Reinforcement Learning as Sequence Modeling

Presenter: Igor Mordatch

Reinforcement Learning as **Sequence Modeling**

- Transformer-based architectures & large-scale pre-training (GPT/BERT/T5/etc)
- A lot of research, compute, and infrastructure investment
- How can we get the most out of this investment?

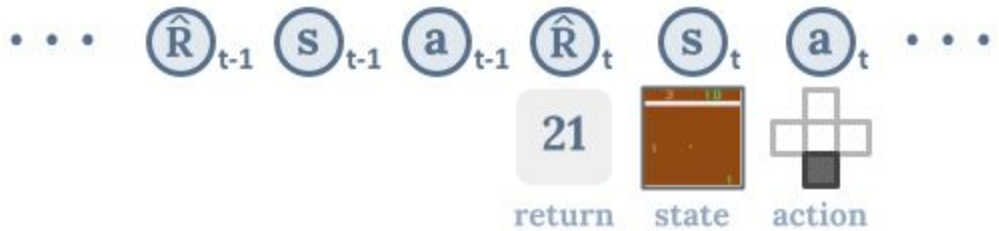
Reinforcement Learning as Sequence Modeling

- Not just using transformer network for policy/value, but replace RL algorithms with sequence modeling
- Drastic problem reduction, but how far can it take us?

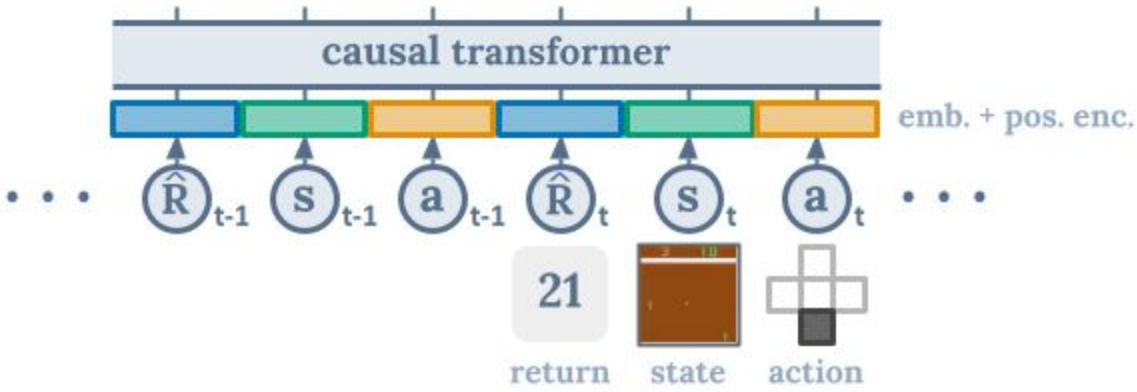
Decision Transformer: Reinforcement Learning via Sequence Modeling

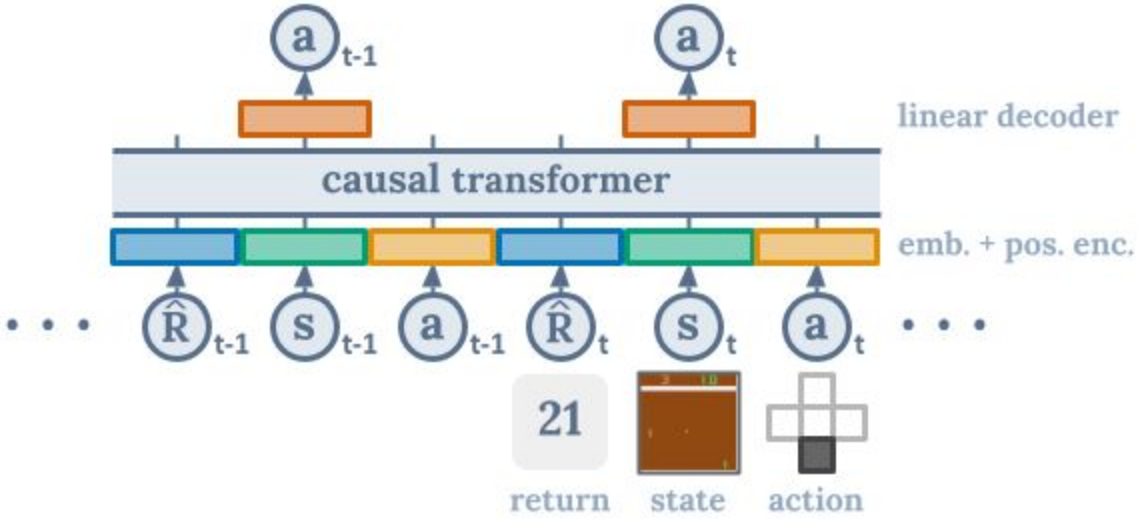
Lili Chen*, Kevin Lu*, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas*, Igor Mordatch*

