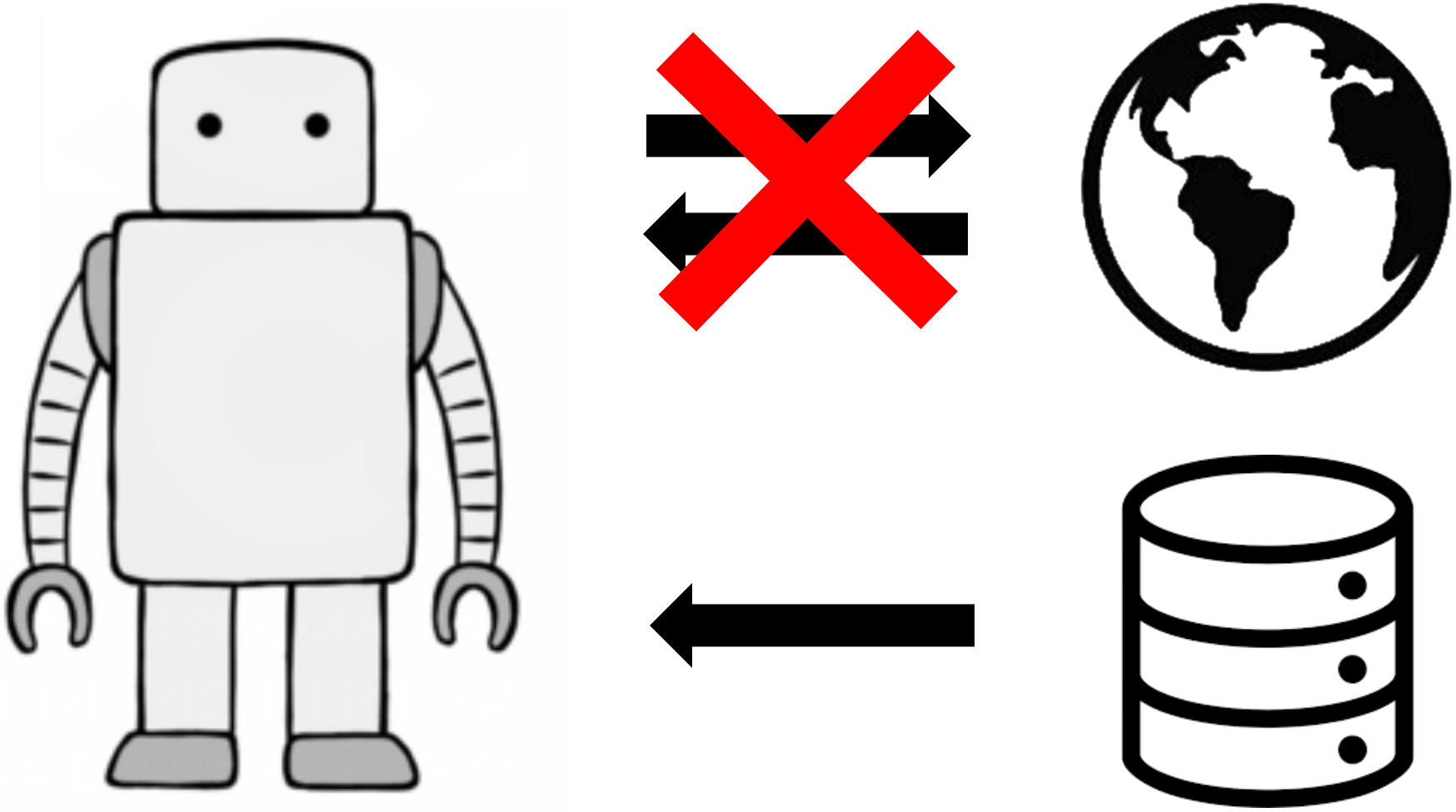
Model-based Offline Reinforcement Learning with Lower Expectile Q-Learning

