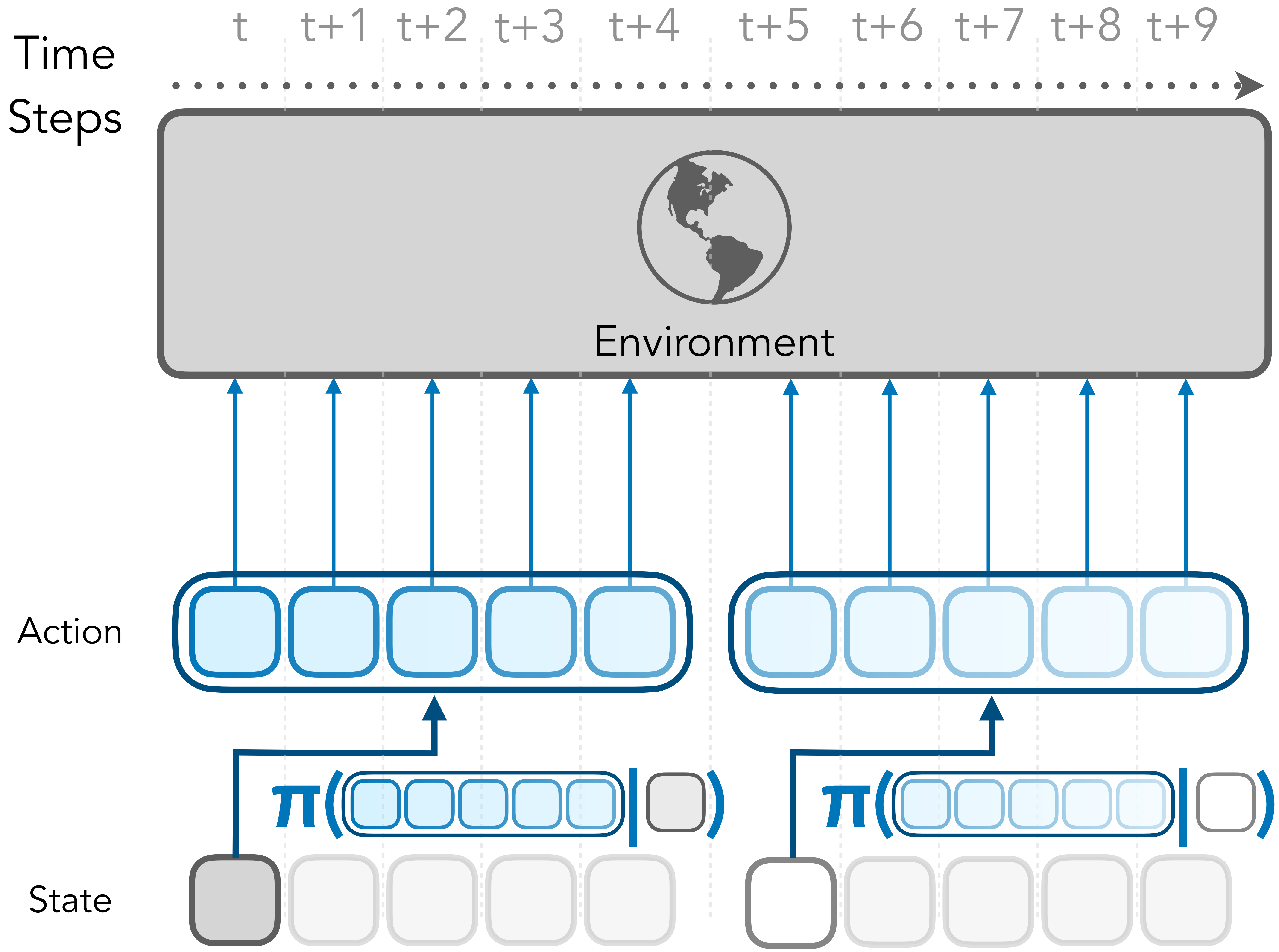
Q-learning with Action Chunking

an action chunk: a sequence of actions



Q-learning with Action Chunking

an action chunk: a sequence of actions



Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

Ingredient #1:

Chunked Critic and Policy

Unbiased **n-step Backup**

1-step TD:

Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

Ingredient #1:

Chunked Critic and Policy

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

Unbiased n-step Backup

Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

$$Q(s_t, a_t) \leftarrow \sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}} + \gamma^h Q(s_{t+h}, a_{t+h} \sim \pi(s_{t+h}))$$

Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

$$Q(s_t, a_t) \leftarrow \underbrace{\sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}}}_{\textbf{biased}} + \gamma^h Q(s_{t+h}, a_{t+h} \sim \pi(s_{t+h}))$$

Ingredient #1:

Chunked Critic
and Policy

Unbiased
n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

$$Q(s_t, a_t) \leftarrow \underbrace{\sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}}}_{\text{biased}} + \gamma^h Q(s_{t+h}, a_{t+h} \sim \pi(s_{t+h}))$$

Q-chunking:

Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

$$Q(s_t, a_t) \leftarrow \underbrace{\sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}}}_{\textbf{biased}} + \gamma^h Q(s_{t+h}, a_{t+h} \sim \pi(s_{t+h}))$$

Q-chunking:

$$Q(s_t, \mathbf{a}_{t:t+h}) \leftarrow \sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}} + \gamma^h Q(s_{t+h}, \mathbf{a}_{t+h:t+2h} \sim \pi(s_{t+h}))$$

Ingredient #1:

Chunked Critic and Policy

Unbiased

n-step Backup

1-step TD:

$$Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a_{t+1} \sim \pi(s_{t+1}))$$

n-step return:

$$Q(s_t, a_t) \leftarrow \underbrace{\sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}}}_{\textbf{biased}} + \gamma^h Q(s_{t+h}, a_{t+h} \sim \pi(s_{t+h}))$$