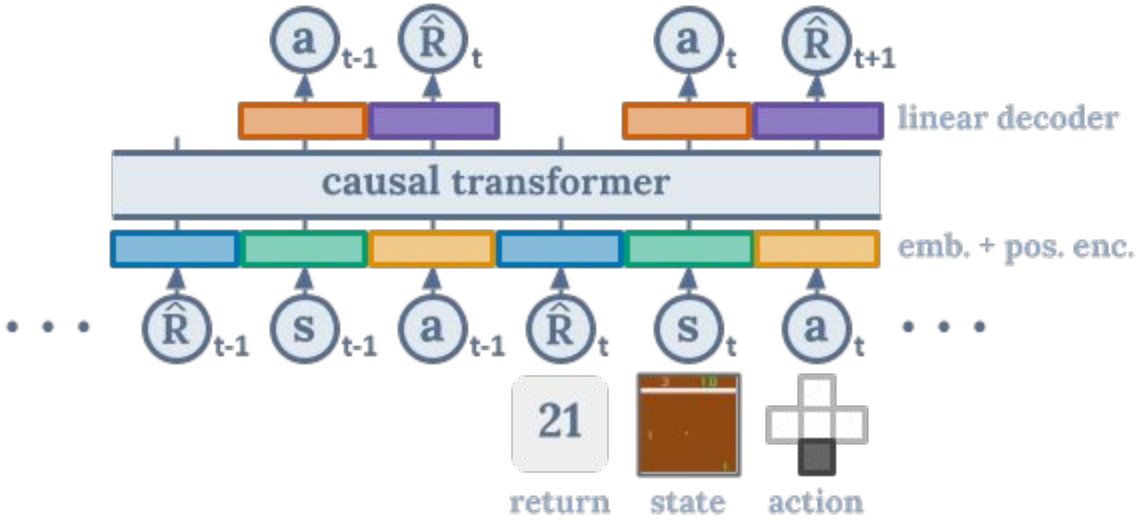

$\hat{R}_{t-1}$ s_{t-1} a_{t-1} $\hat{R}_t$ s_t a_t



$$R_t = \sum_{t'=t}^{T} r_{t'}$$ (incorporates hindsight information)


causal transformer
emb. + pos. enc.
$\hat{R}_{t-1}$
s_{t-1}
a_{t-1}
$\hat{R}_t$
s_t
a_t
21
return
state
action

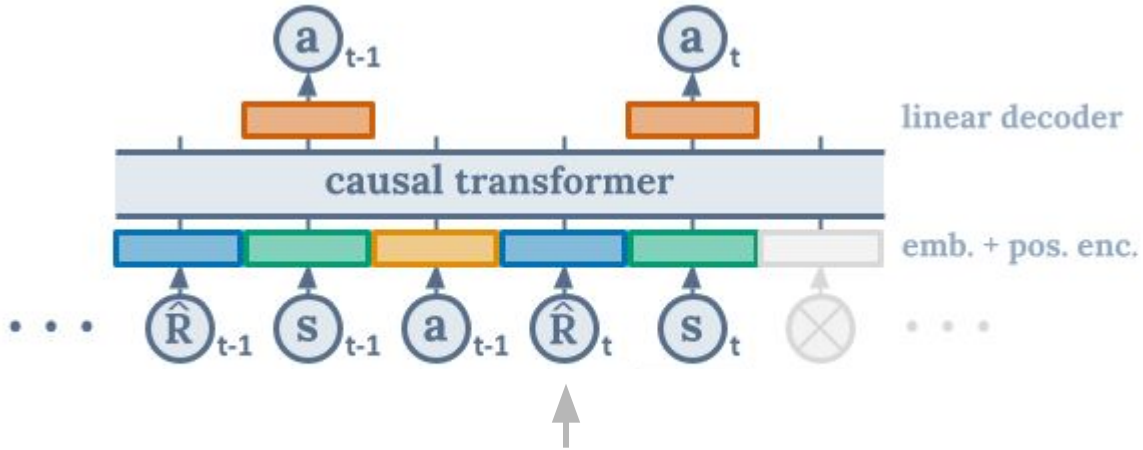

a t-1
a t
linear decoder
causal transformer
emb. + pos. enc.
R̂ t-1
s t-1
a t-1
R̂ t
s t
a t
21
return
state
action