Kwanyoung Park, Youngwoon Lee

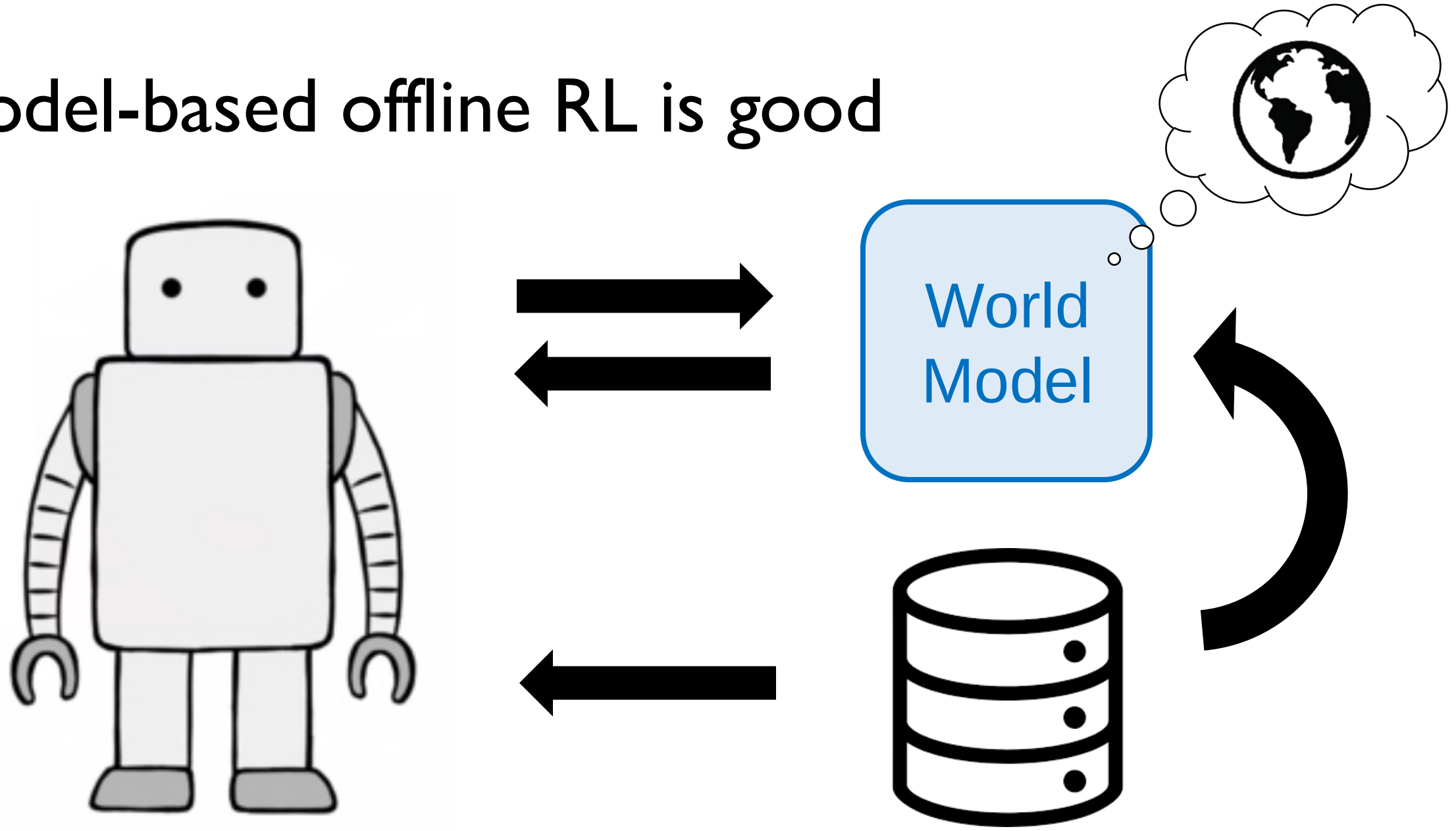
ICLR 2025

https://kwanyoungpark.github.io/LEQ/

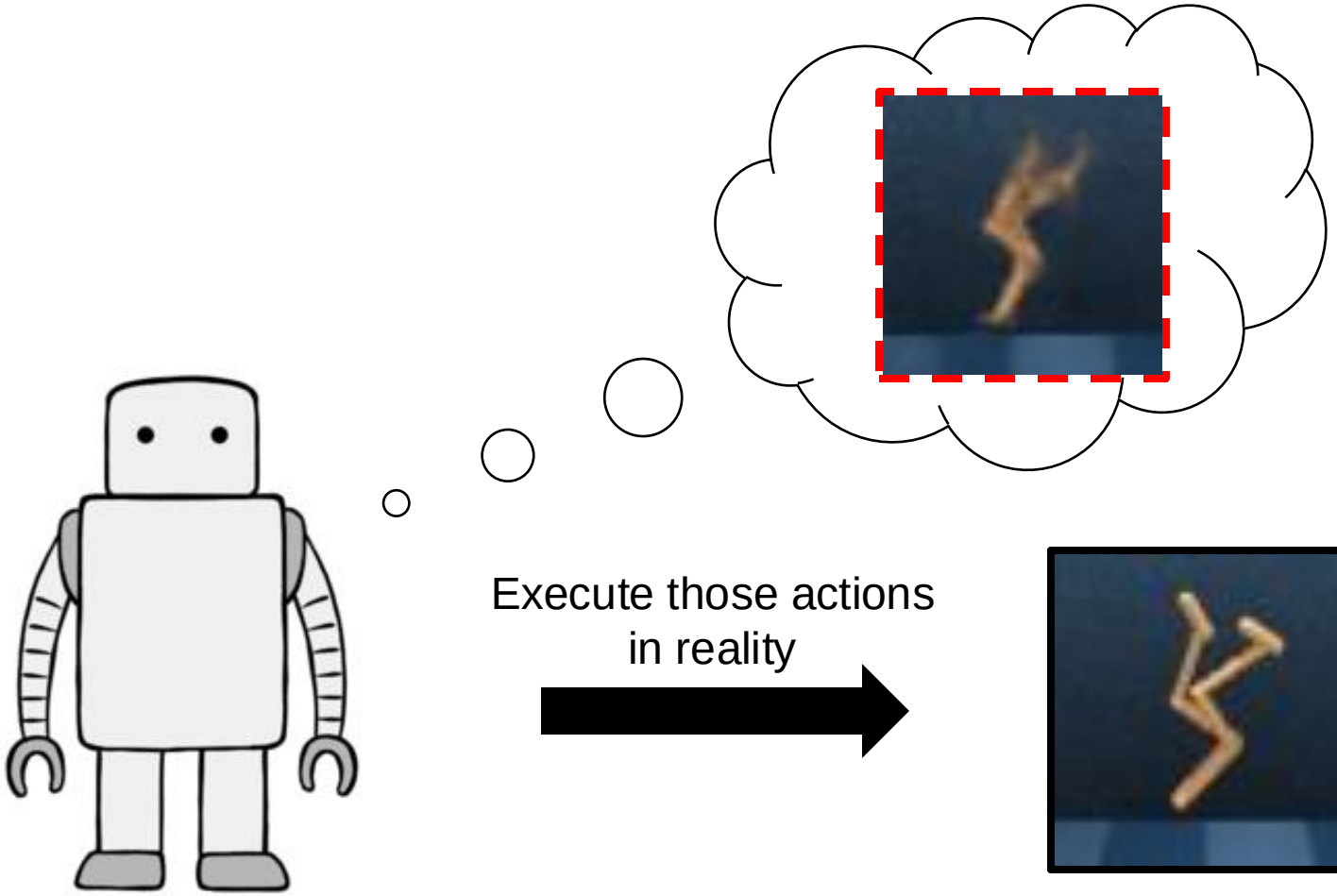
Why offline RL is hard?