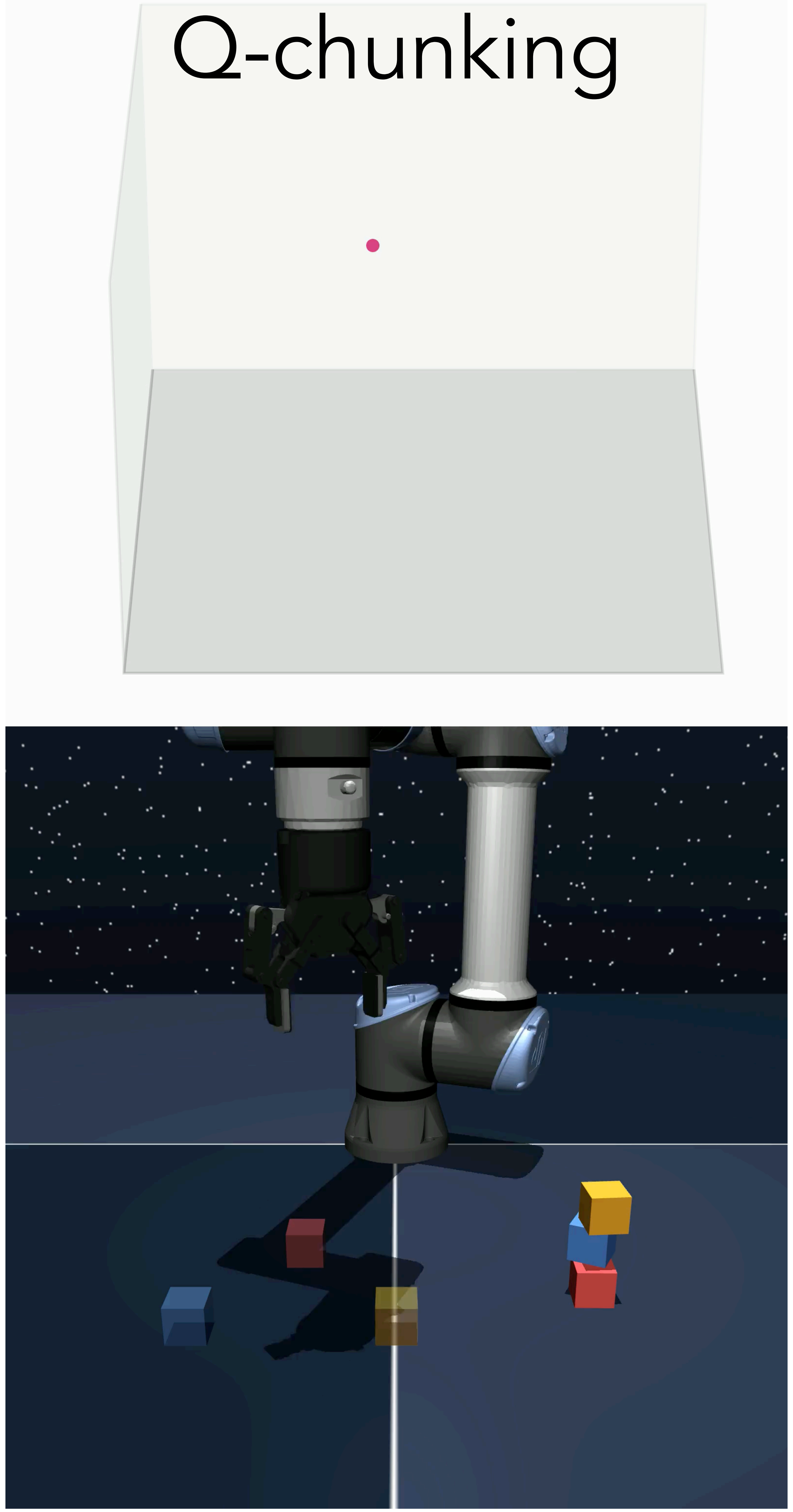
Q-chunking:

$$Q(s_t, \mathbf{a}_{t:t+h}) \leftarrow \underbrace{\sum_{\bar{t}=t}^{t+h-1} \gamma^{\bar{t}-t} r_{\bar{t}}}_{\textbf{unbiased}} + \gamma^h Q(s_{t+h}, \mathbf{a}_{t+h:t+2h} \sim \pi(s_{t+h}))$$

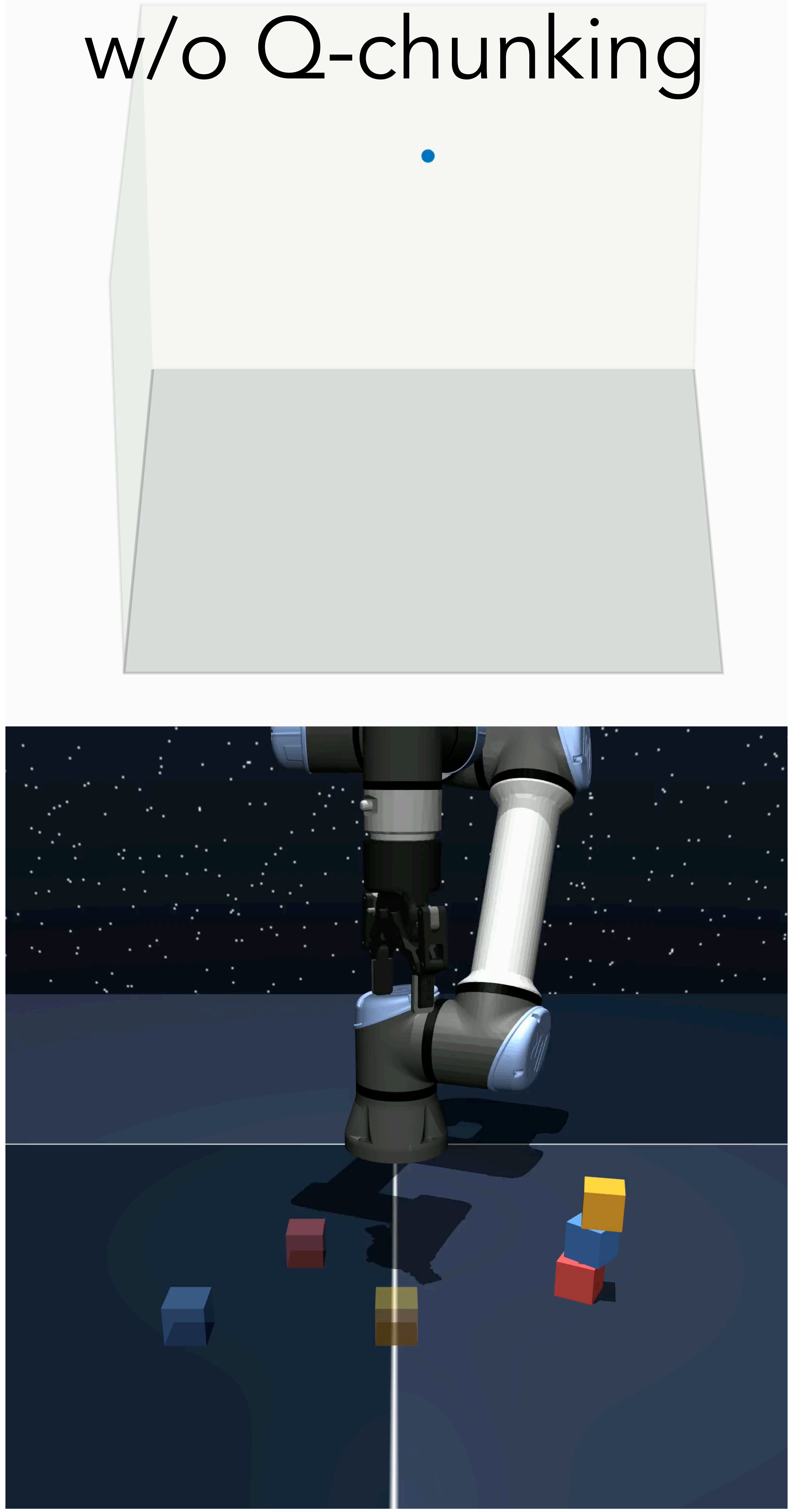
Ingredient #2:

Expressive Behavior Constraint

Better **Temporal Coherency**



Q-chunking

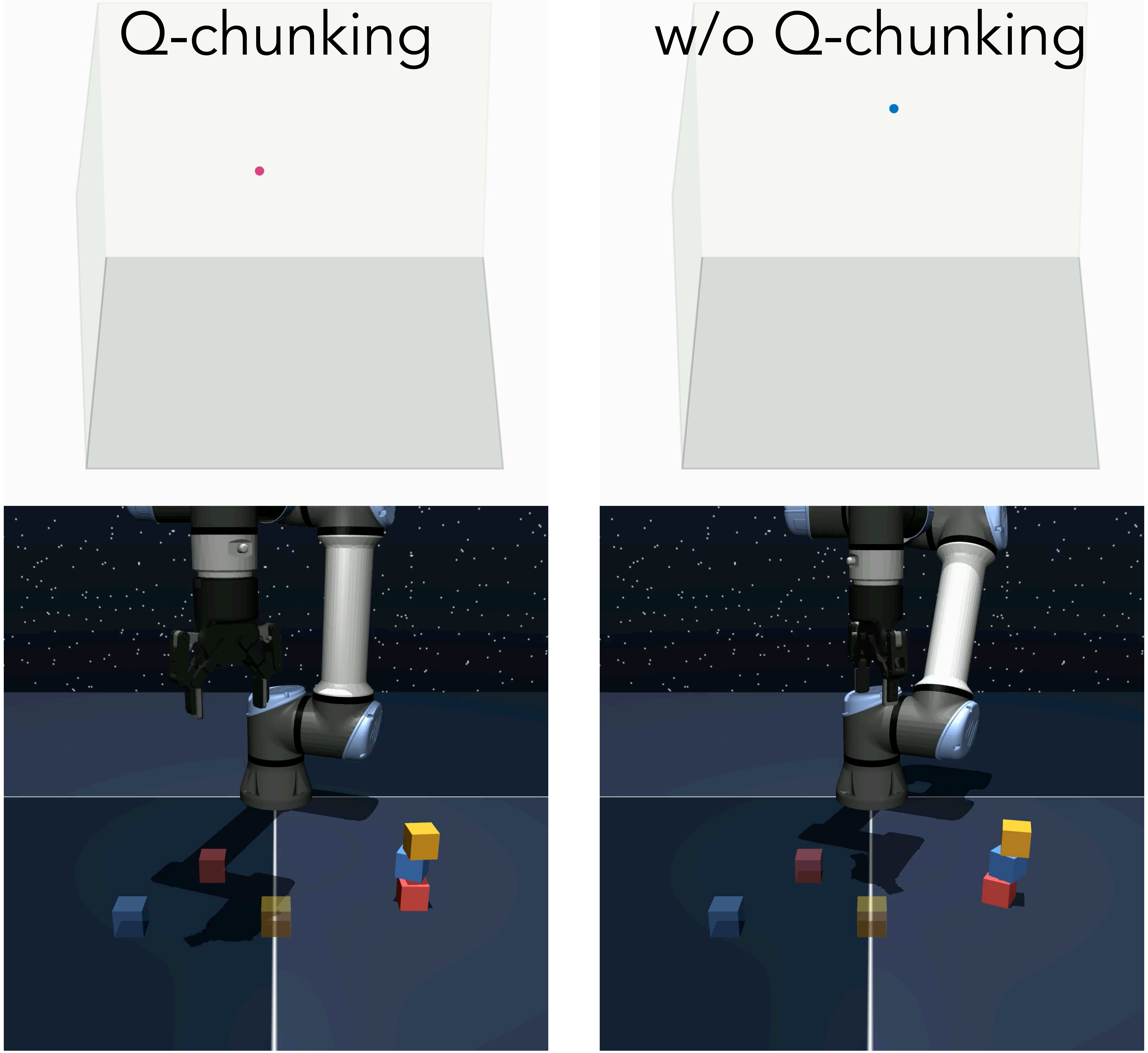


w/o Q-chunking

Ingredient #2:

