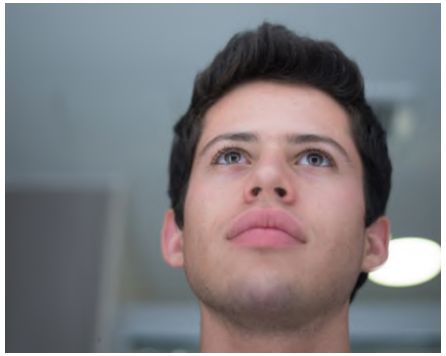
Exploiting Latent Structure and Bisimulation Metrics for Better Generalization

Amy Zhang

Arxiv: 2006.10742

Learning Invariant Representations for Reinforcement Learning without Reconstruction

Amy Zhang[*1,2] **Rowan McAllister**[*3] **Roberto Calandra**[2] **Yarin Gal**[4] **Sergey Levine**[3]
[1]McGill University
[2]Facebook AI Research
[3]University of California, Berkeley
[4]OATML group, University of Oxford



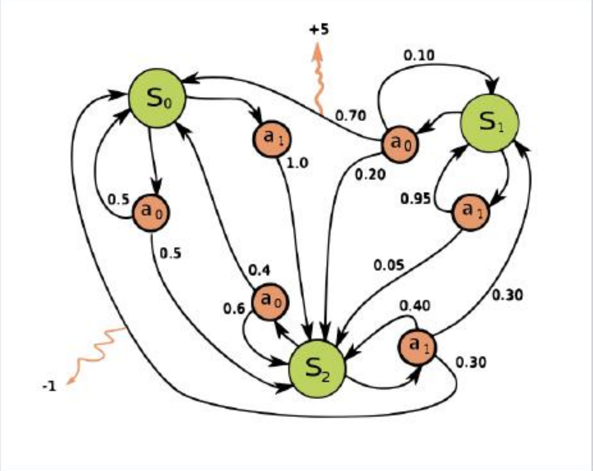
* Equal contribution

Markov Decision Processes

Definition [edit]

A Markov decision process is a 4-tuple (S, A, P_a, R_a), where

- S is a finite set of states,
- A is a finite set of actions (alternatively, A_s is the finite set of actions available from state s),
- $P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t+1$,
- $R_a(s, s')$ is the immediate reward (or expected immediate reward) received after transitioning from state s to state s', due to action a

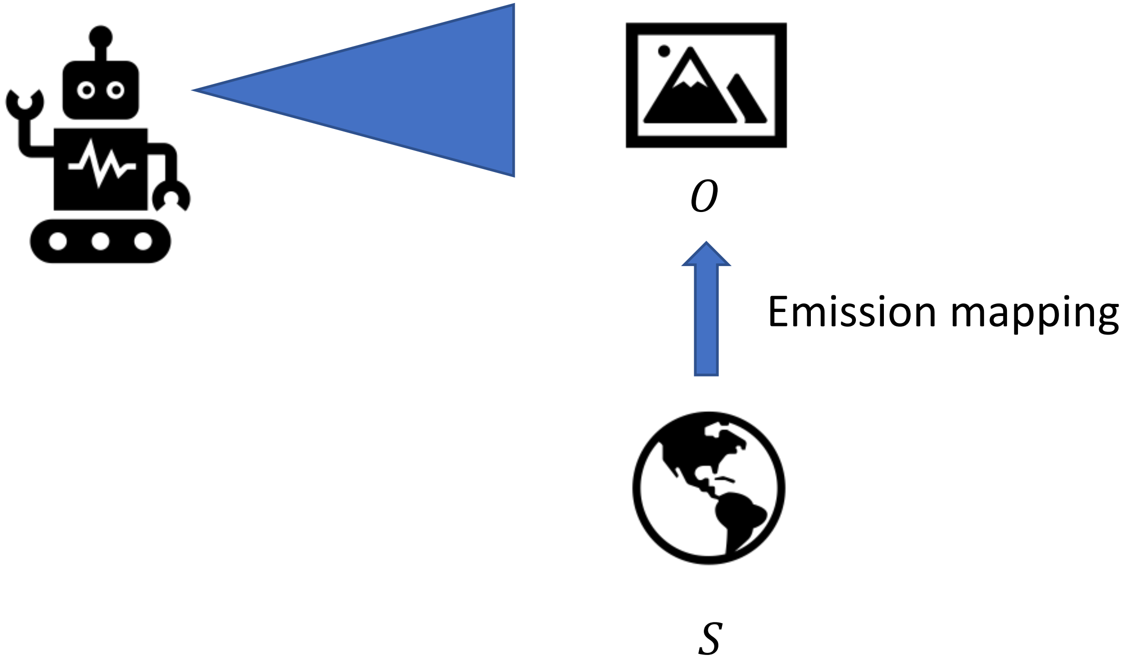


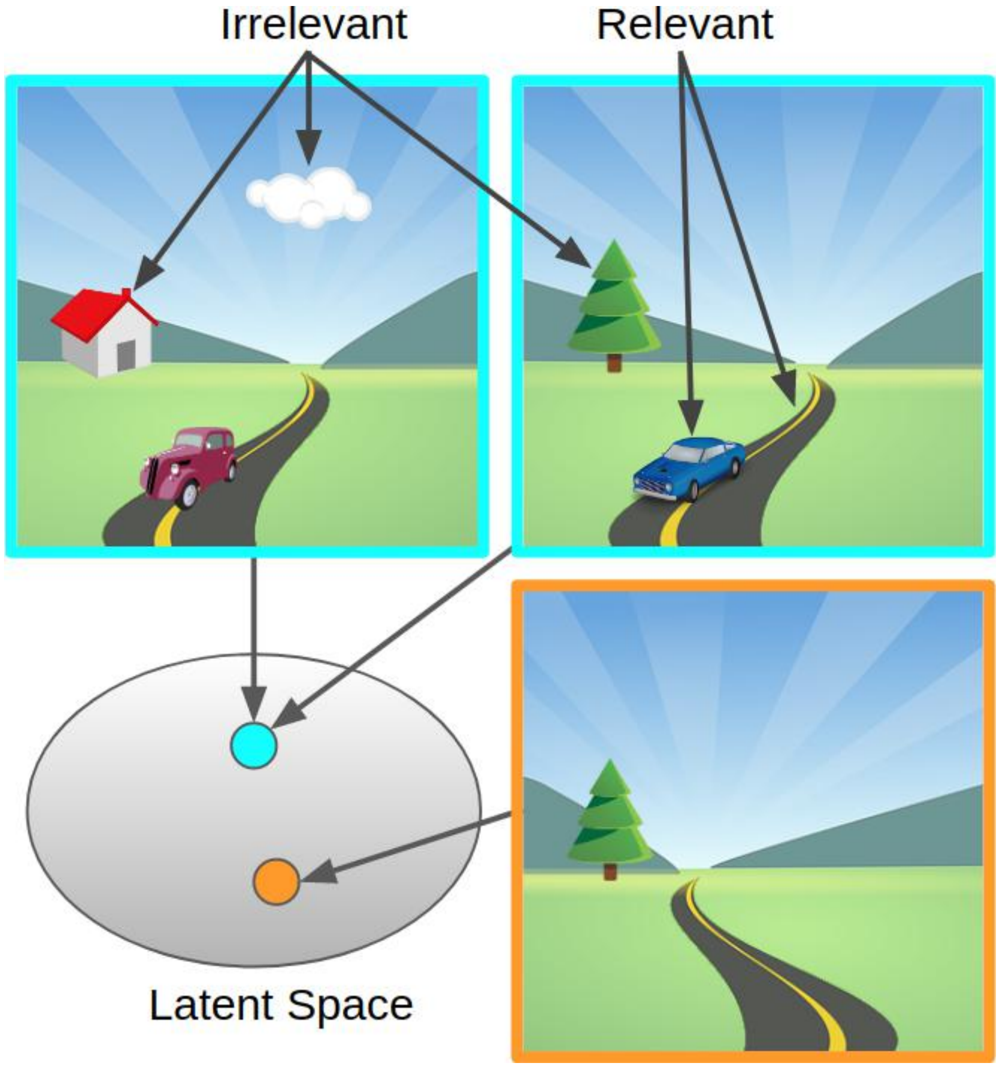
Example of a simple MDP with three states (green circles) and two actions (orange circles), with two rewards (orange arrows).

What kind of additional structure is reasonable to assume in MDPs ?

From wikipedia

A realistic additional assumption





Goal: Generalization to new observations *where the underlying MDP is the same*
Solution: Ignore irrelevant information