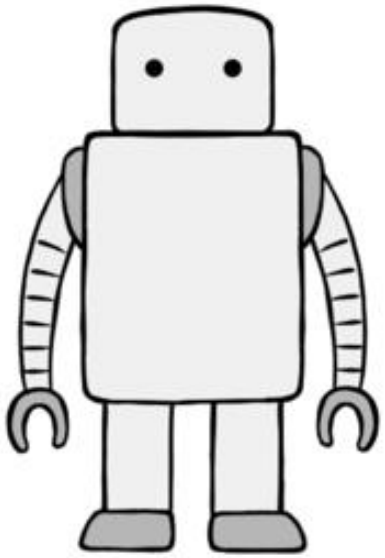
Model-based offline RL is good

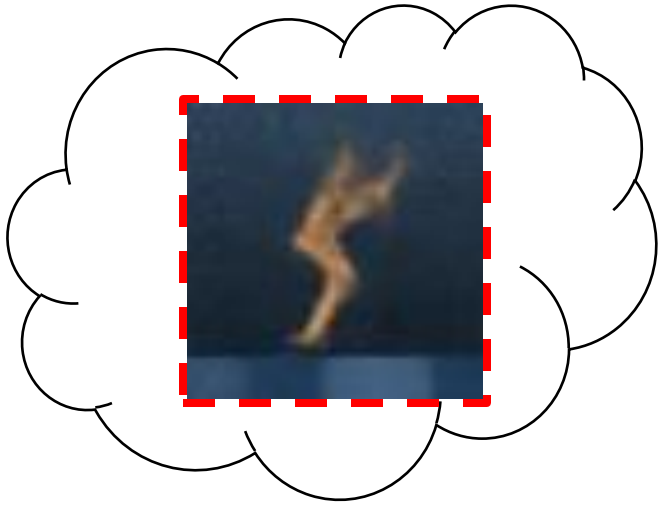


Model can be exploited



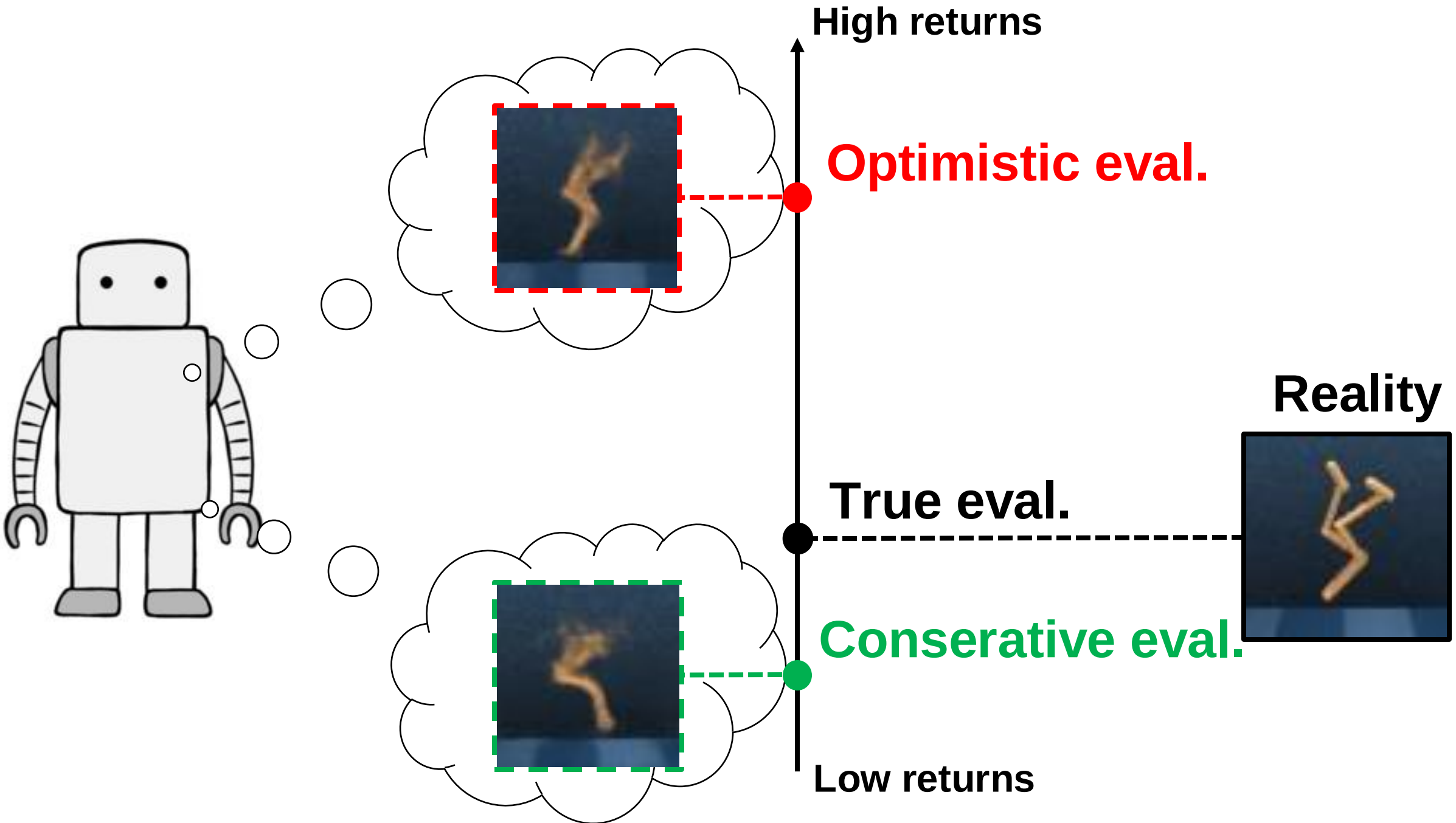
Reality

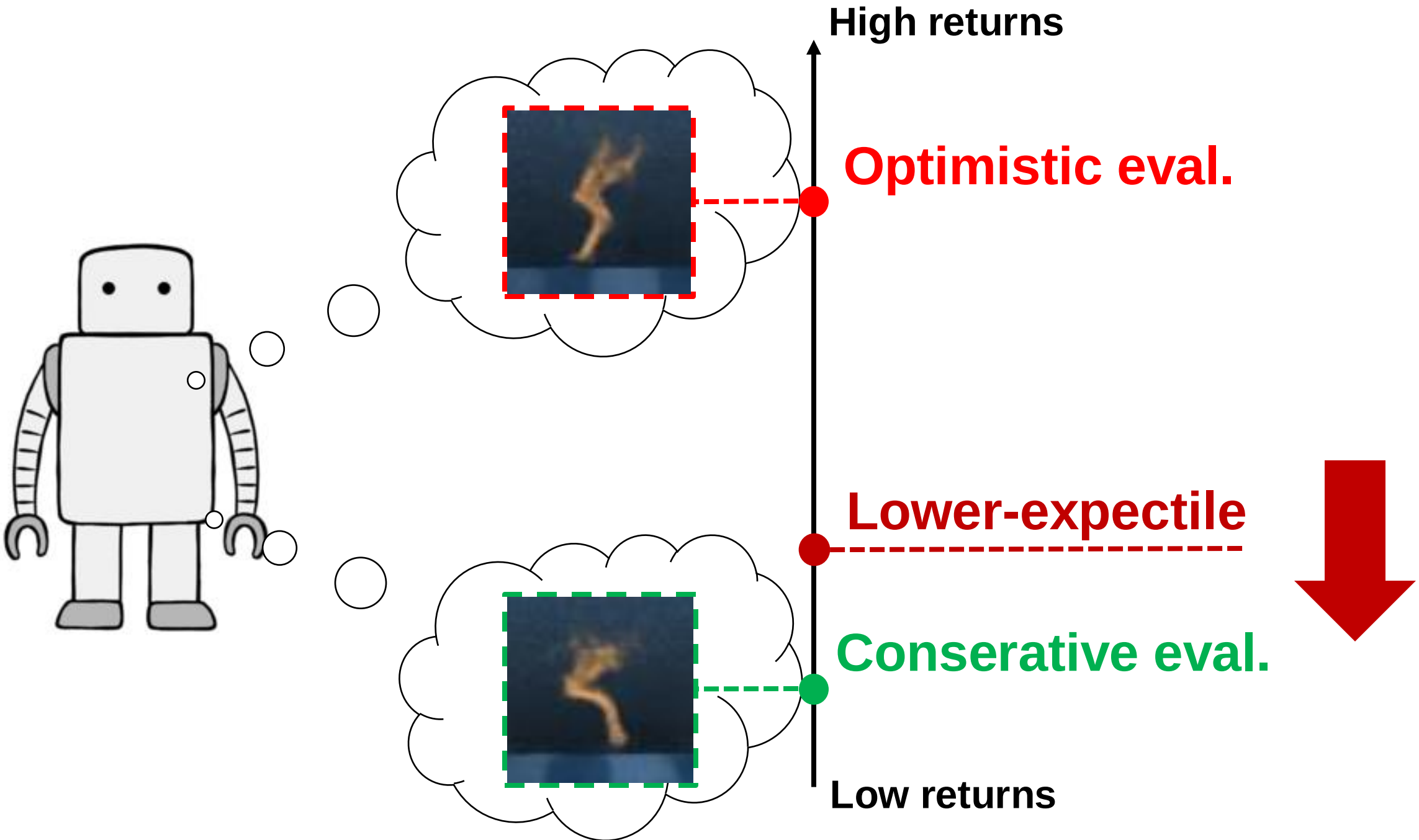




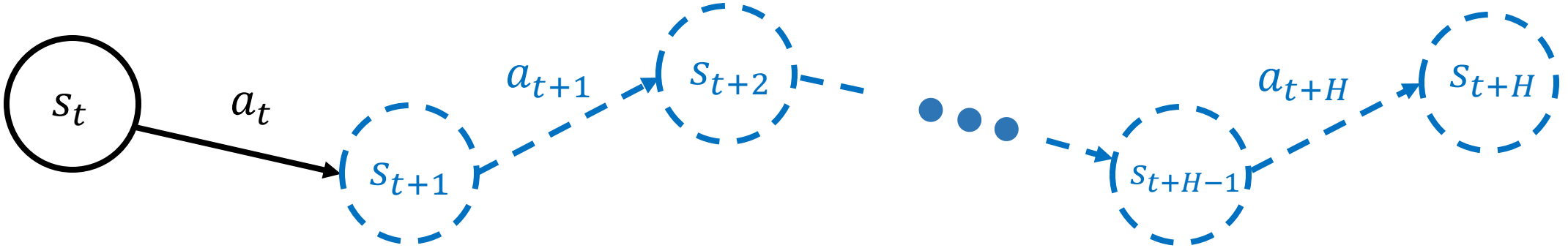
Optimistic

Conservative


High returns
Optimistic eval.
Reality
True eval.