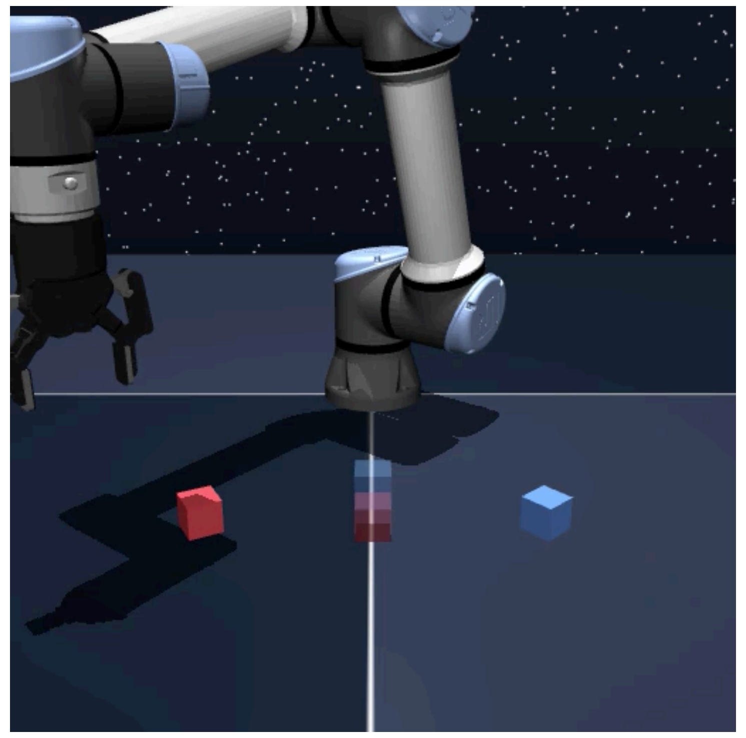
Expressive Behavior Constraint

Better **Temporal Coherency**

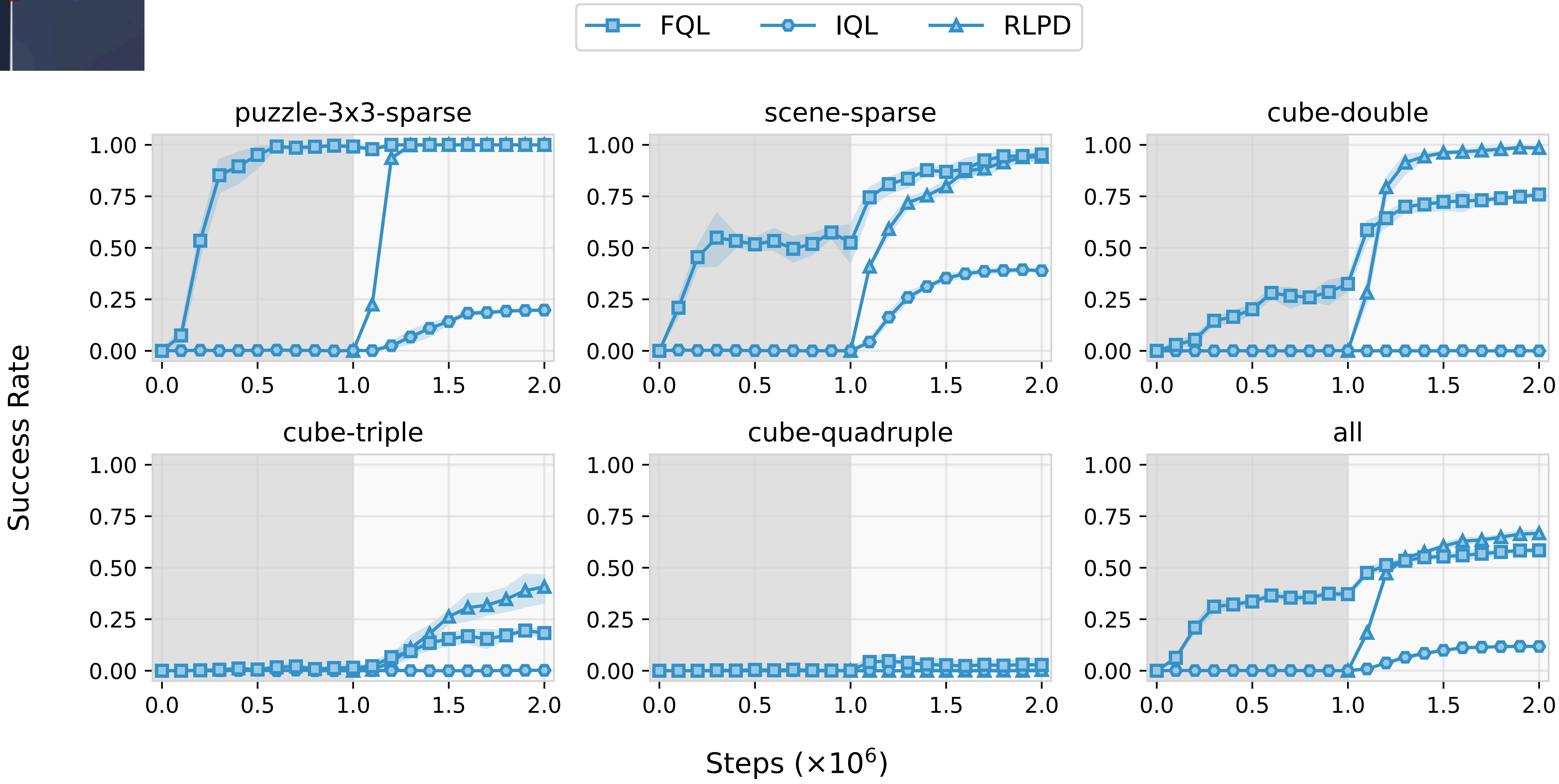


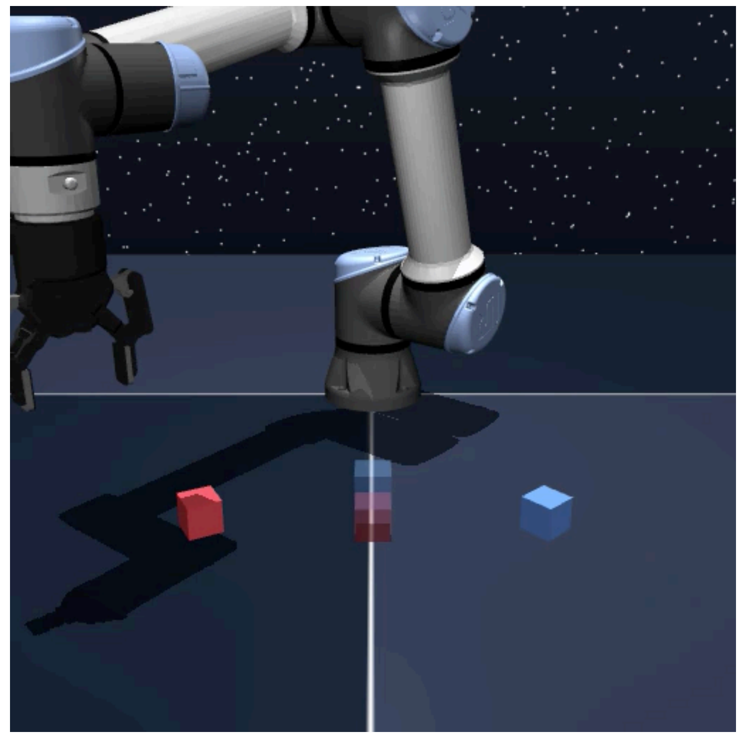