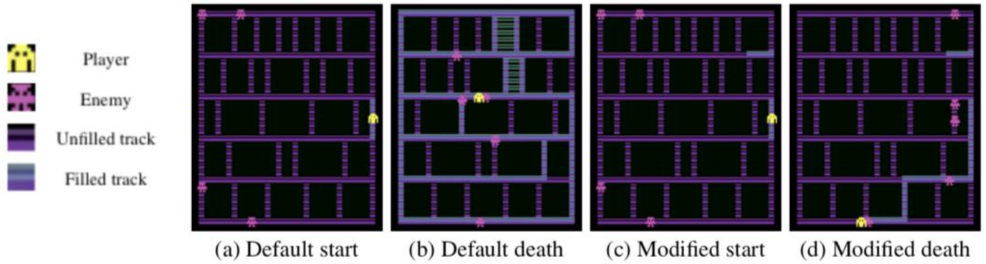


Figure: Train and Test on Atari proposed by Witty et al. 2018

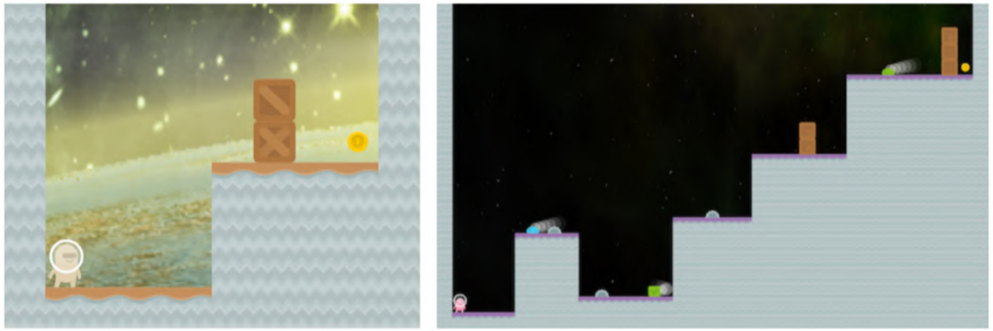


Figure: Train and Test on CoinRun proposed by Cobbe et al. 2019

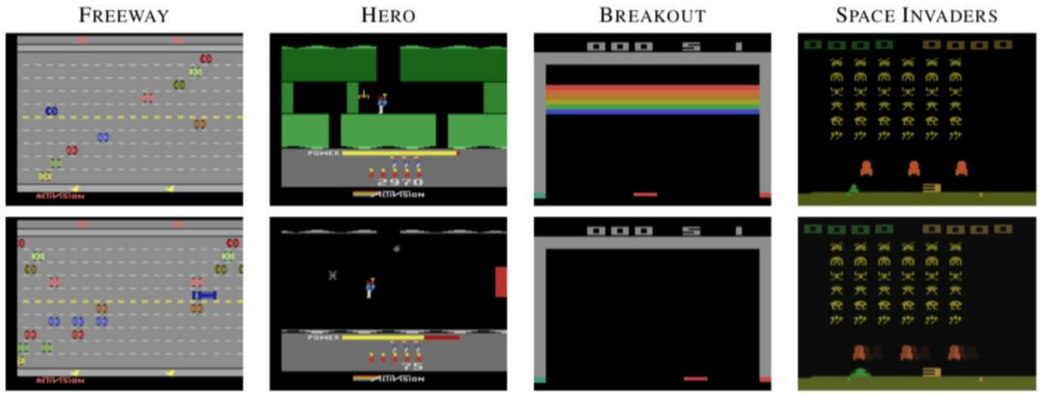


Figure: Train and Test on Atari proposed by Farebrother, Machado, and Bowling 2018

State Abstractions and Bisimulation

State abstractions have been studied as a way to distinguish relevant from irrelevant information in order to create a more compact representation for easier decision making and planning.

Definition 1 (Bisimulation Relations (Givan et al., 2003)). *Given an MDP $\mathcal{M}$, an equivalence relation B between states is a bisimulation relation if for all states $s_1, s_2 \in \mathcal{S}$ that are equivalent under B (i.e. $s_1 B s_2$), the following conditions hold for all actions $a \in \mathcal{A}$:*

$$R(s_1, a) = R(s_2, a)$$
$$\mathcal{P}(G|s_1, a) = \mathcal{P}(G|s_2, a), \forall G \in \mathcal{S}/B$$

Where $\mathcal{S}/B$ denotes the partition of $\mathcal{S}$ under the relation B, the set of all groups of equivalent states, and where $\mathcal{P}(G|s, a) = \sum_{s' \in G} \mathcal{P}(s'|s, a)$.

Bisimulation Metrics

State abstraction only groups equivalent states. What about a metric for state similarity?

Definition 3 (Bisimulation Metric (Theorem 2.6 in Ferns et al. [7])). *Let $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ be a finite MDP and let $c \in (0, 1)$ be a discount factor. Let $\mathfrak{met}$ be the space of bounded pseudometrics on $\mathcal{S}$ equipped with the metric induced by the uniform norm. Define $F : \mathfrak{met} \mapsto \mathfrak{met}$ by*

$$F(\mathbf{s}, \mathbf{s}') = \max_{\mathbf{a} \in \mathcal{A}} (1 - c)|r^{\mathbf{a}}_{\mathbf{s}} - r^{\mathbf{a}}_{\mathbf{s}'}| + cW(\mathcal{P}^{\mathbf{a}}_{\mathbf{s}}, \mathcal{P}^{\mathbf{a}}_{\mathbf{s}'}). \quad (2)$$

Then F has a unique fixed point $\tilde{d}$ which is the bisimulation metric.

On-Policy Bisimulation Metrics

Let's modify the previous definition to get rid of the max over actions:

Theorem 1. *Let* $\mathfrak{met}$ *be the space of bounded pseudometrics on* S *and* π *a policy that is continuously improving. Define* $\mathcal{F} : \mathfrak{met} \mapsto \mathfrak{met}$ *by*

$$\mathcal{F}(d, \pi)(\mathbf{s}_i, \mathbf{s}_j) = (1 - c) \cdot |\mathcal{R}^{\pi}_{\mathbf{s}_i} - \mathcal{R}^{\pi}_{\mathbf{s}_j}| + c \cdot W(d)(\mathcal{P}^{\pi}_{\mathbf{s}_i}, \mathcal{P}^{\pi}_{\mathbf{s}_j}). \tag{5}$$

Then $\mathcal{F}$ *has a least fixed point* $\tilde{d}$ *which is a* π^**-bisimulation metric.*

Another issue…

Computing the empirical Wasserstein of a generative model is difficult.

However, there are closed form solutions for Gaussian distributions:

$$W_2(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j))^2 = ||\mu_i - \mu_j||_2^2 + ||\Sigma_i^{1/2} - \Sigma_j^{1/2}||_\mathcal{F}^2$$

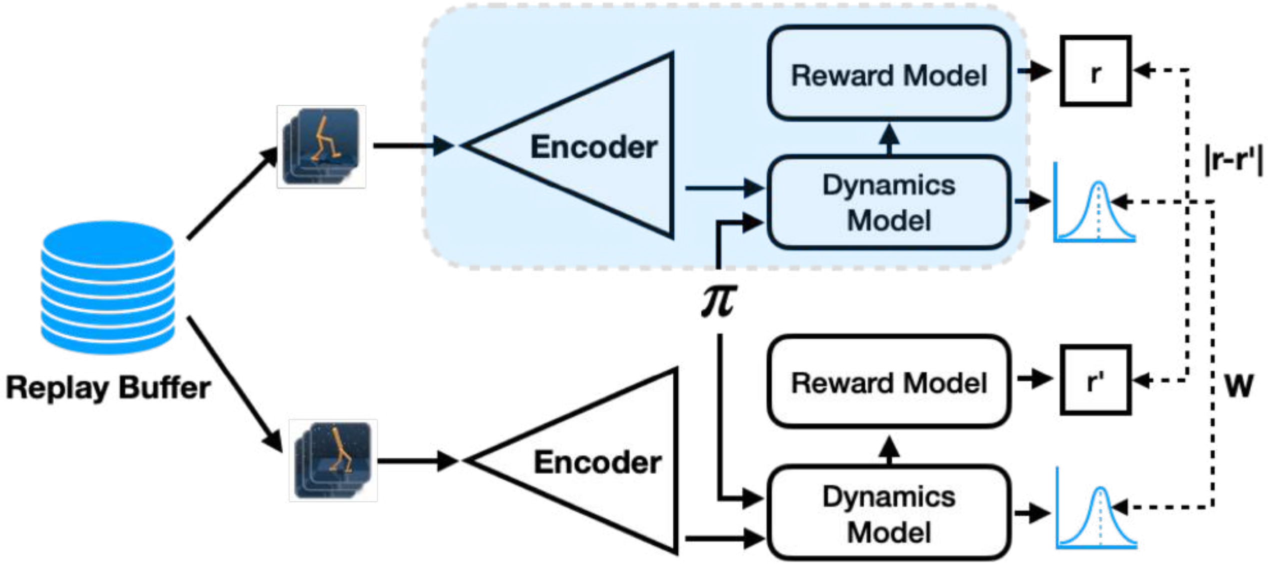
Frobenius norm

The representation learning objective

Learn a representation where L1 distance between any two states is a measure of their bisimilarity:

$$
\begin{aligned}
J(\phi) &= \Big(|\mathbf{z}_1 - \mathbf{z}_2| - d(\mathbf{o}_1, \mathbf{o}_2)\Big)^2 \\
&= \Big(|\mathbf{z}_1 - \mathbf{z}_2| - \mathbb{E}_{\pi_b}\Big[|r^{\pi_b}_{\mathbf{o}_1} - r^{\pi_b}_{\mathbf{o}_2}| + \gamma \cdot d_P(P^{\mathbf{a}}_{\mathbf{o}_1}, P^{\mathbf{a}}_{\mathbf{o}_2})\Big]\Big)^2 \\
&= \Big(|\phi(\mathbf{o}_1) - \phi(\mathbf{o}_2)| - \mathbb{E}_{\mathbf{a}\sim\pi_b}\Big[|\mathcal{R}(\mathbf{o}_1, \mathbf{a}) - \mathcal{R}(\mathbf{o}_2, \mathbf{a})| \\
&\qquad + \gamma \cdot W\big(q(\phi(\mathbf{o}_1')|\phi(\mathbf{o}_1), \mathbf{a}),\ q(\phi(\mathbf{o}_2')|\phi(\mathbf{o}_2), \mathbf{a})\big)\Big]\Big)^2
\end{aligned}
$$

Deep Bisimulation for Control (DBC)



Algorithm 1 Deep Bisimulation for Control (DBC)

for Time $t = 0$ to ∞ **do**
  Encode observation $\mathbf{z}_t = \phi(\mathbf{s}_t)$
  Execute action $\mathbf{a}_t \sim \pi(\mathbf{z}_t)$
  Record data: $\mathcal{D} \leftarrow \mathcal{D} \cup \{\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_{t+1}\}$
  Sample batch $B_i \sim \mathcal{D}$
  Permute batch randomly: $B_j = \text{permute}(B_i)$
  Train policy: $\mathbb{E}_{B_i}[J(\pi)]$ ▷ Algorithm 2
  Train encoder: $\mathbb{E}_{B_i,B_j}[J(\phi)]$ ▷ Equation (4)
  Train dynamics: $J(\hat{\mathcal{P}},\phi) = (\hat{\mathcal{P}}(\phi(\mathbf{s}_t), \mathbf{a}_t) - \bar{\mathbf{z}}_{t+1})^2$
  Train reward: $J(\hat{\mathcal{R}},\hat{\mathcal{P}},\phi) = (\hat{\mathcal{R}}(\hat{\mathcal{P}}(\phi(\mathbf{s}_t),\mathbf{a}_t)) - r_{t+1})^2$

Algorithm 2 Train Policy (changes to SAC in blue)

Get value: $V = \min_{i=1,2} \hat{Q}_i(\hat{\phi}(\mathbf{s})) - \alpha \log \pi(\mathbf{a}|\phi(\mathbf{s}))$
Train critics: $J(Q_i, \phi) = (Q_i(\phi(\mathbf{s})) - r - \gamma V)^2$
Train actor: $J(\pi) = \alpha \log p(\mathbf{a}|\phi(\mathbf{s})) - \min_{i=1,2} Q_i(\phi(\mathbf{s}))$
Train alpha: $J(\alpha) = -\alpha \log p(\mathbf{a}|\phi(\mathbf{s}))$
Update target critics: $\hat{Q}_i \leftarrow \tau_Q Q_i + (1 - \tau_Q)\hat{Q}_i$
Update target encoder: $\hat{\phi} \leftarrow \tau_\phi \phi + (1 - \tau_\phi)\hat{\phi}$