[Representation learning with reward prediction errors, Alexander and Gershman, 2021]

[On The Effect of Auxiliary Tasks on Representation Dynamics, Lyle et al, 2021]

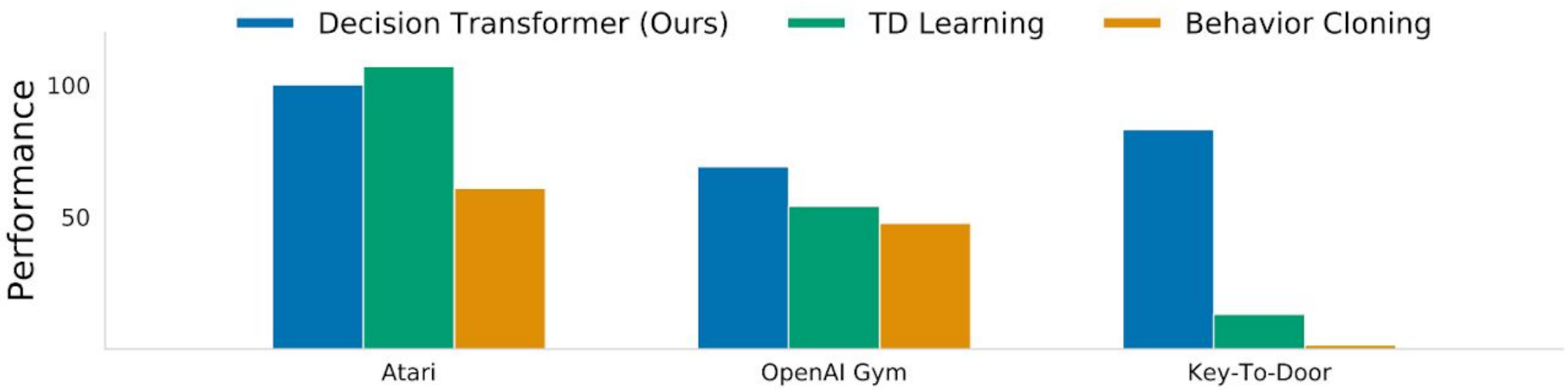
At deployment step t:



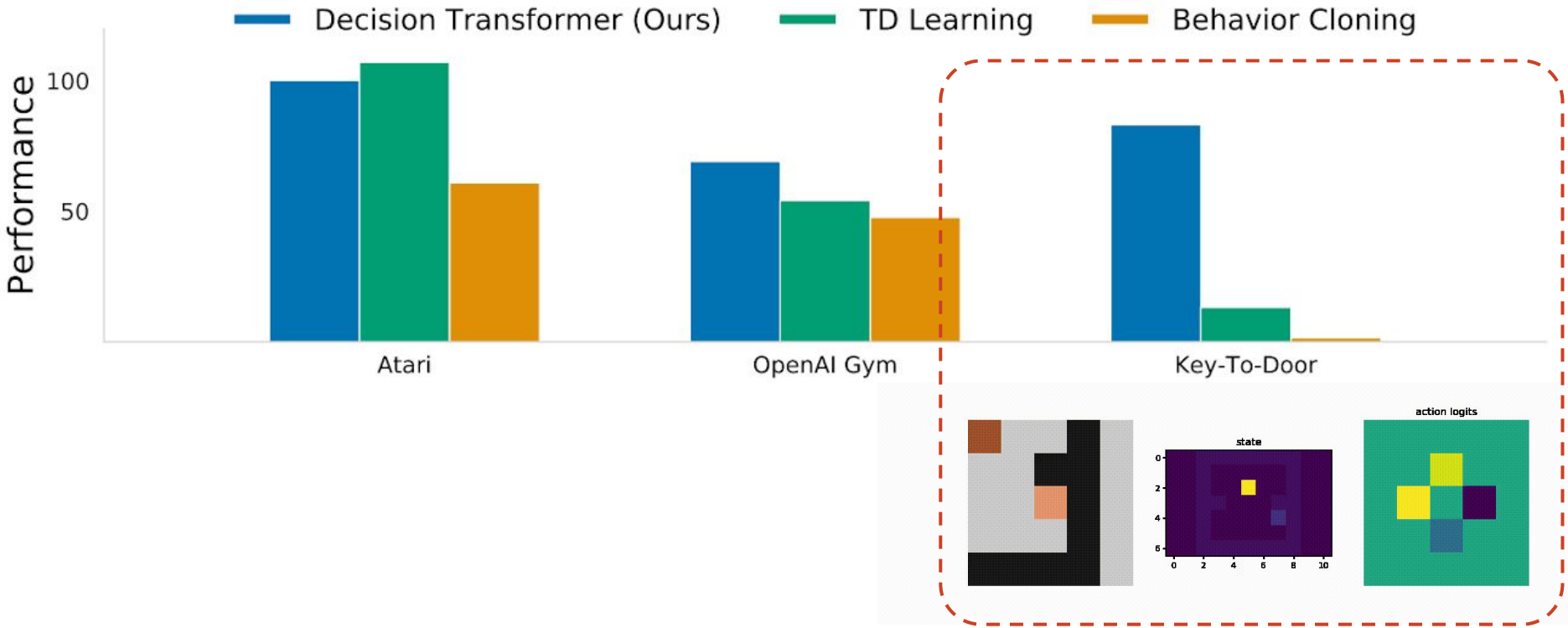
condition on target return/optimality

[Probabilistic Inference and Influence Diagrams, Shachter, 1988]
[Robot Trajectory Optimization using Approximate Inference, Toussaint, 2009]
[Optimal control as a graphical model inference problem, Kappen et al, 2012]
[Reinforcement Learning Upside Down, Schmidhuber, 2019]