Q-chunking

w/o Q-chunking

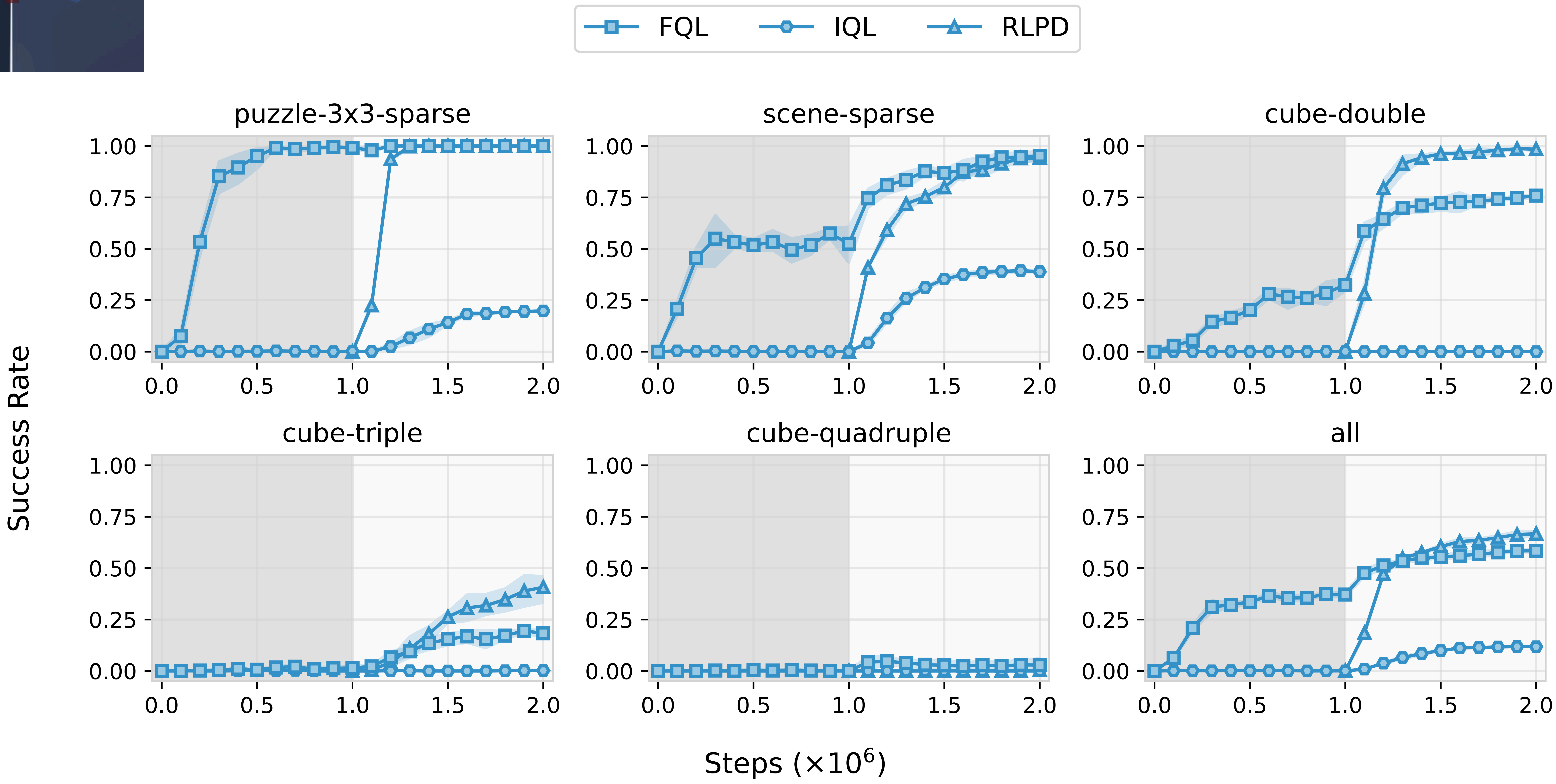


Results

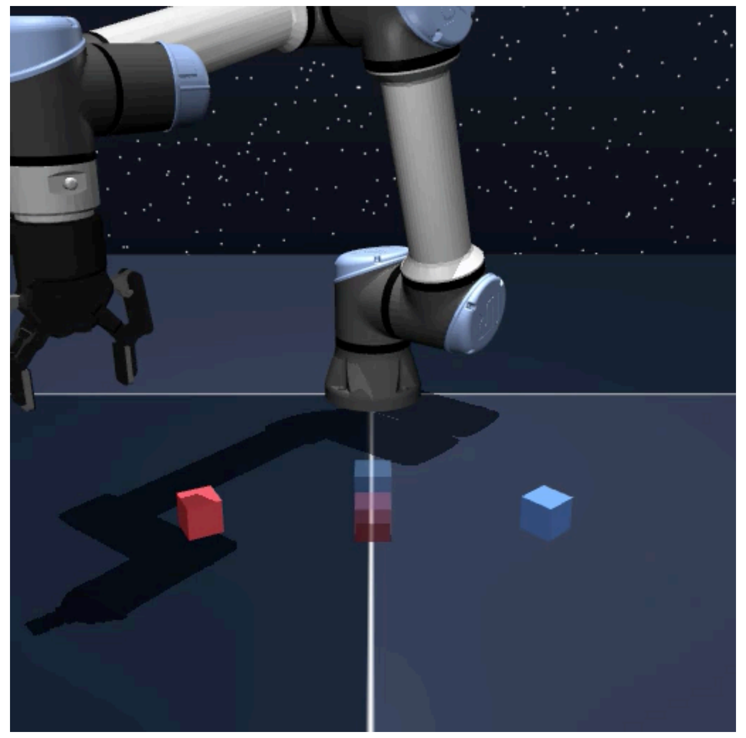

FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
Steps ($\times 10^6$)