Conserative eval.
Low returns

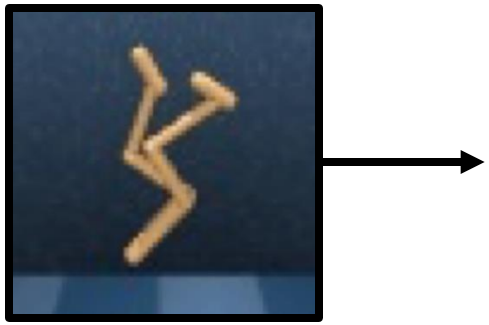

High returns
Optimistic eval.
Lower-expectile
Conserative eval.
Low returns

Model-generated trajectory (Imagination)

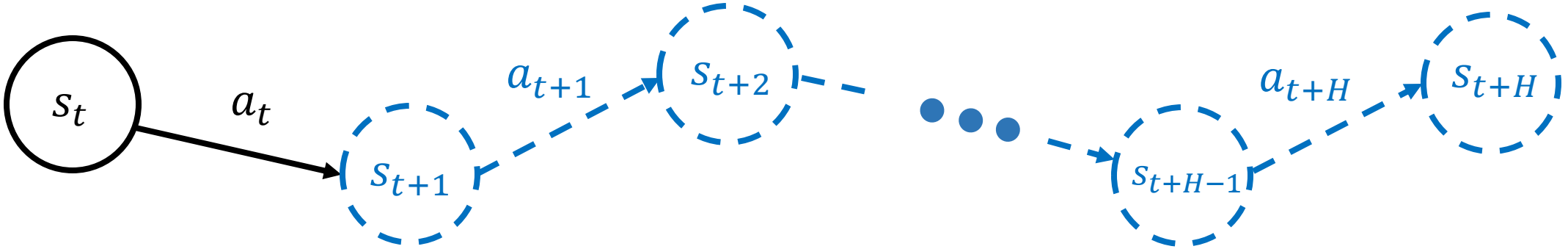


Q-values: $\mathrm{Q}(s_t, a_t)$

Imagined returns: $\hat{Q}(s_{t:t+H}, a_{t:t+H})$



Model-generated trajectory (Imagination)