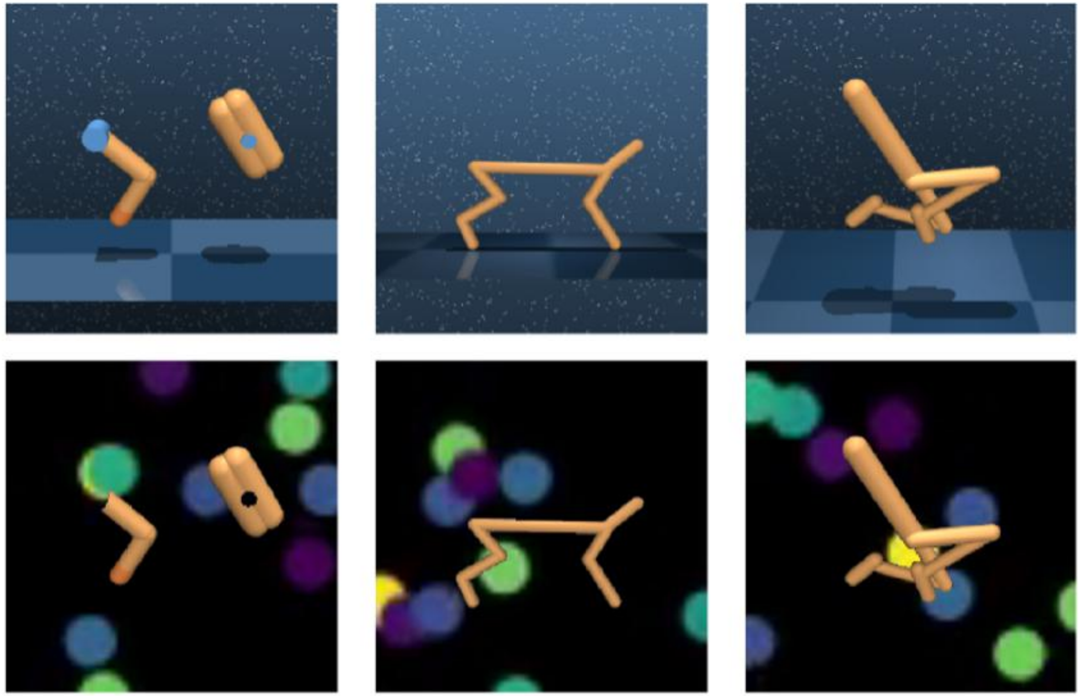
Representation Learning with Bisimulation Metrics

Distractions:

No Background

Representation Learning with Bisimulation Metrics

Distractions:

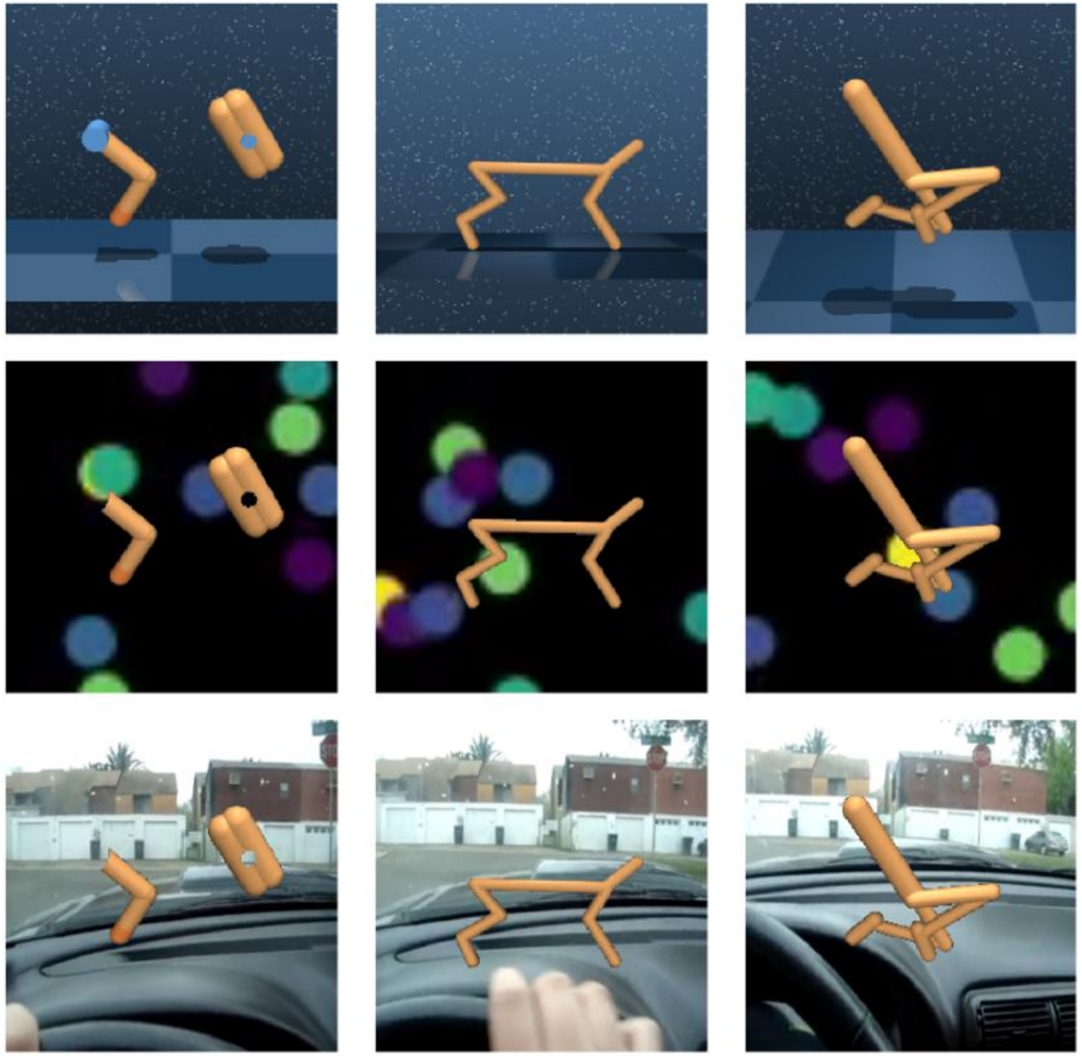


No Background

Simple Distractors

Representation Learning with Bisimulation Metrics

Distractions:

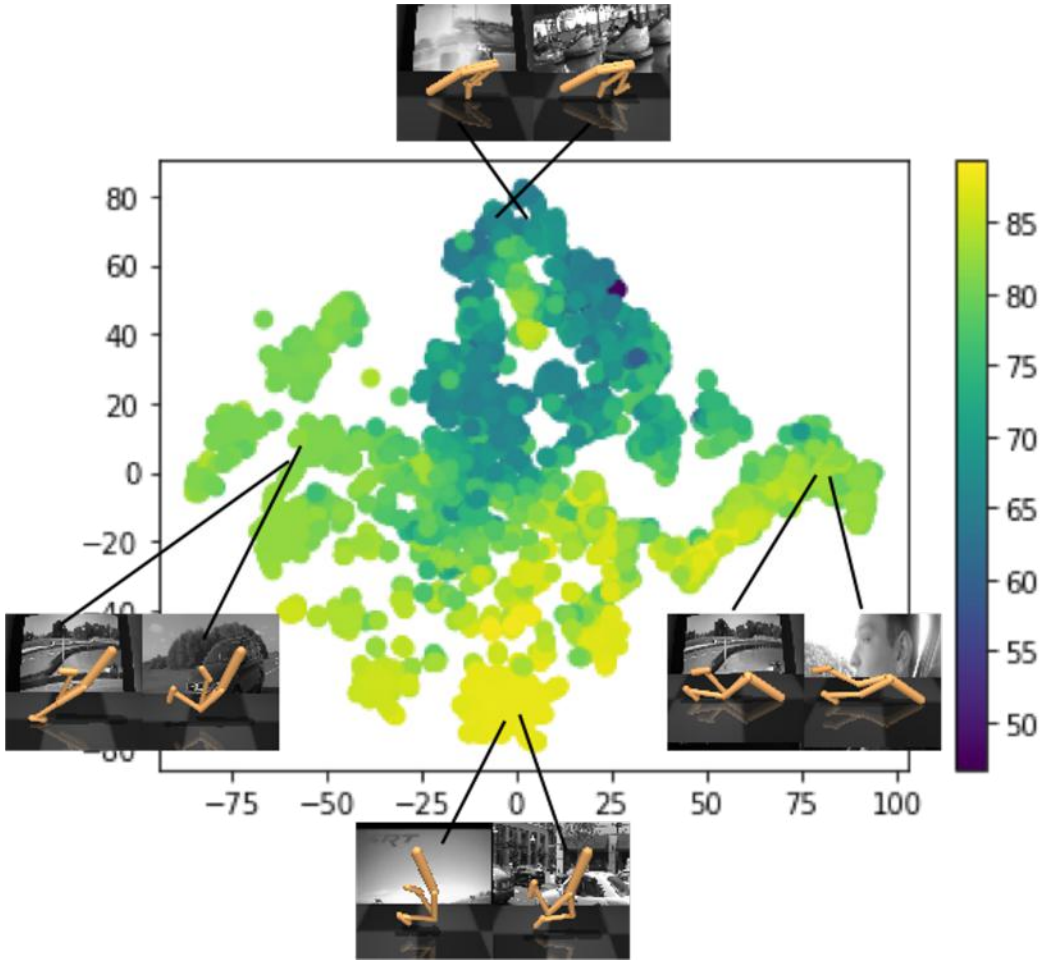


No Background

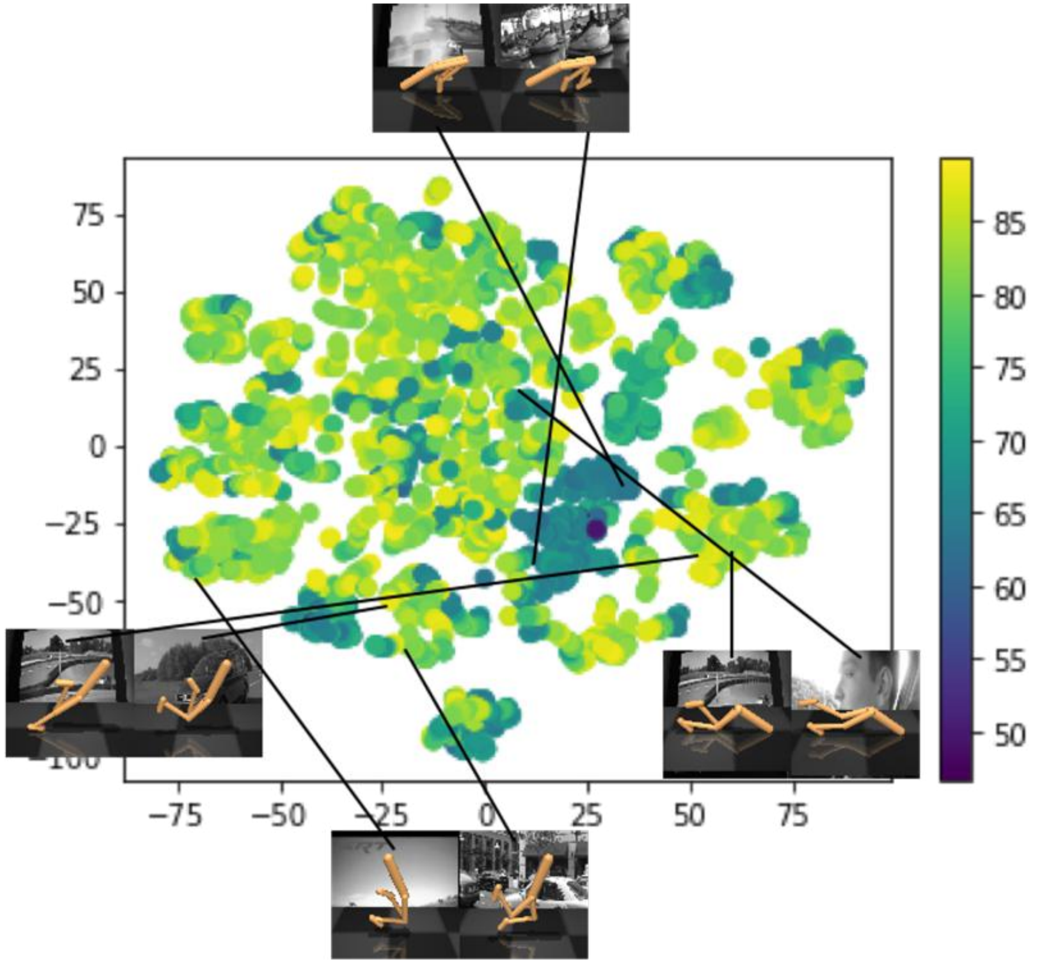
Simple Distractors

Natural Video

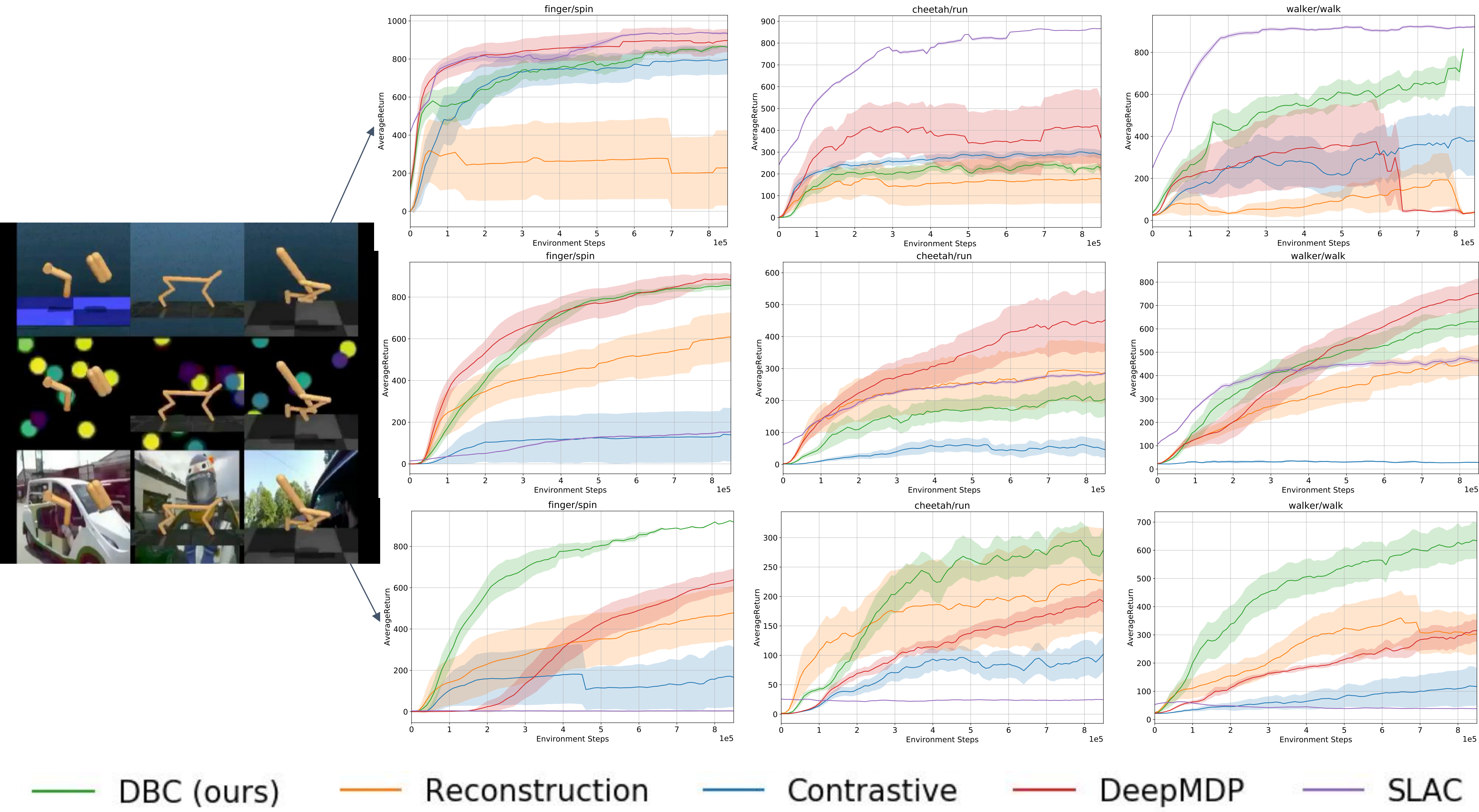
Representation Learning with Bisimulation Metrics



t-SNE of **Bisimulation** codes



t-SNE of **VAE** codes



finger/spin
cheetah/run
walker/walk
AverageReturn
Environment Steps
1e5
DBC (ours)
Reconstruction
Contrastive
DeepMDP
SLAC

Connections to Causal Inference

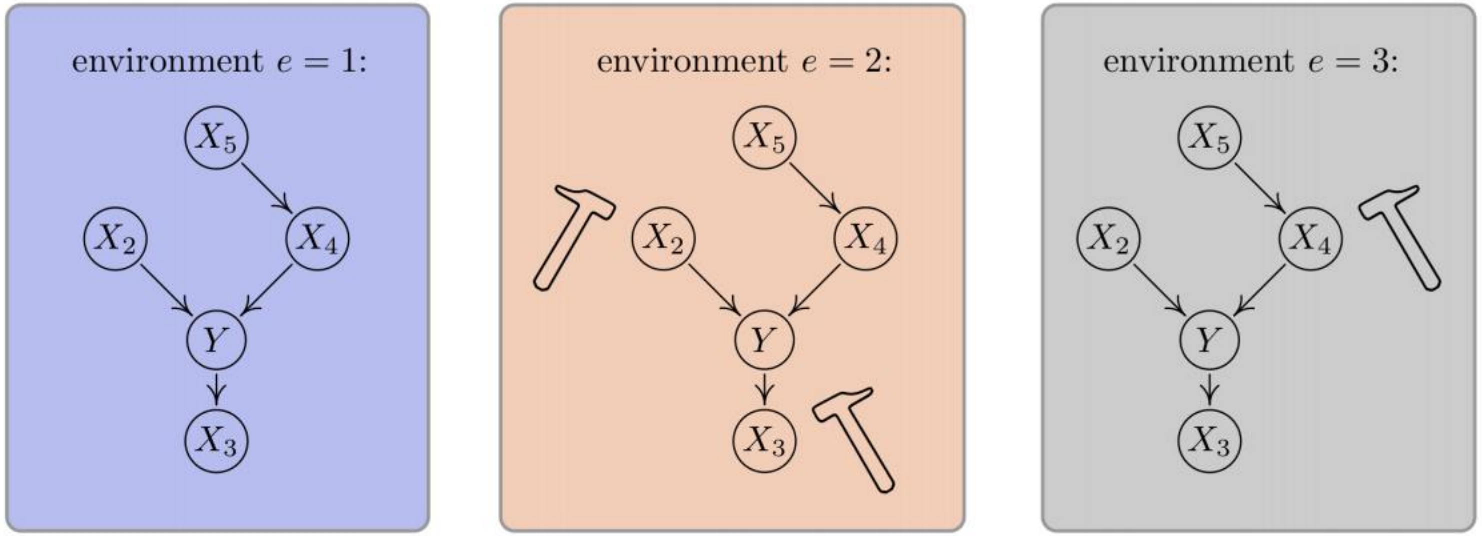
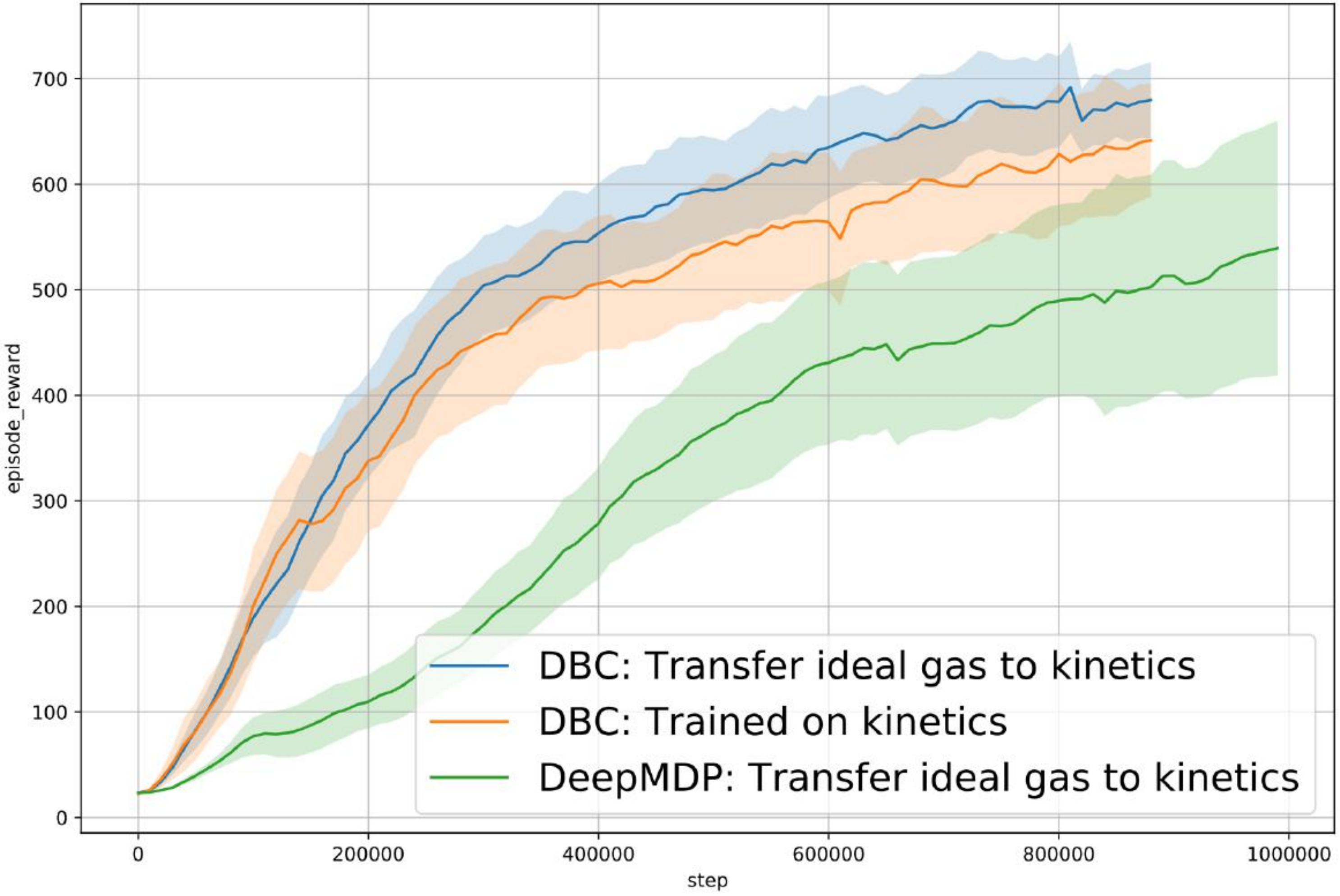