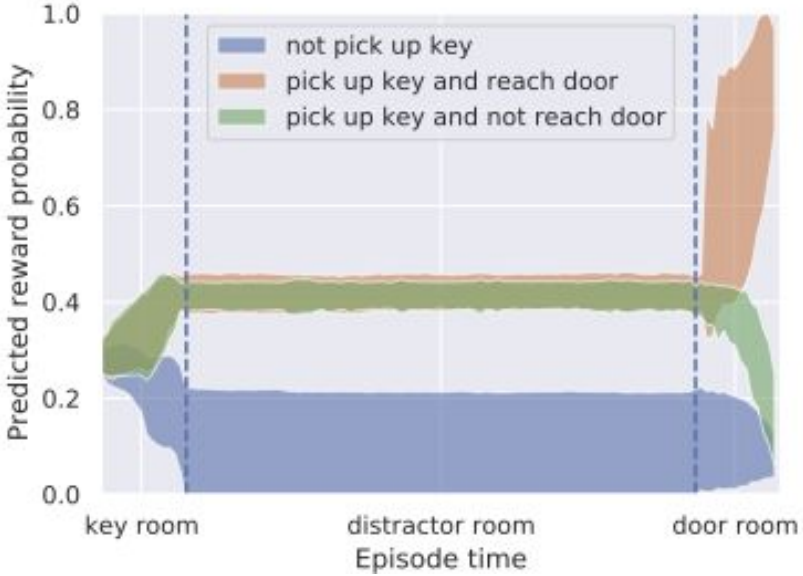
Very Promising Offline RL Results



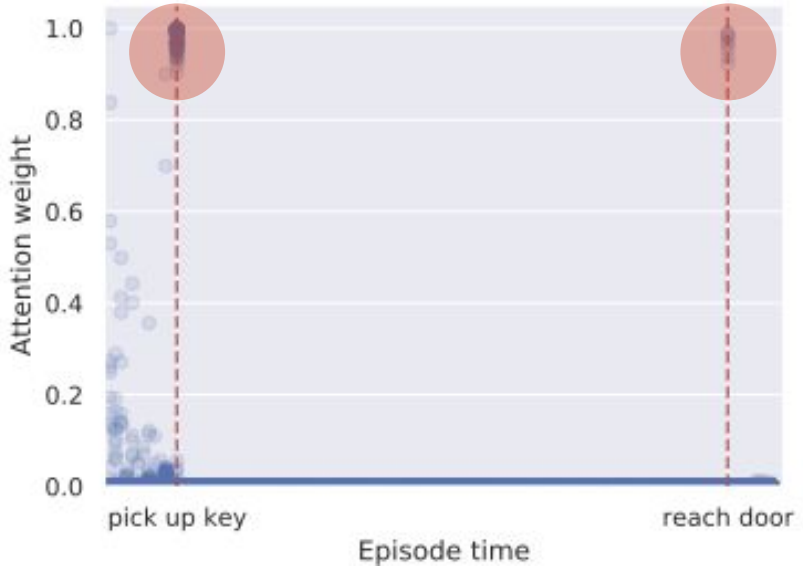
Very Promising Offline RL Results

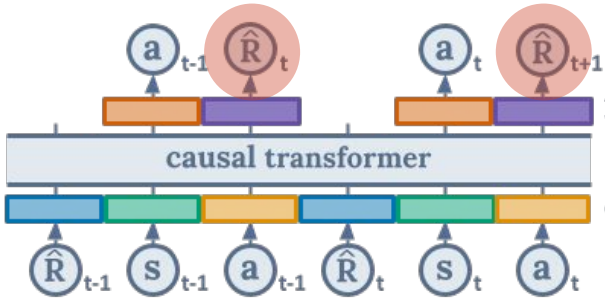


Future reward prediction

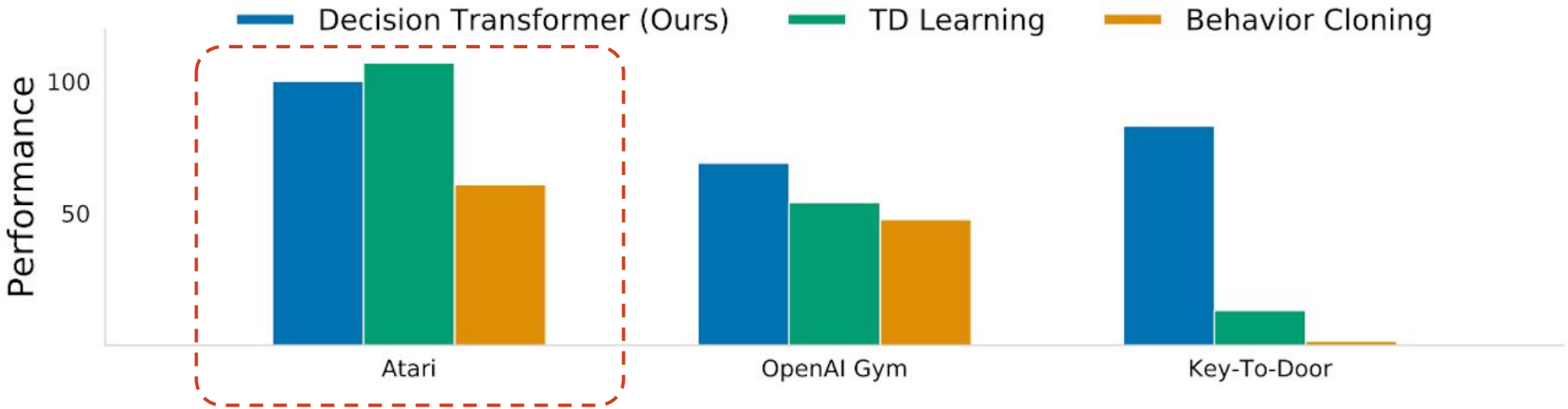


Attention to past events