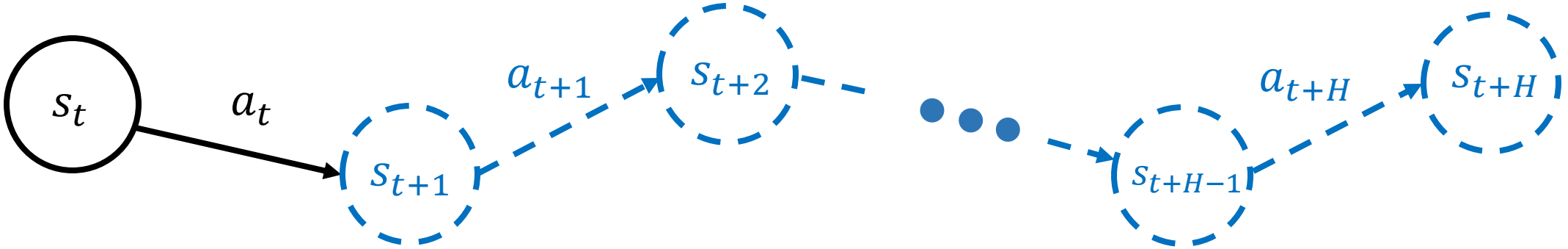
Q-values: $\mathrm{Q}(s_t, a_t)$

Imagined returns: $\hat{Q}(s_{t:t+H}, a_{t:t+H})$

Optimistic

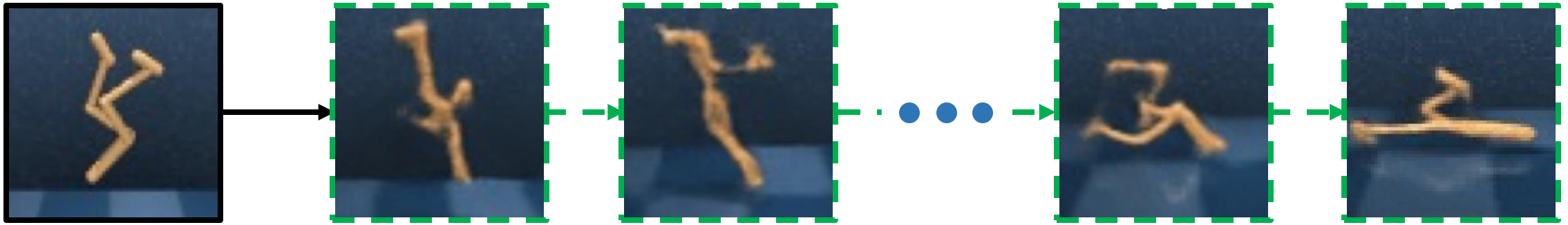
$(Q < \hat{Q})$

Model-generated trajectory (Imagination)



Q-values: $\mathrm{Q}(s_t, a_t)$

Imagined returns: $\hat{Q}(s_{t:t+H}, a_{t:t+H})$



Conservative

$(Q > \hat{Q})$

Lower-Expectile Q-Learning (LEQ)

Optimistic

Conservative

$$0.1 \cdot \underbrace{\left(Q_\phi(s_t, a_t) - \hat{Q}\right)^2}_{\text{Bellman loss}}$$

$$0.9 \cdot \underbrace{\left(Q_\phi(s_t, a_t) - \hat{Q}\right)^2}_{\text{Bellman loss}}$$

Lower-Expectile Q-Learning (LEQ)

Optimistic

$$0.1 \cdot \underbrace{\nabla_\theta \hat{Q}(s_{t:t+H}, a_{t:t+H})}_{\text{Policy Gradient}}$$

Conservative

$$0.9 \cdot \underbrace{\nabla_\theta \hat{Q}(s_{t:t+H}, a_{t:t+H})}_{\text{Policy Gradient}}$$

Lower-Expectile Q-Learning (LEQ)

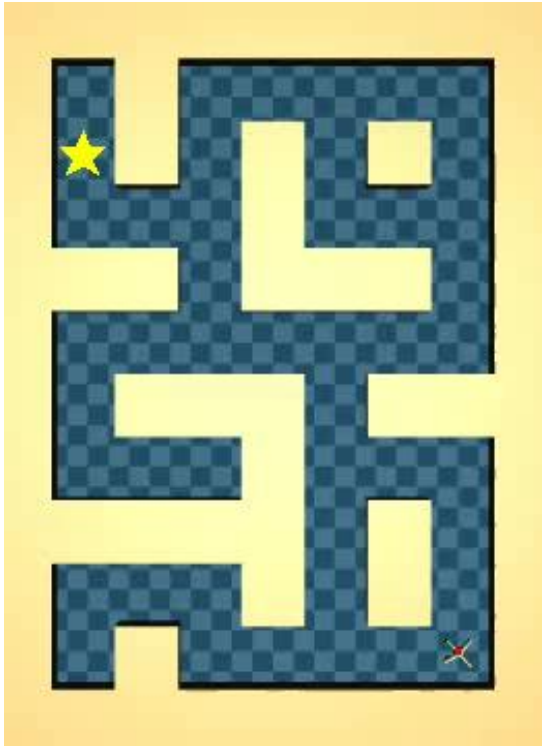
Optimistic → $0.1 \cdot$Bellman loss + $0.1 \cdot$Policy gradient