Figure 1: An example including three environments. The invariance (1) and (2) holds if we consider $S^* = \{X_2, X_4\}$. Considering indirect causes instead of direct ones (e.g. $\{X_2, X_5\}$) or an incomplete set of direct causes (e.g. $\{X_4\}$) may not be sufficient to guarantee invariant prediction.

Figure from Peters et al. (2016)

Theorem 3 (Connections to causal feature sets (Thm 1 in Zhang et al. [36])). *If we partition observations using the bisimulation metric, those clusters (a bisimulation partition) correspond to the causal feature set of the observation space with respect to current and future reward.*

Generalization to new observations



Walker Walk

Generalization to new reward functions

Theorem 4 (Task Generalization). *Given an encoder $\phi : \mathcal{S} \mapsto \mathcal{Z}$ that maps observations to a latent bisimulation metric representation where $||\phi(\mathbf{s}_i) - \phi(\mathbf{s}_j)||_2 := \tilde{d}(\mathbf{s}_i, \mathbf{s}_j)$, ϕ encodes information about all the causal ancestors of the reward $AN(R)$.*

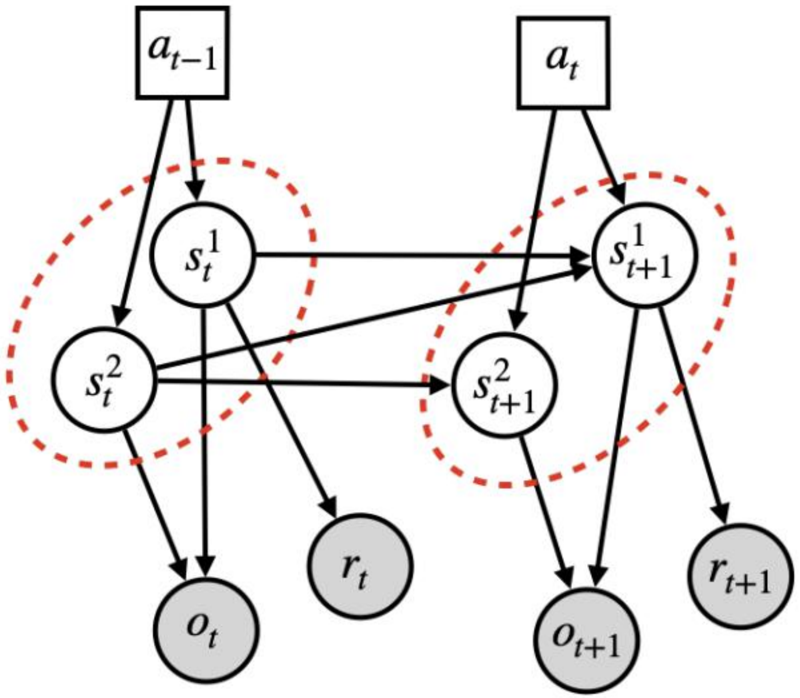
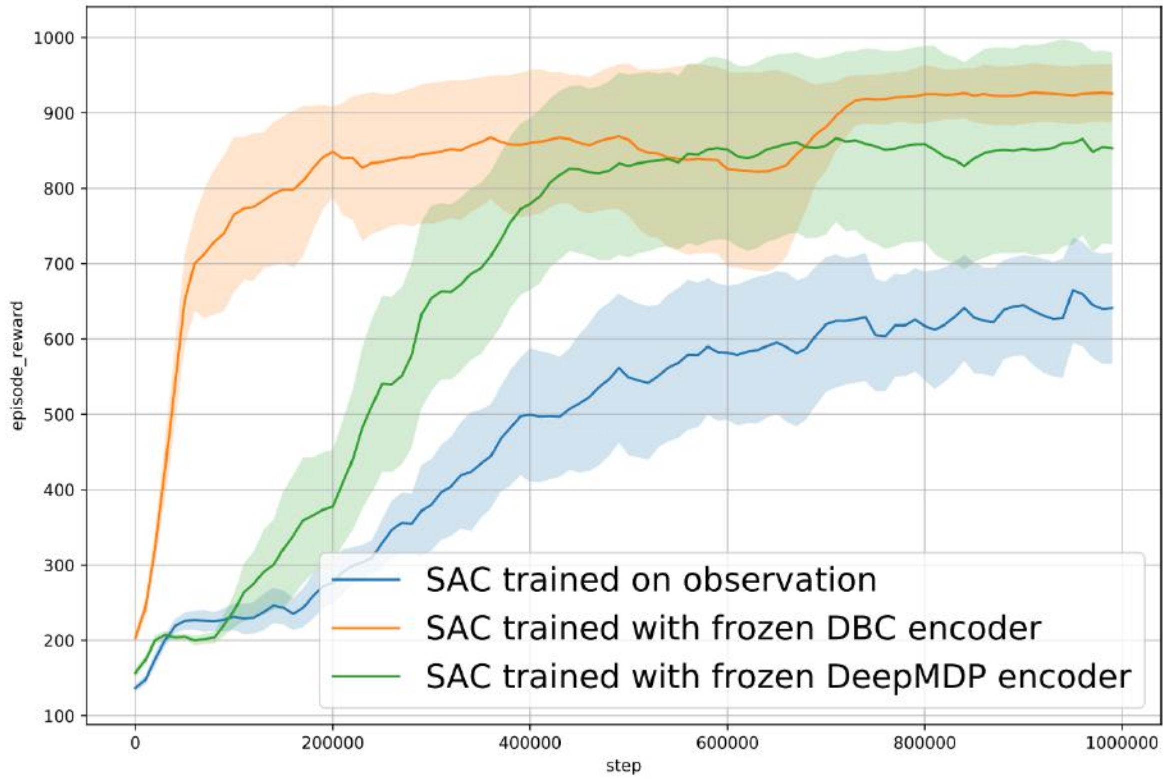


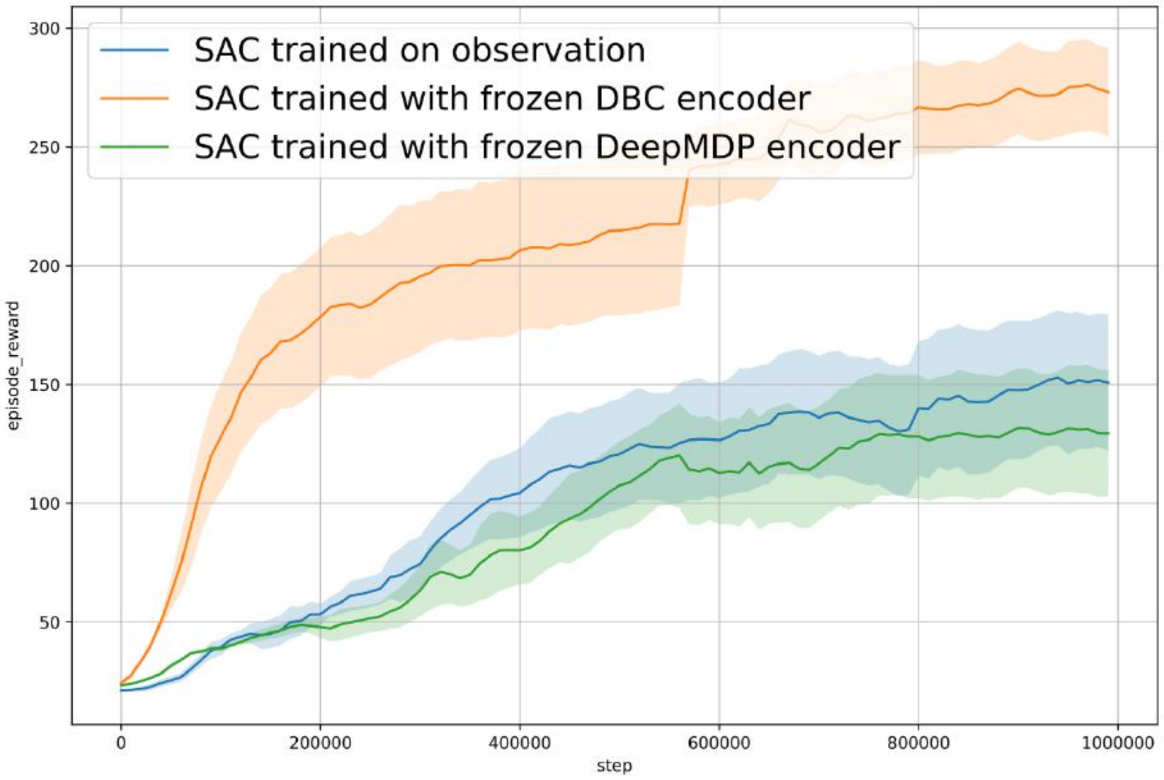
Figure 3: Causal graph of two time steps. Reward depends only on $\mathbf{s}^1$ as a causal parent, but $\mathbf{s}^1$ causally depends on $\mathbf{s}^2$, so AN(R) is the set $\{\mathbf{s}^1, \mathbf{s}^2\}$.