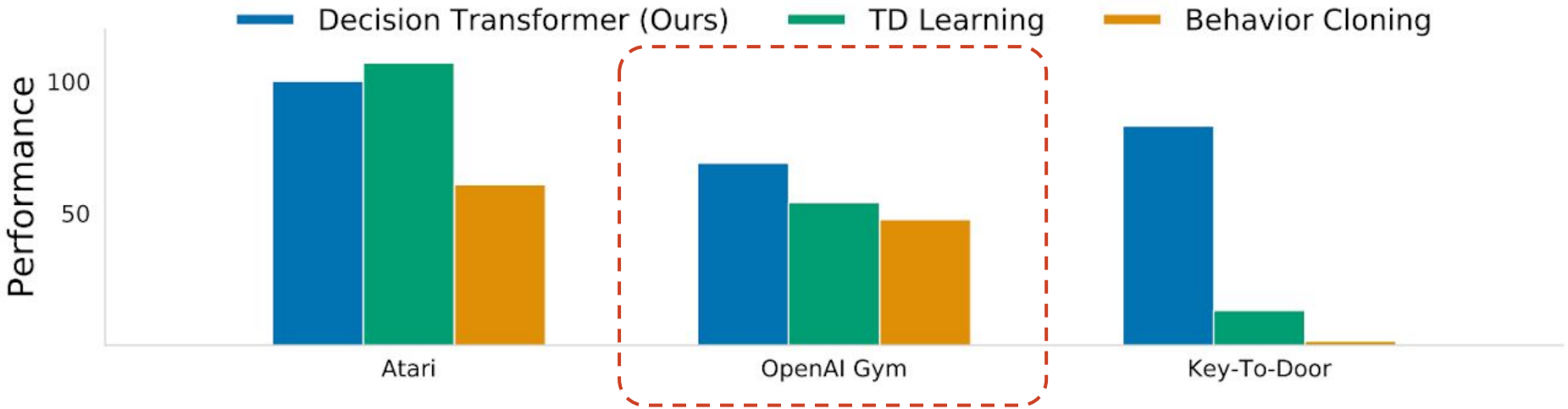


Very Promising Offline RL Results



Game	DT (Ours)	CQL	QR-DQN	REM	BC
Breakout	**267.5 ± 97.5**	211.1	21.1	32.1	138.9 ± 61.7
Qbert	25.1 ± 18.1	**104.2**	1.7	1.4	17.3 ± 14.7
Pong	106.1 ± 8.1	**111.9**	20.0	39.1	85.2 ± 20.0
Seaquest	**2.4 ± 0.7**	1.7	1.4	1.0	2.1 ± 0.3