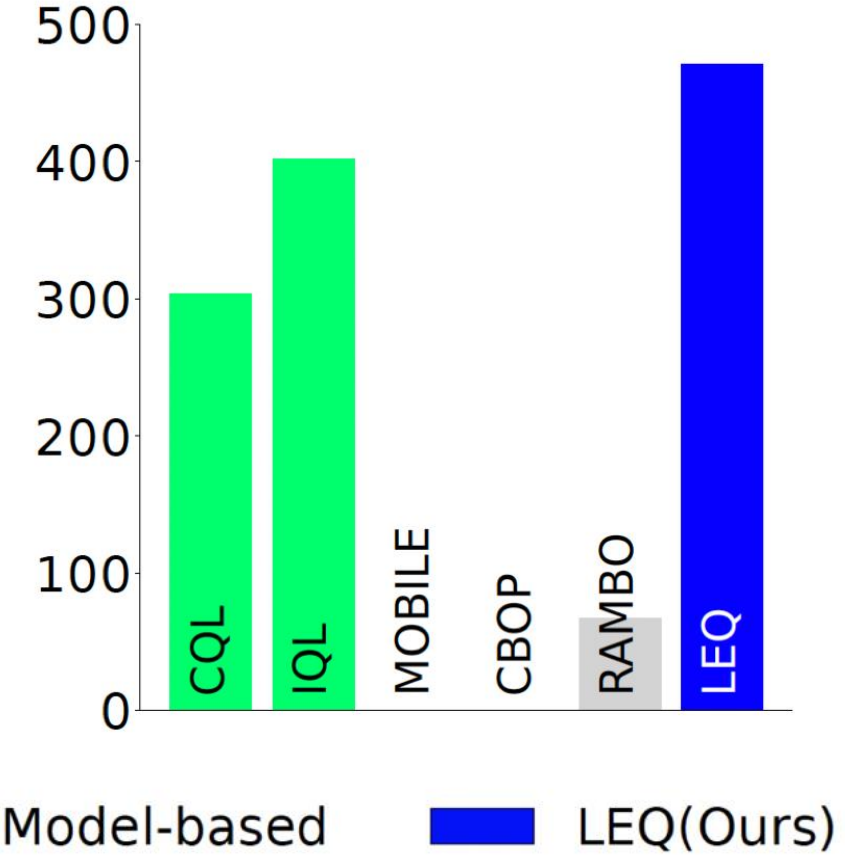
Conservative → $0.9 \cdot$Bellman loss + $0.9 \cdot$Policy gradient

$Q_\phi(s_t, a_t) =$ Lower-expectile

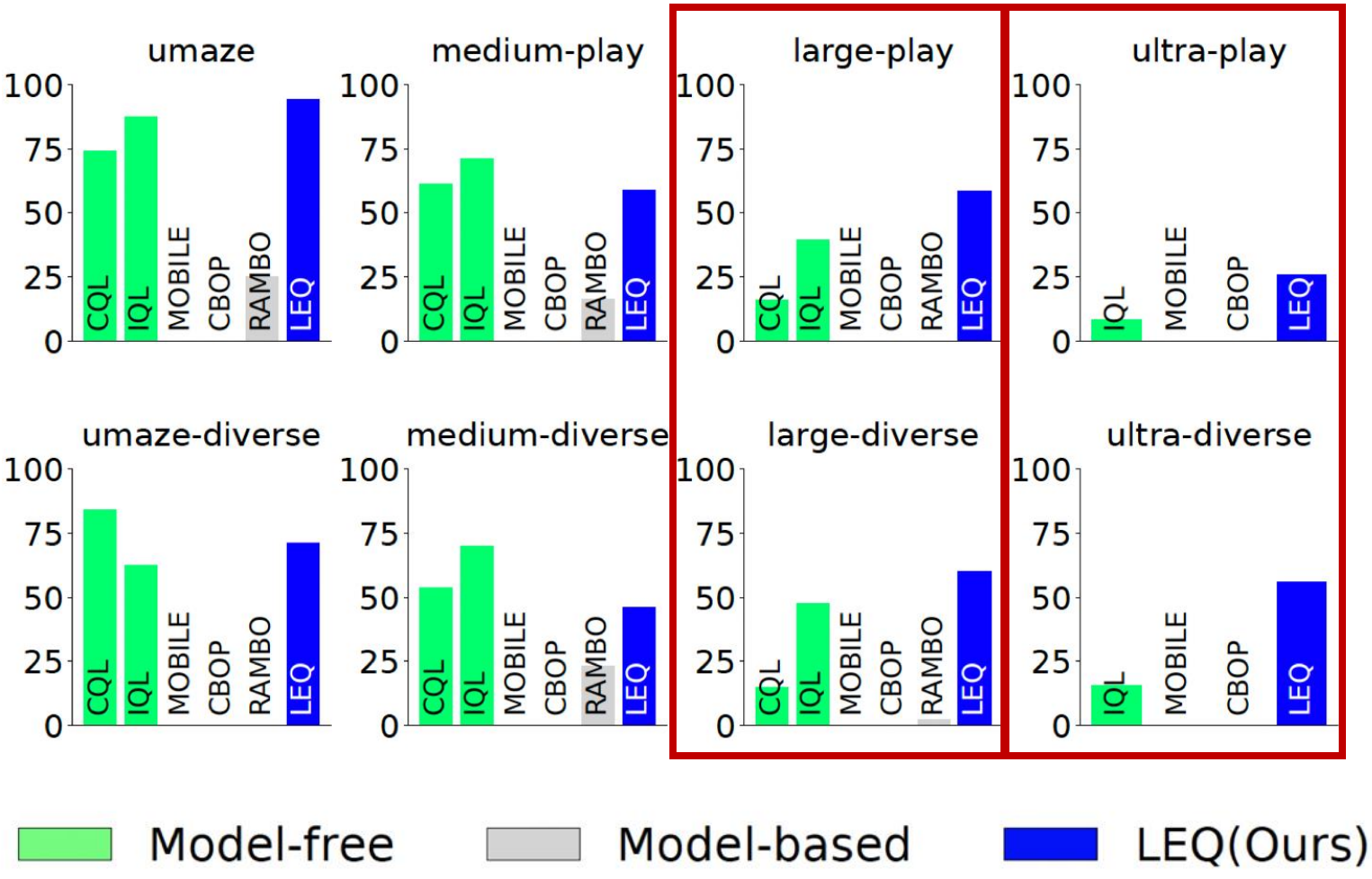
$\nabla_\theta L_\pi(\theta) \approx \nabla_\theta$(Lower-expectile)