Generalization to new reward functions

Frozen encoders trained on Walker walk.



Walker stand



Walker run

Representation Learning with Bisimulation Metrics

CARLA highway with traffic

Representation Learning with Bisimulation Metrics

vehicle reward = highway progression (meters)
- collision penalty
+ throttle
- brake

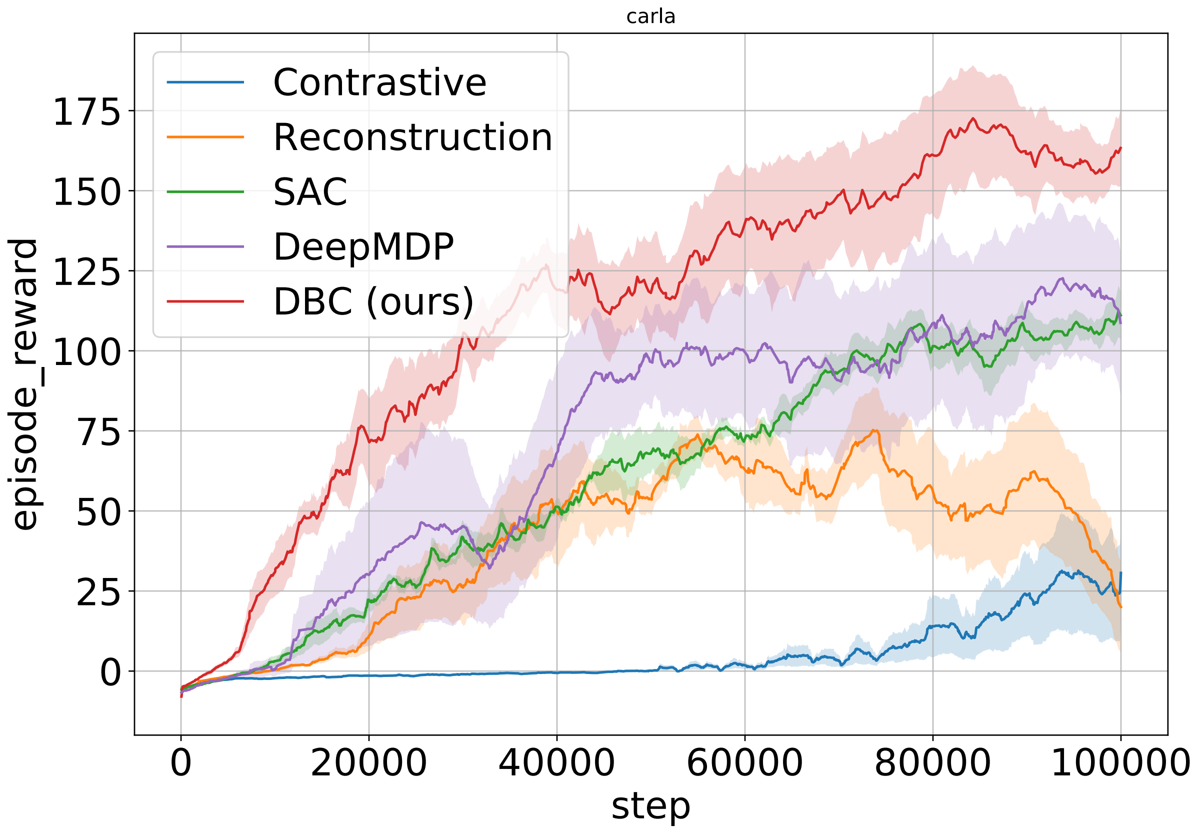
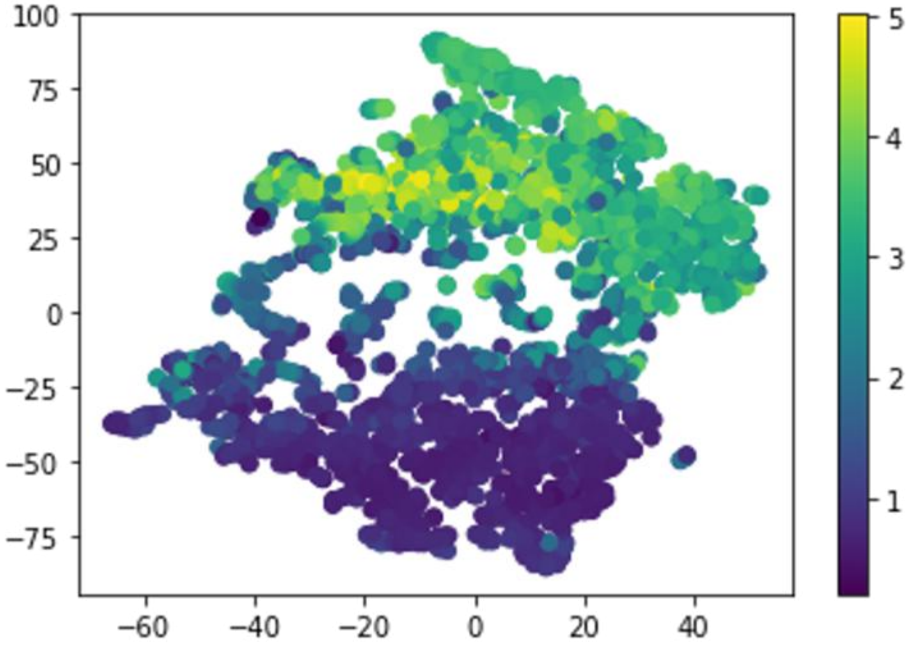
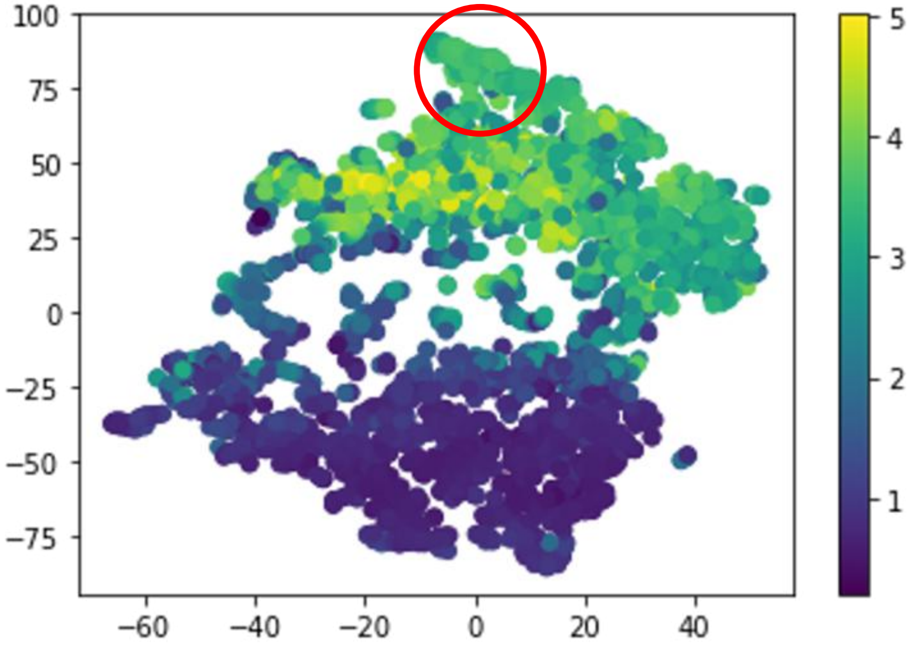
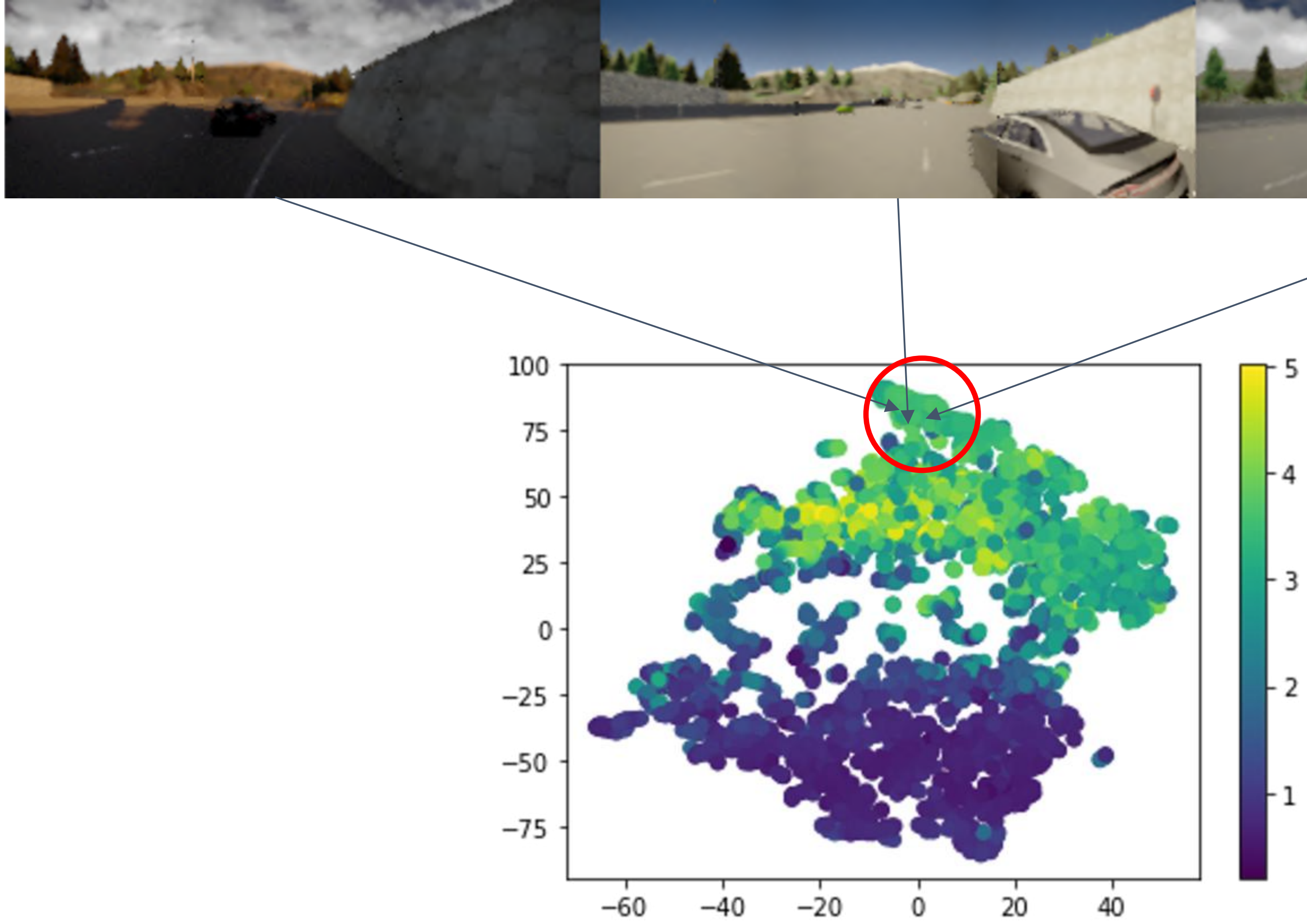