Very Promising Offline RL Results

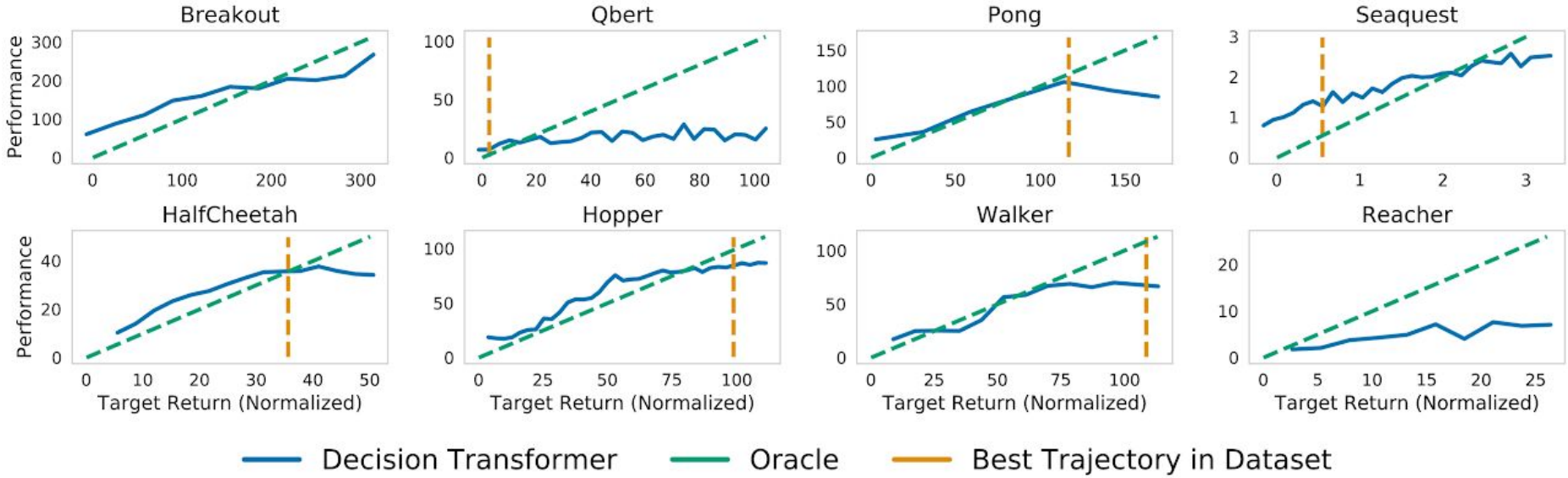


Very Promising Offline RL Results

Dataset	Environment	DT (Ours)	CQL	BEAR	BRAC-v	AWR	BC
Medium-Expert	HalfCheetah	**86.8 ± 1.3**	62.4	53.4	41.9	52.7	59.9
Medium-Expert	Hopper	107.6 ± 1.8	**111.0**	96.3	0.8	27.1	79.6
Medium-Expert	Walker	**108.1 ± 0.2**	98.7	40.1	81.6	53.8	36.6
Medium-Expert	Reacher	**89.1 ± 1.3**	30.6	-	-	-	73.3
Medium	HalfCheetah	42.6 ± 0.1	44.4	41.7	**46.3**	37.4	43.1
Medium	Hopper	**67.6 ± 1.0**	58.0	52.1	31.1	35.9	63.9
Medium	Walker	74.0 ± 1.4	79.2	59.1	**81.1**	17.4	77.3
Medium	Reacher	**51.2 ± 3.4**	26.0	-	-	-	**48.9**
Medium-Replay	HalfCheetah	36.6 ± 0.8	46.2	38.6	**47.7**	40.3	4.3
Medium-Replay	Hopper	**82.7 ± 7.0**	48.6	33.7	0.6	28.4	27.6
Medium-Replay	Walker	**66.6 ± 3.0**	26.7	19.2	0.9	15.5	36.9
Medium-Replay	Reacher	**18.0 ± 2.4**	**19.0**	-	-	-	5.4
Average (Without Reacher)		**74.7**	63.9	48.2	36.9	34.3	46.4
Average (All Settings)		**69.2**	54.2	-	-	-	47.7

Investigations

- How well does it model the distribution of returns?



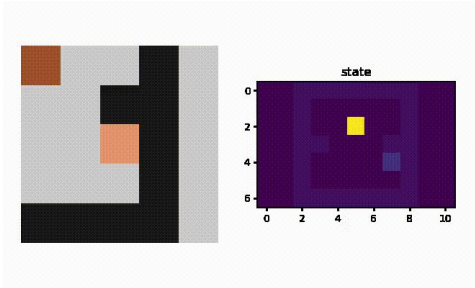
Investigations

- How well does it model the distribution of returns?
- What is the benefit of using a longer context length?

Game	DT (Ours)	DT with no context ($K = 1$)
Breakout	$\mathbf{267.5 \pm 97.5}$	73.9 ± 10
Qbert	$\mathbf{25.1 \pm 18.1}$	13.7 ± 6.5
Pong	$\mathbf{106.1 \pm 8.1}$	2.5 ± 0.2
Seaquest	$\mathbf{2.4 \pm 0.7}$	0.5 ± 0.0

Investigations

- How well does it model the distribution of returns?
- What is the benefit of using a longer context length?
- Does it perform effective long-term credit assignment?



Dataset	DT (Ours)	CQL	BC	%BC	Random
1K Random Trajectories	**71.8%**	13.1%	1.4%	69.9%	3.1%
10K Random Trajectories	94.6%	13.3%	1.6%	**95.1%**	3.1%

Investigations

- How well does it model the distribution of returns?
- What is the benefit of using a longer context length?
- Does it perform effective long-term credit assignment?
- Does it perform well in sparse reward settings?

		Delayed (Sparse)		Agnostic		Original (Dense)	
Dataset	**Environment**	**DT (Ours)**	**CQL**	**BC**	**%BC**	**DT (Ours)**	**CQL**
Medium-Expert	Hopper	**107.3 ± 3.5**	9.0	59.9	102.6	107.6	111.0
Medium	Hopper	60.7 ± 4.5	5.2	63.9	**65.9**	67.6	58.0
Medium-Replay	Hopper	**78.5 ± 3.7**	2.0	27.6	70.6	82.7	48.6

Reinforcement Learning as Sequence Modeling

- Did not introduce new algorithms or models
- Analysis and problem reduction to whittle down the space of methods
- Build on our community's efforts more effectively

Links

arXiv: arxiv.org/abs/2106.01345

Github: github.com/kzl/decision-transformer