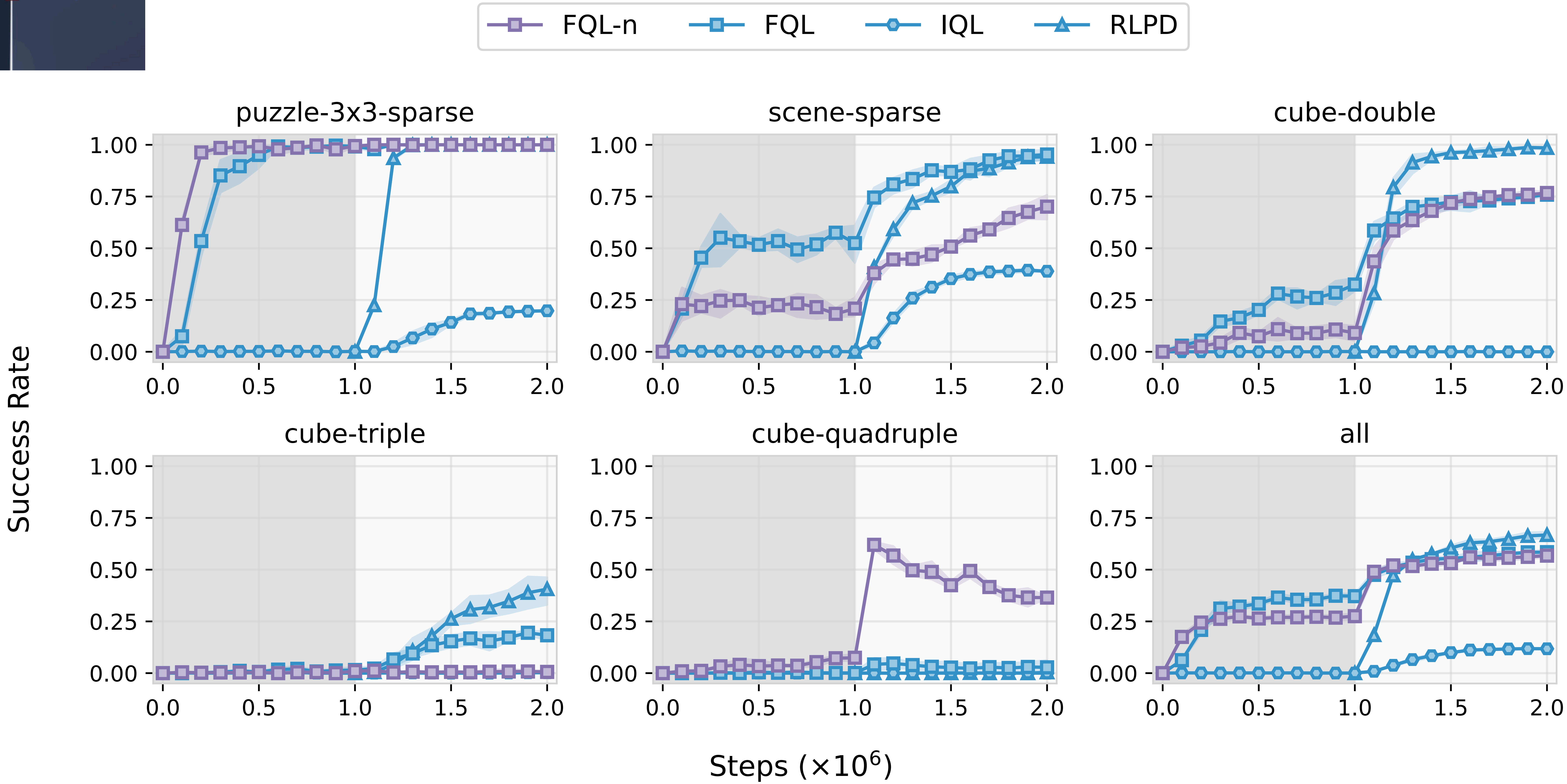


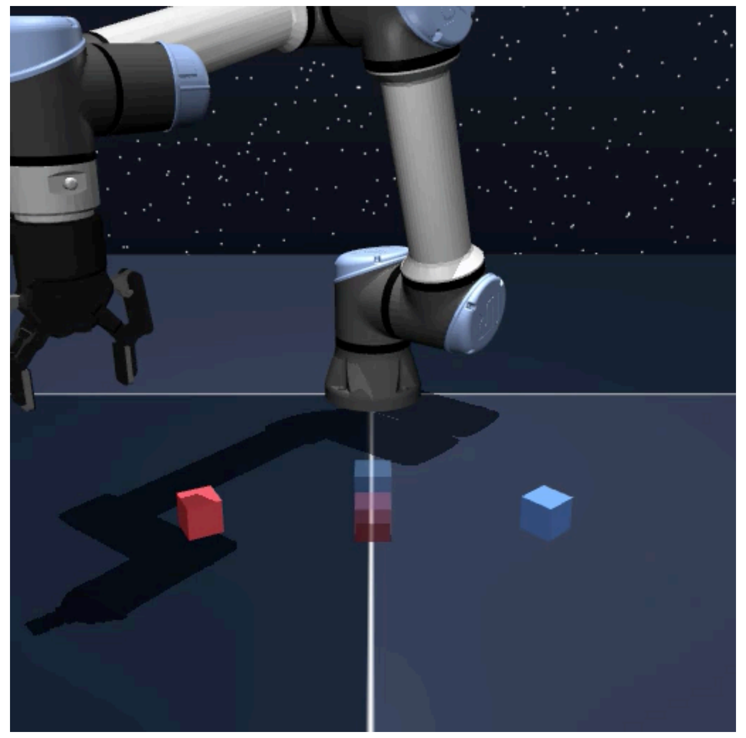
Results


FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
Steps (×10^6)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0
1.5
2.0

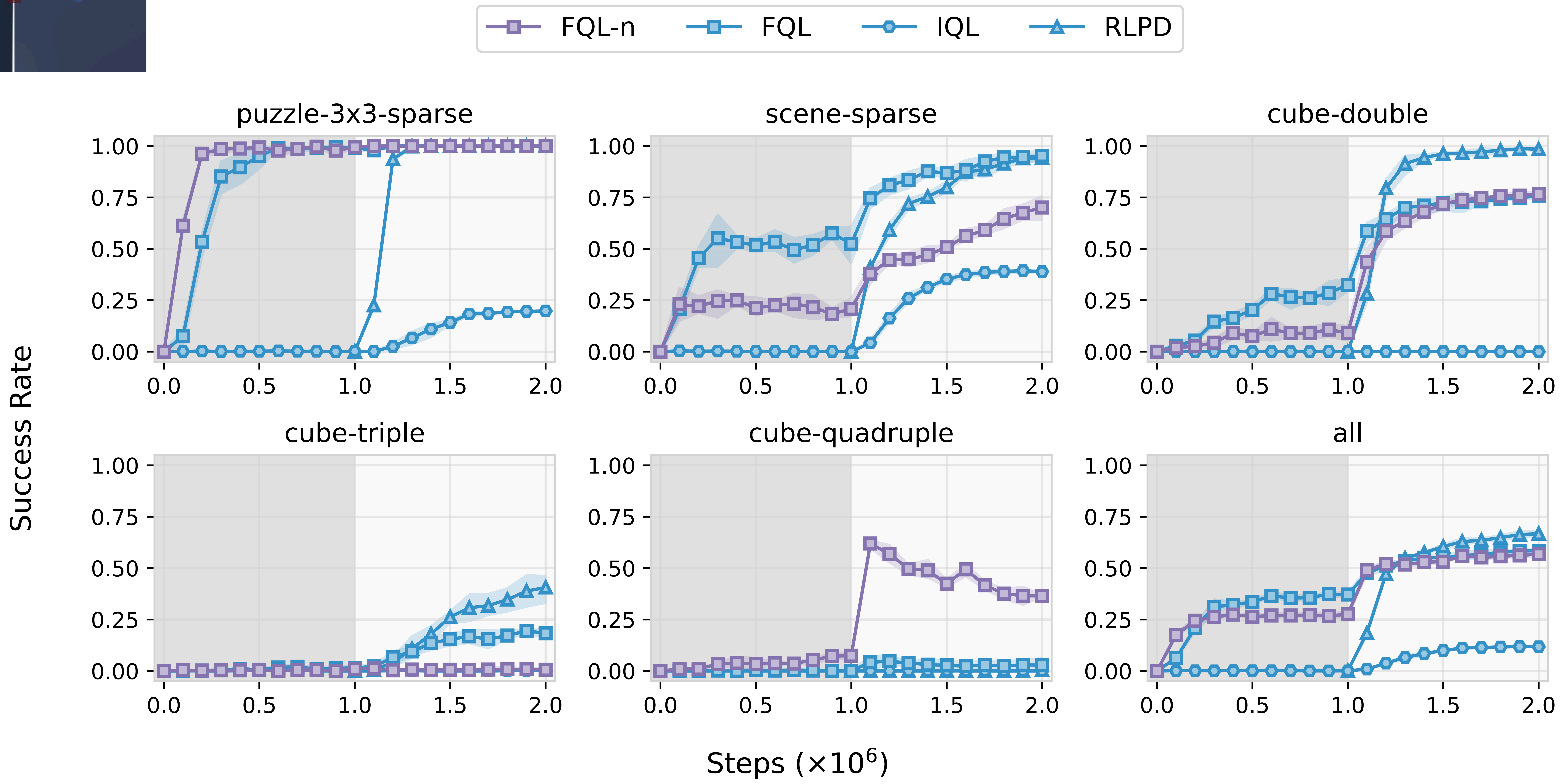
Results



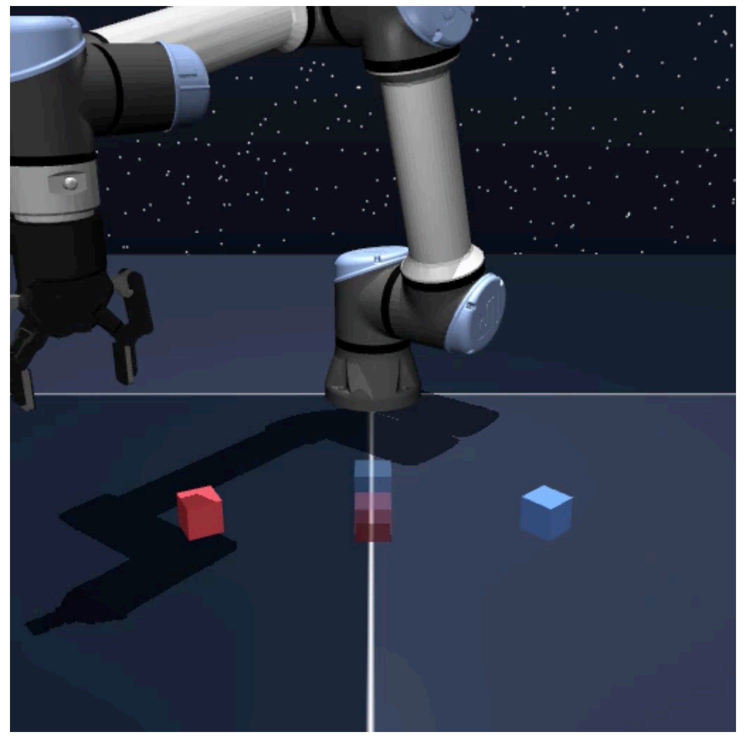

FQL-n
FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0
1.5
2.0
Steps (×10^6)