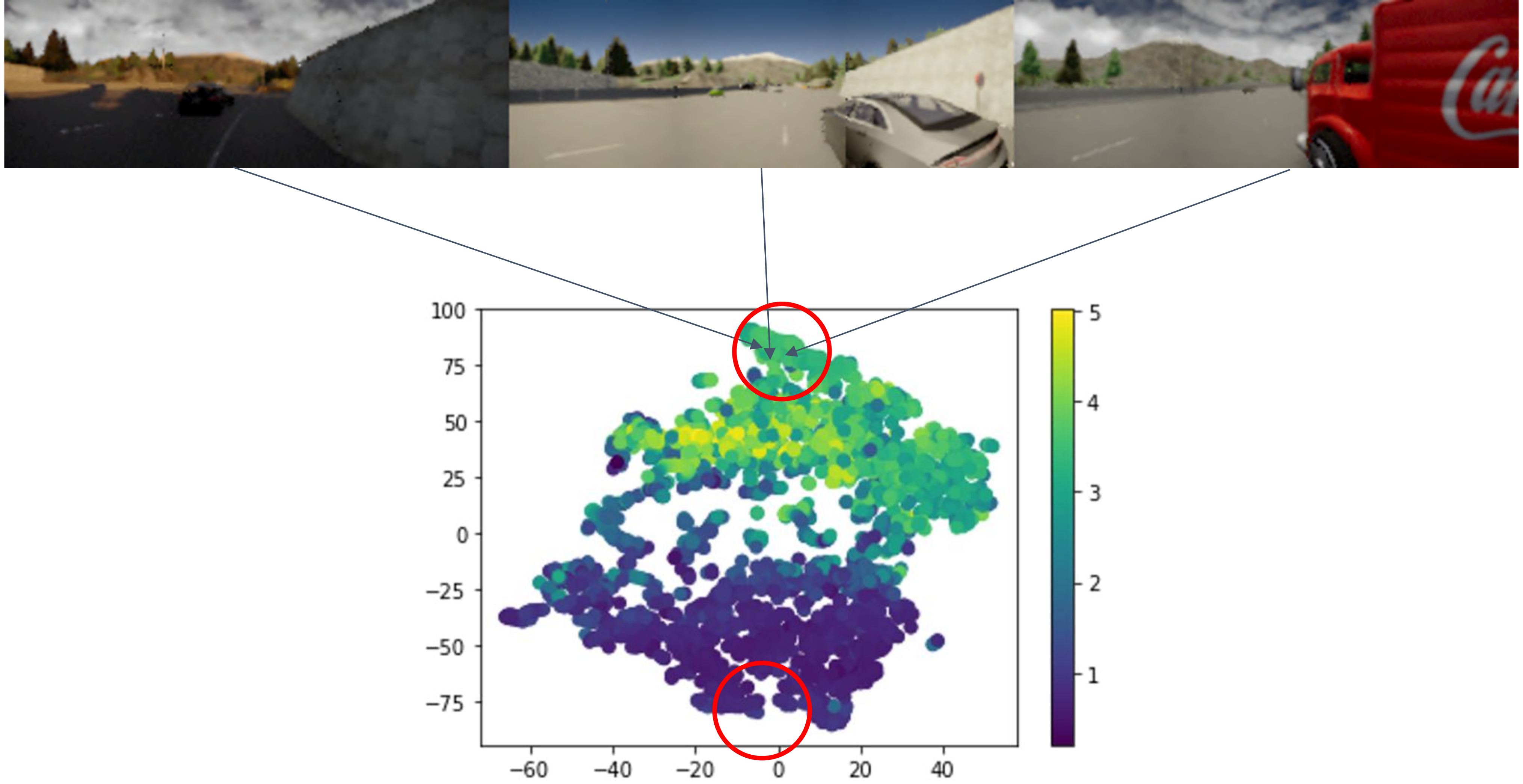
Table 1: Driving metrics, averaged over 100 episodes, after 100k training steps. Standard error shown. Arrow direction indicates if we desire the metric larger or smaller.

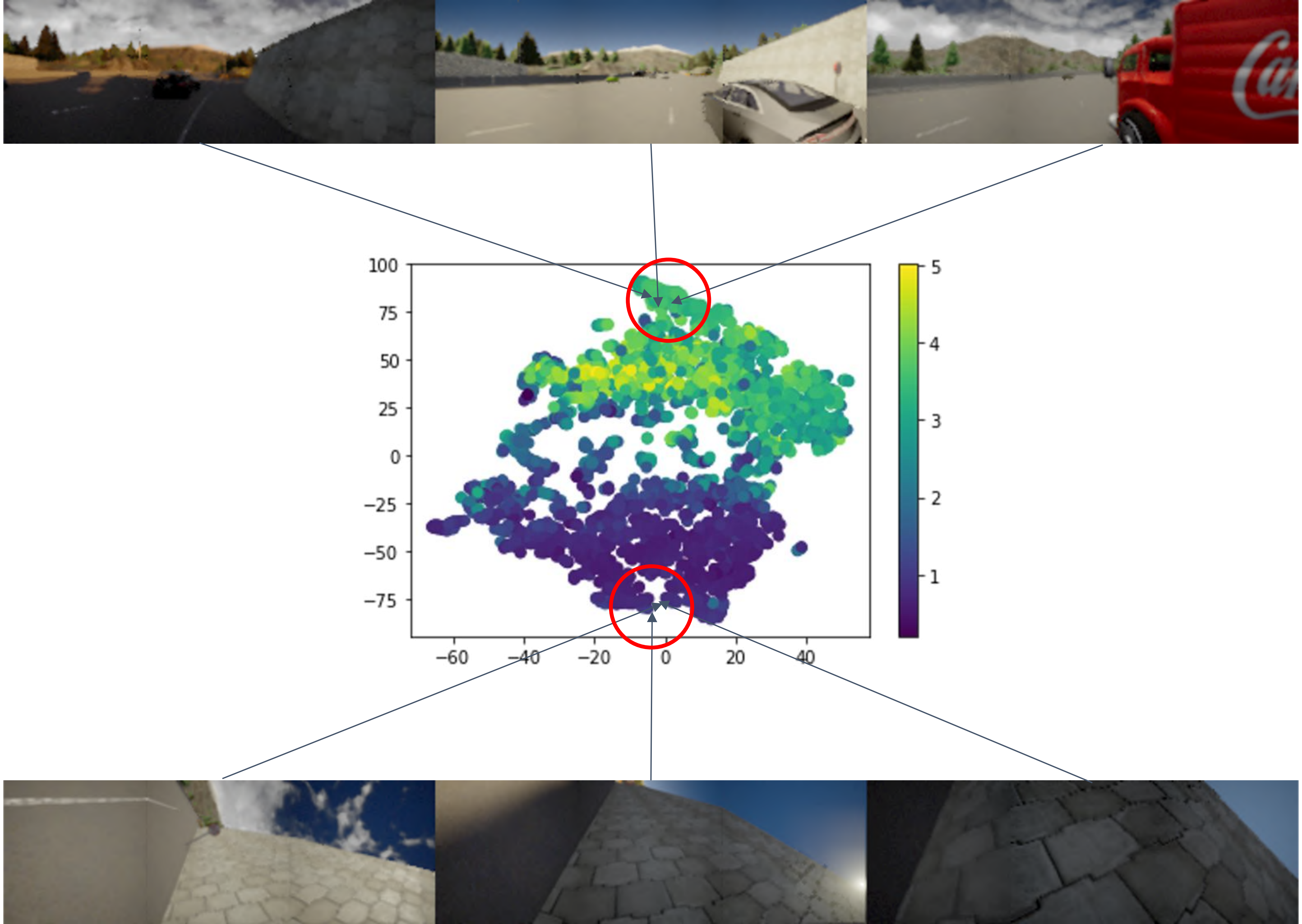
		SAC	DeepMDP	**DBC (ours)**
trials succeeded (100m)	$\uparrow$	12%	17%	**24%**
highway progression (m)	$\uparrow$	123.2 ± 7.43	106.7 ± 11.1	$\mathbf{179.0} \pm 11.4$
crash intensity	$\downarrow$	4604 ± 30.7	$\mathbf{1958} \pm 15.6$	2673 ± 38.5
average steer	$\downarrow$	$16.6\% \pm 0.019\%$	$10.4\% \pm 0.015\%$	$\mathbf{7.3\%} \pm 0.012\%$
average brake	$\downarrow$	$\mathbf{1.3\%} \pm 0.006\%$	$4.3\% \pm 0.033\%$	$1.6\% \pm 0.022\%$