AntMaze

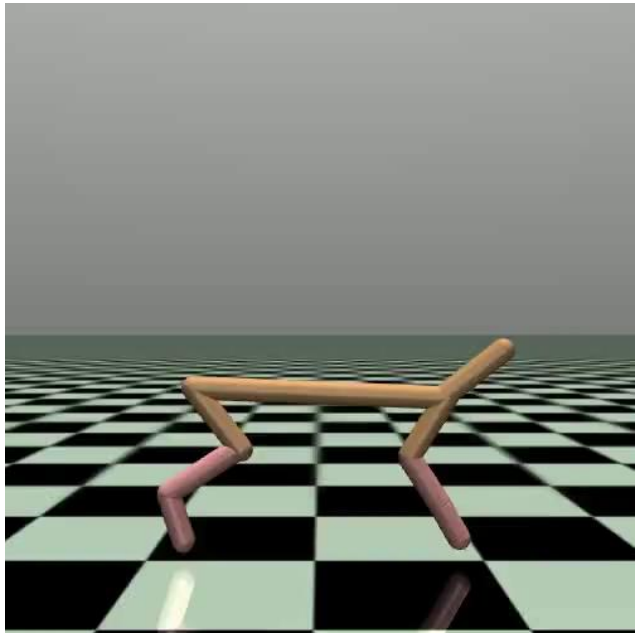



500
400
300
200
100
0
CQL
IQL
MOBILE
CBOP
RAMBO
LEQ
Model-free
Model-based
LEQ(Ours)

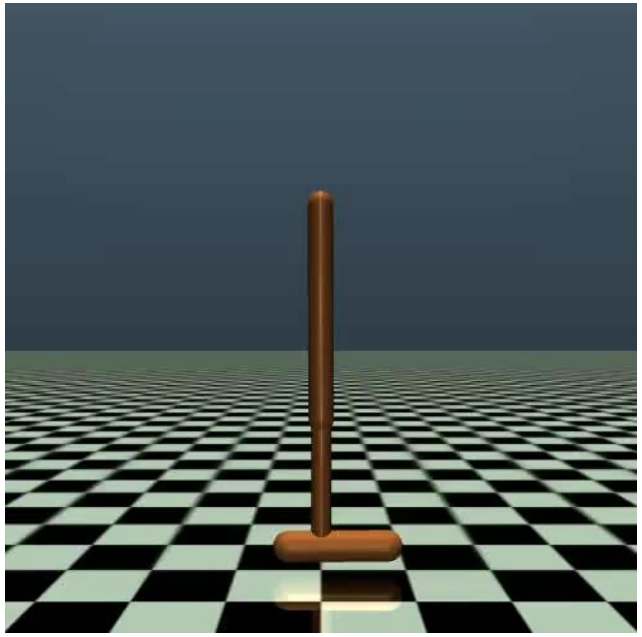
AntMaze


umaze
medium-play
large-play
ultra-play
umaze-diverse
medium-diverse
large-diverse
ultra-diverse
100
75
50
25
0
CQL
IQL
MOBILE
CBOP
RAMBO
LEQ
Model-free
Model-based
LEQ(Ours)

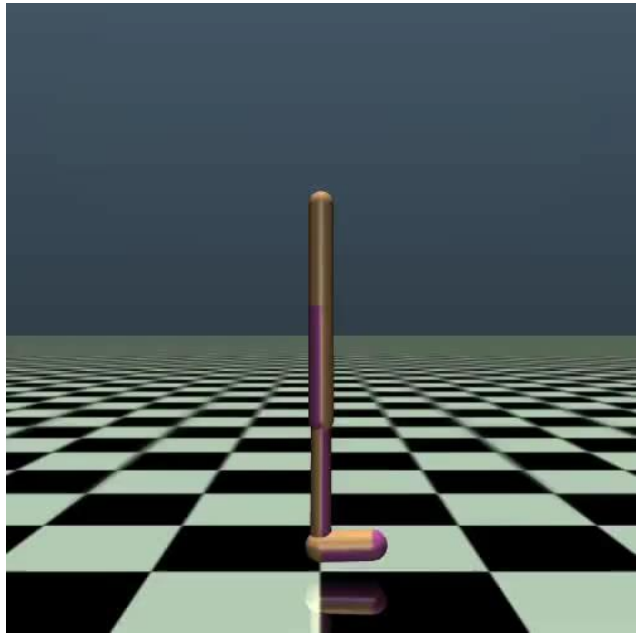
Mujoco Gym (NeoRL, D4RL)