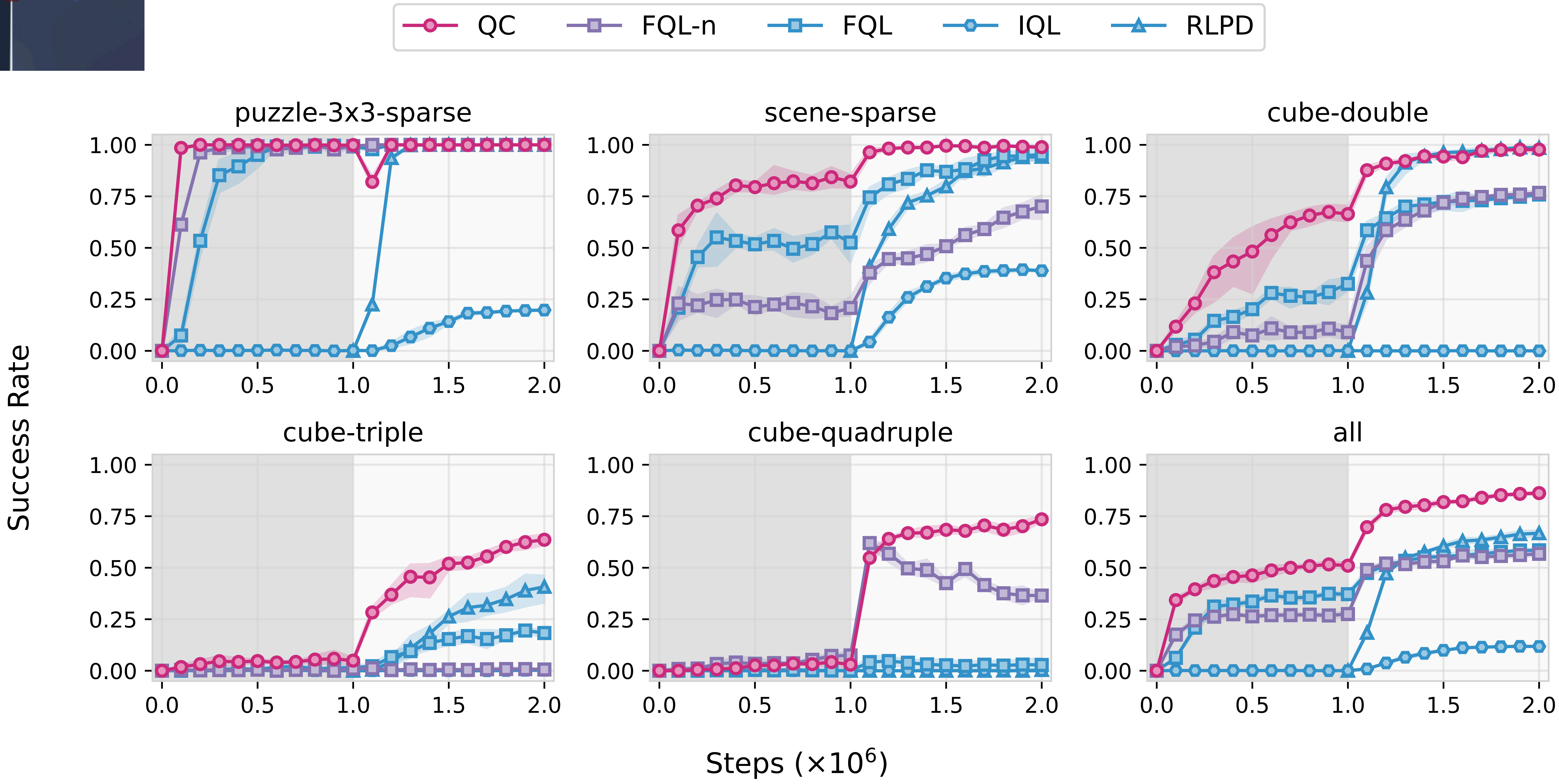


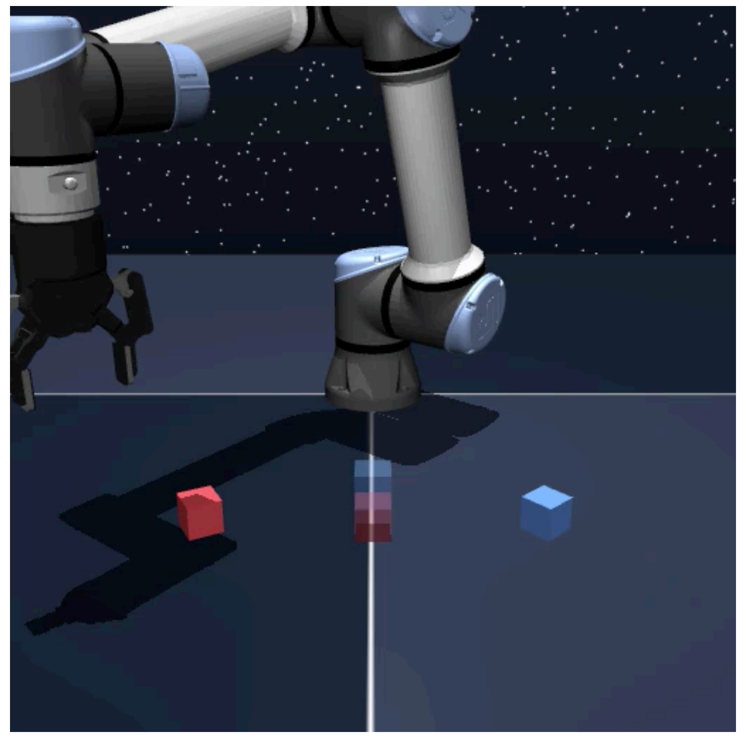
Results


FQL-n
FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
Steps (×10^6)