100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1

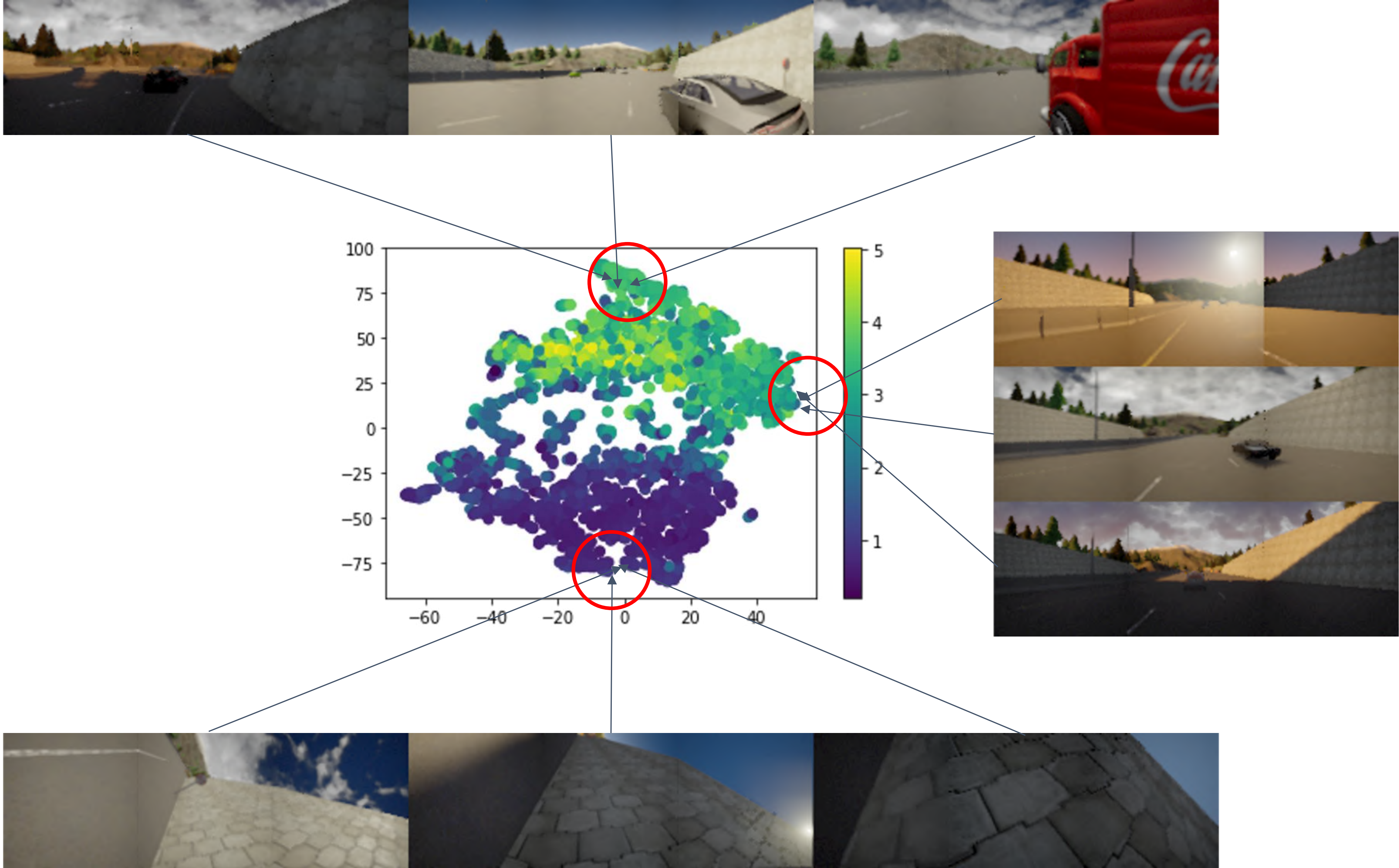

100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1


100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1

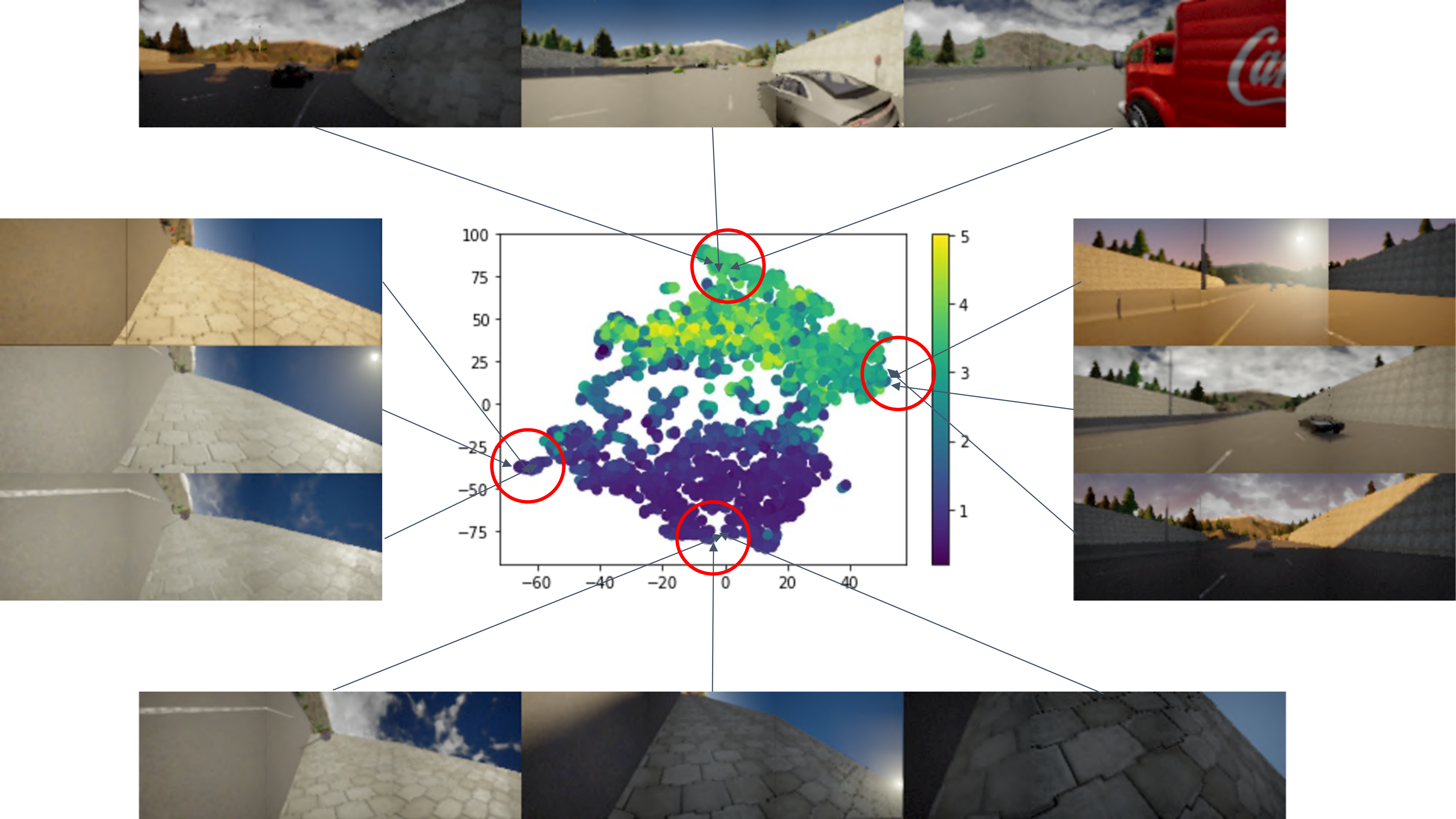

100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1



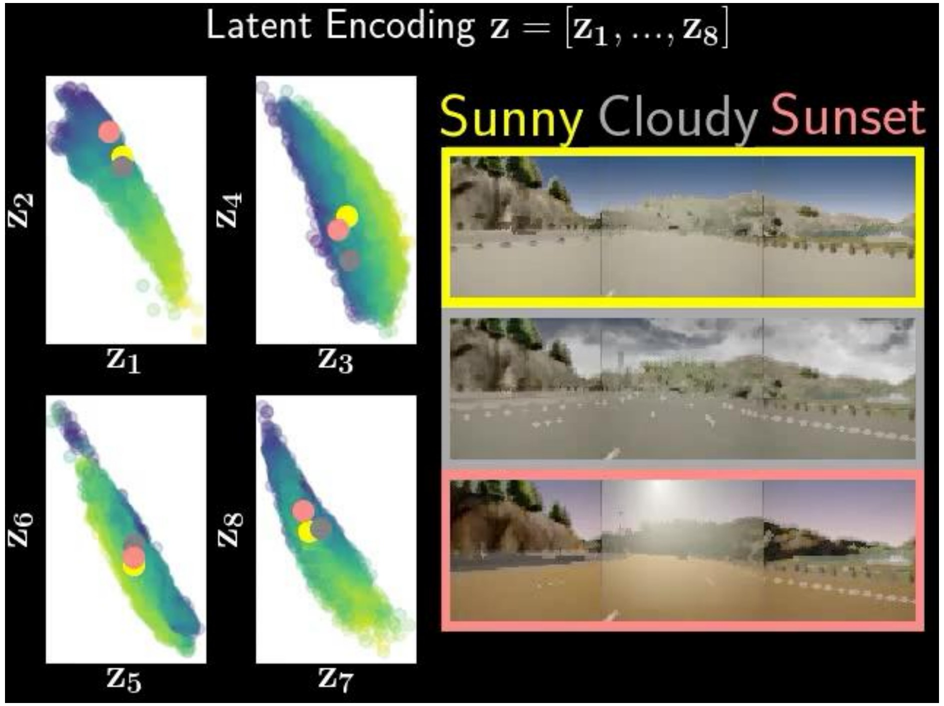
100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1



100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1



100
75
50
25
0
−25
−50
−75
−60
−40
−20
0
20
40
5
4
3
2
1



Mapping of latent encodings in different settings

DBC Agent POV during episode

Conclusions

- Goal was to learn lossy representations that only capture relevant information.
- We do this by learning a representation where L1 distance is bisimilarity between states.
- We show policy optimization on this representation improves generalization.

Arxiv: 2006.10742