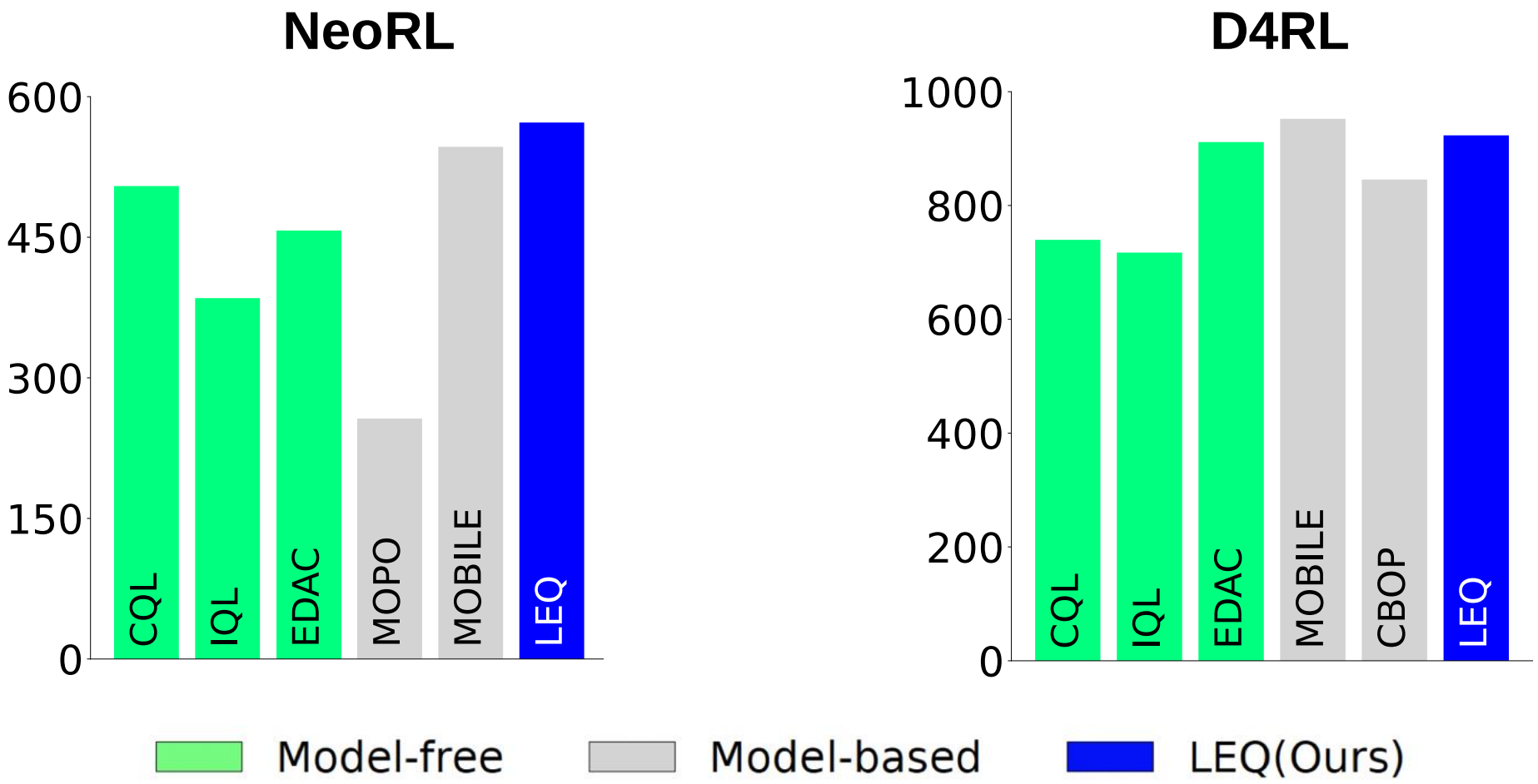
HalfCheetah



Hopper



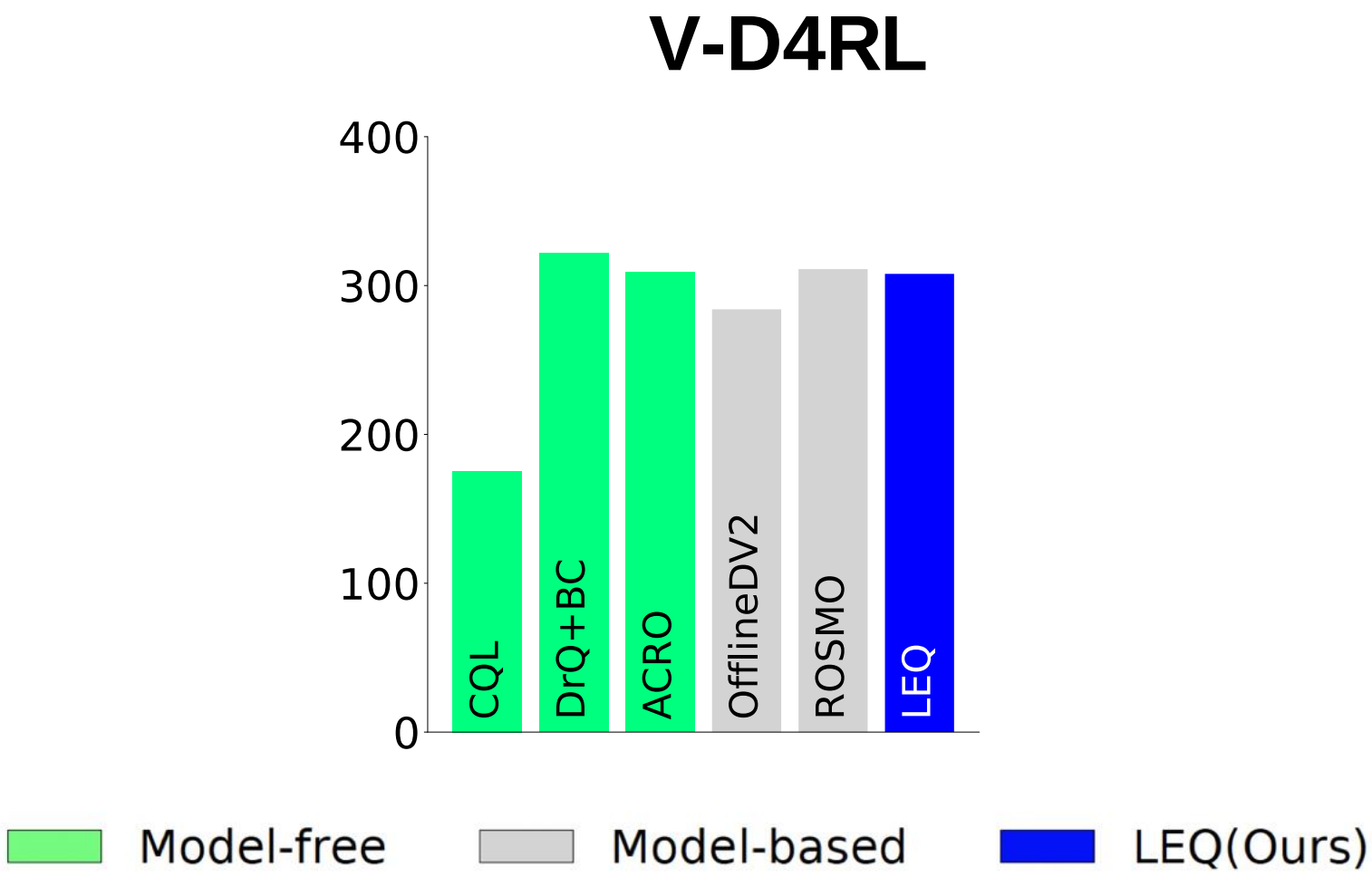
Walker


NeoRL
600
450
300
150
0
CQL
IQL
EDAC
MOPO
MOBILE
LEQ
D4RL
1000
800
600
400
200
0
CQL
IQL
EDAC
MOBILE
CBOP
LEQ
Model-free
Model-based
LEQ(Ours)

DMControl (V-D4RL)

Cheetah-Run

Walker-Walk


V-D4RL
400
300
200
100
0
CQL
DrQ+BC
ACRO
OfflineDV2
ROSMO
LEQ
Model-free
Model-based
LEQ(Ours)

Summary

- LEQ is an offline **model-based** RL algorithm.
- LEQ optimizes the **lower-expectile** of the returns.
  - Assign **larger** weights to **conservative** trajectories
- LEQ works well in various domains.

Project Page