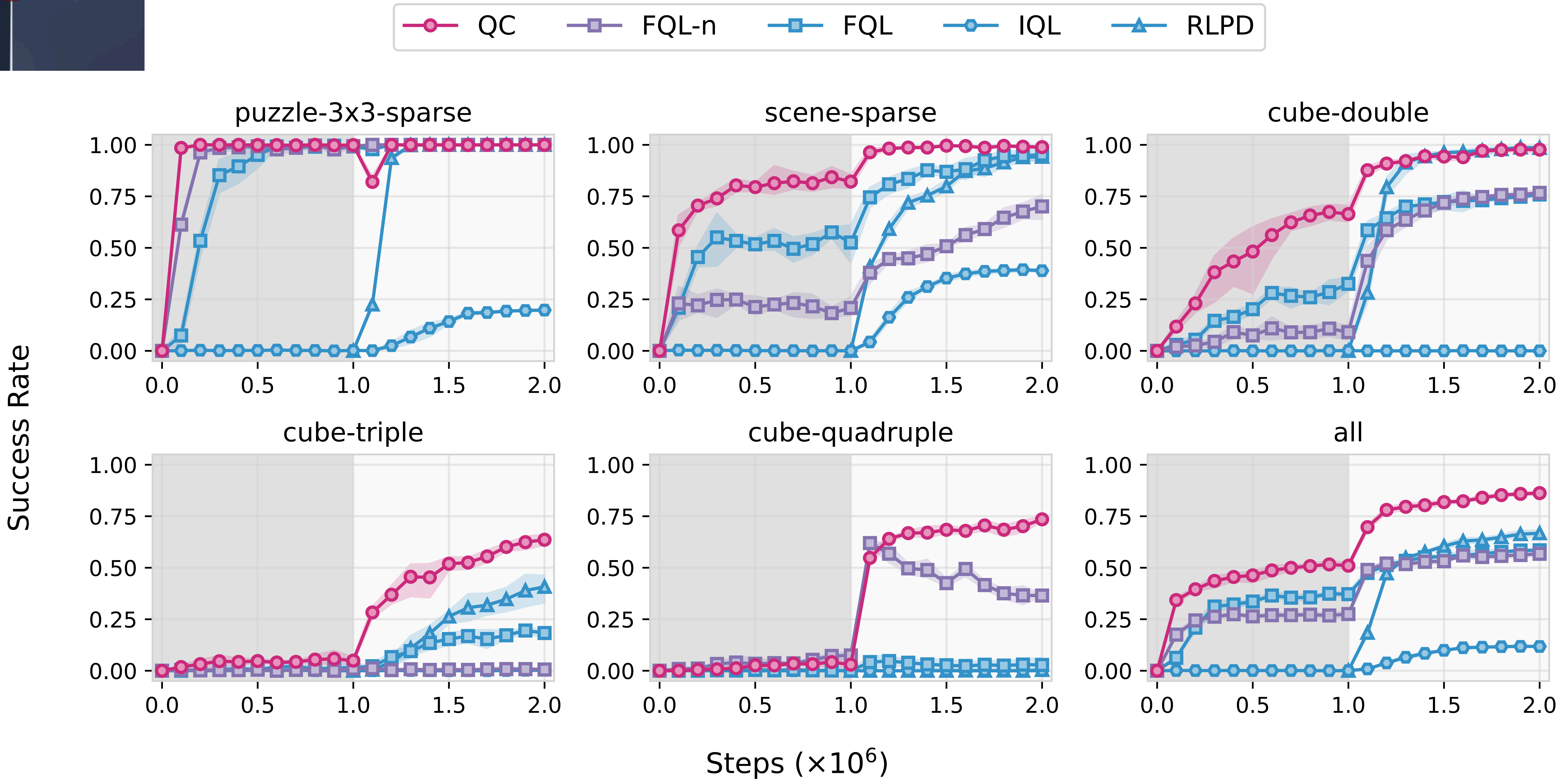
Results




QC
FQL-n
FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
Steps (×10^6)
0.00
0.25
0.50
0.75
1.00
0.0
0.5
1.0
1.5
2.0



Results

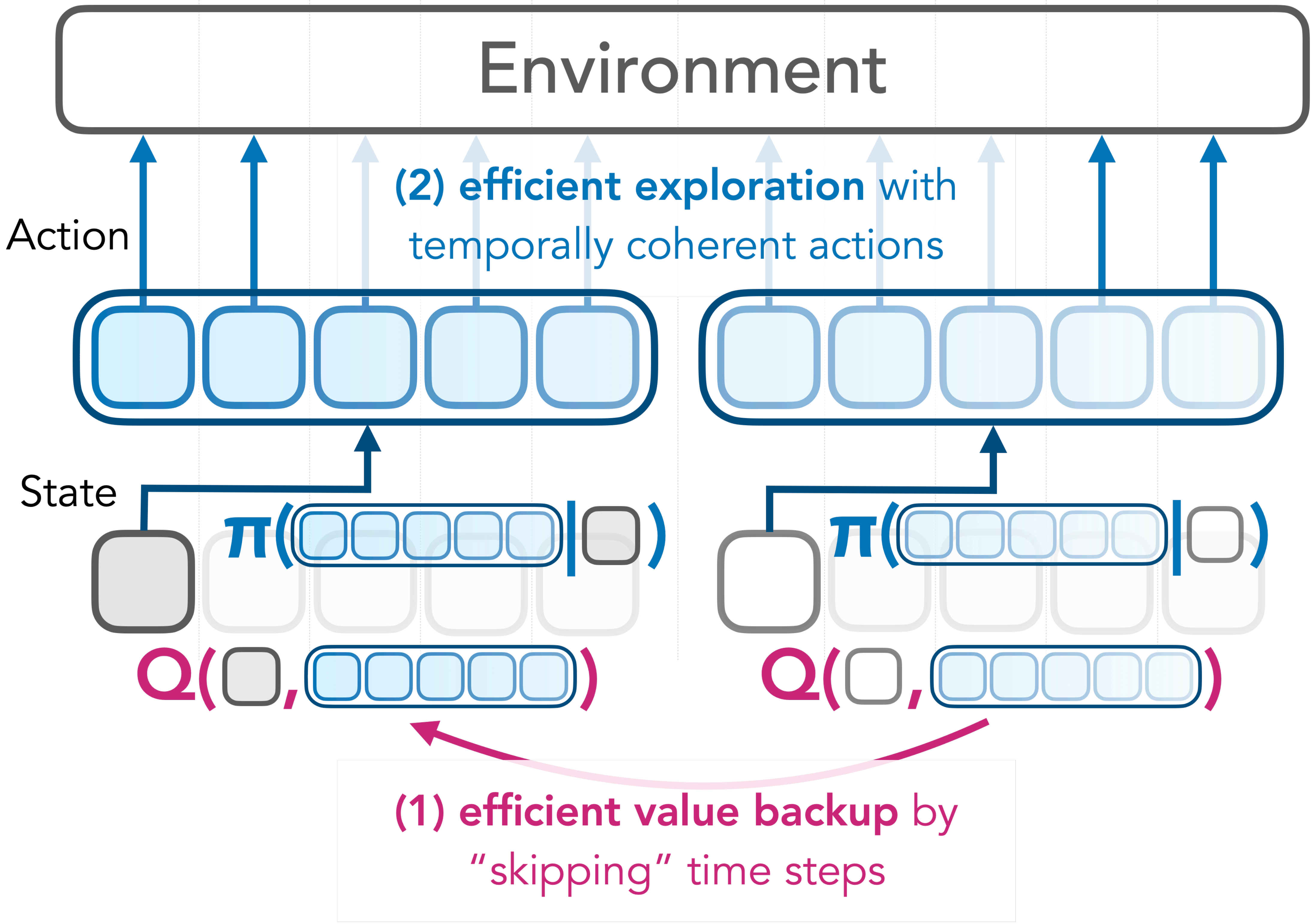


QC
FQL-n
FQL
IQL
RLPD
puzzle-3x3-sparse
scene-sparse
cube-double
cube-triple
cube-quadruple
all
Success Rate
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
Steps (×10^6)

Summary

Q: How to speedup **offline-to-online** RL on manipulation tasks?

A: Apply **action chunking** to both policy and critic and use an **expressive** policy with **BC constraint**.



Environment
Action
State
(2) efficient exploration with temporally coherent actions
π(
)
Q(
)
(1) efficient value backup by "skipping" time steps

Thank you!!

Zhiyuan (Paul) Zhou

Sergey Levine

Code: github.com/ColinQiyangLi/qc

website arXiv code