DRIBO: Robust Deep Reinforcement Learning via Multi-View Information Bottleneck

Jiameng Fan and Wenchao Li

tl;dr: a robust representation learning approach for DRL to *extract only task-relevant features from raw pixels* based on the multi-view information bottleneck principle.

Train DRL agents that are robust to *task-irrelevant visual distractions*.

github.com/BU-DEPEND-Lab/DRIBO

What information is relevant for DRL?

Task-relevant visual details

What information is relevant for DRL?

Task-irrelevant visual details

Task-relevant visual details

What information is relevant for DRL?

temporally relevant visual details

Task-irrelevant visual details

Task-relevant visual details

What information is relevant for DRL?

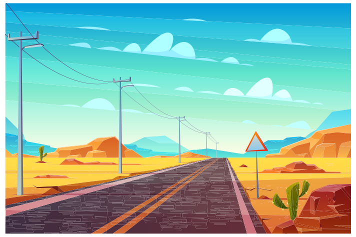
temporally relevant visual details

Task-irrelevant visual details

Task-relevant visual details

Goal: learn latent state representations that maximize *task-relevant information* while compressing away *task-irrelevant information*.

Sequential nature of RL



$\boldsymbol{o}_{t=1}$

…

$\boldsymbol{o}_{t=T}$

→ Encoder

$p_\theta(s_{1:T} | o_{1:T}, a_{1:T})$

→ $s_{1:T}$ →

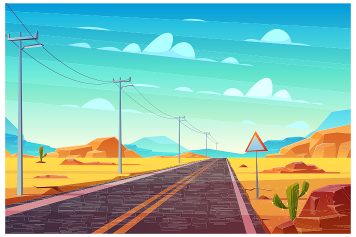
Reinforcement Learning Objective

Challenge: minimizing this term requires optimal actions to ***be known a priori***

$$I(O_{1:T}; S_{1:T}) = \underbrace{I(S_{1:T}; O_{1:T} | A^*_{1:T})}_{\text{Task-irrelevant}} + \underbrace{I(S_{1:T}; A^*_{1:T})}_{\text{Task-relevant}}$$

BU

Sequential multi-view observations in unsupervised settings

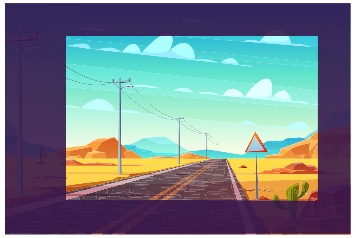


$o_{t=1}$

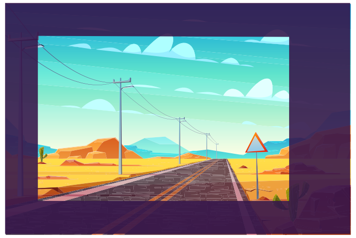
…

$o_{t=T}$

Sequential multi-view observations in unsupervised settings

 … 

$o_{t=1}^{(1)}$ $o_{t=T}^{(1)}$

 … 

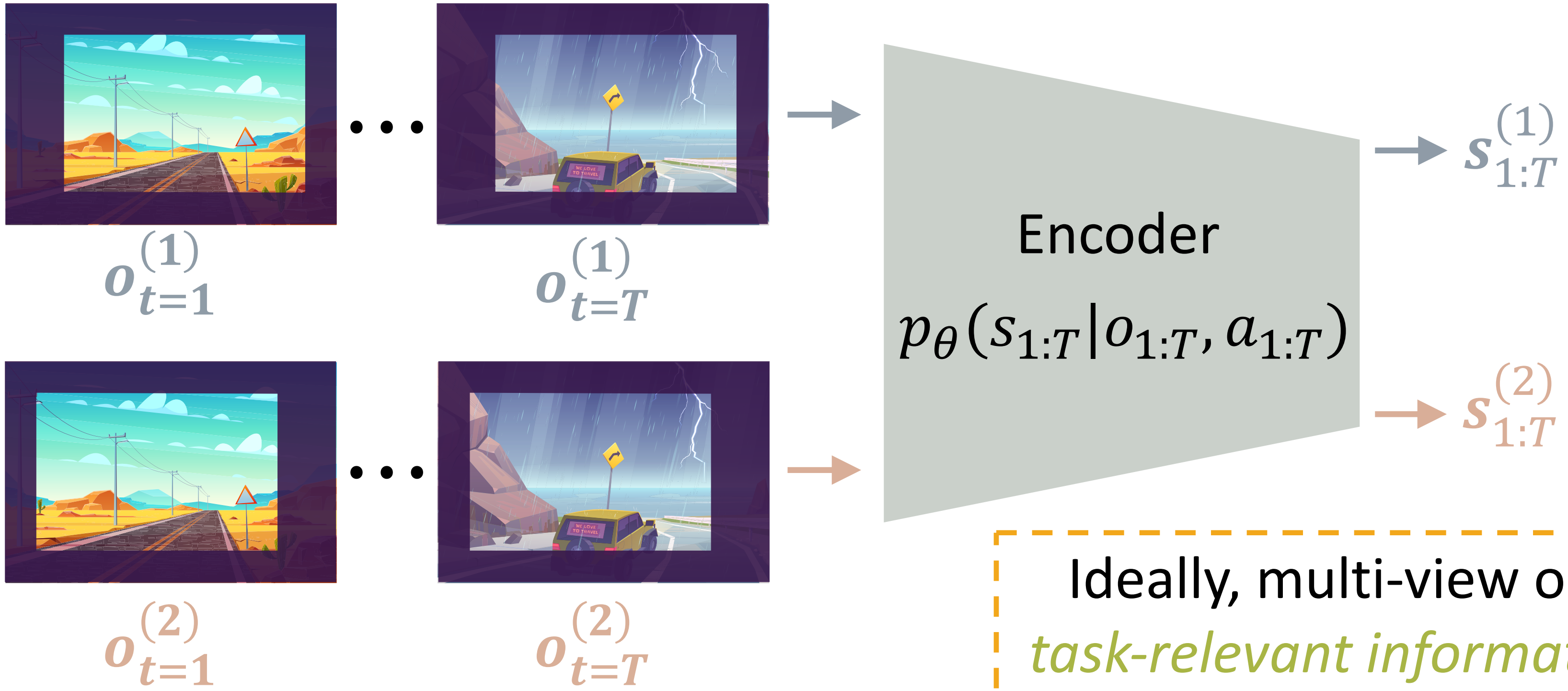
$o_{t=1}^{(2)}$ $o_{t=T}^{(2)}$

Random augmentations

Ideally, multi-view observations share *the same task-relevant information* while all the information *not shared by them is task-irrelevant*

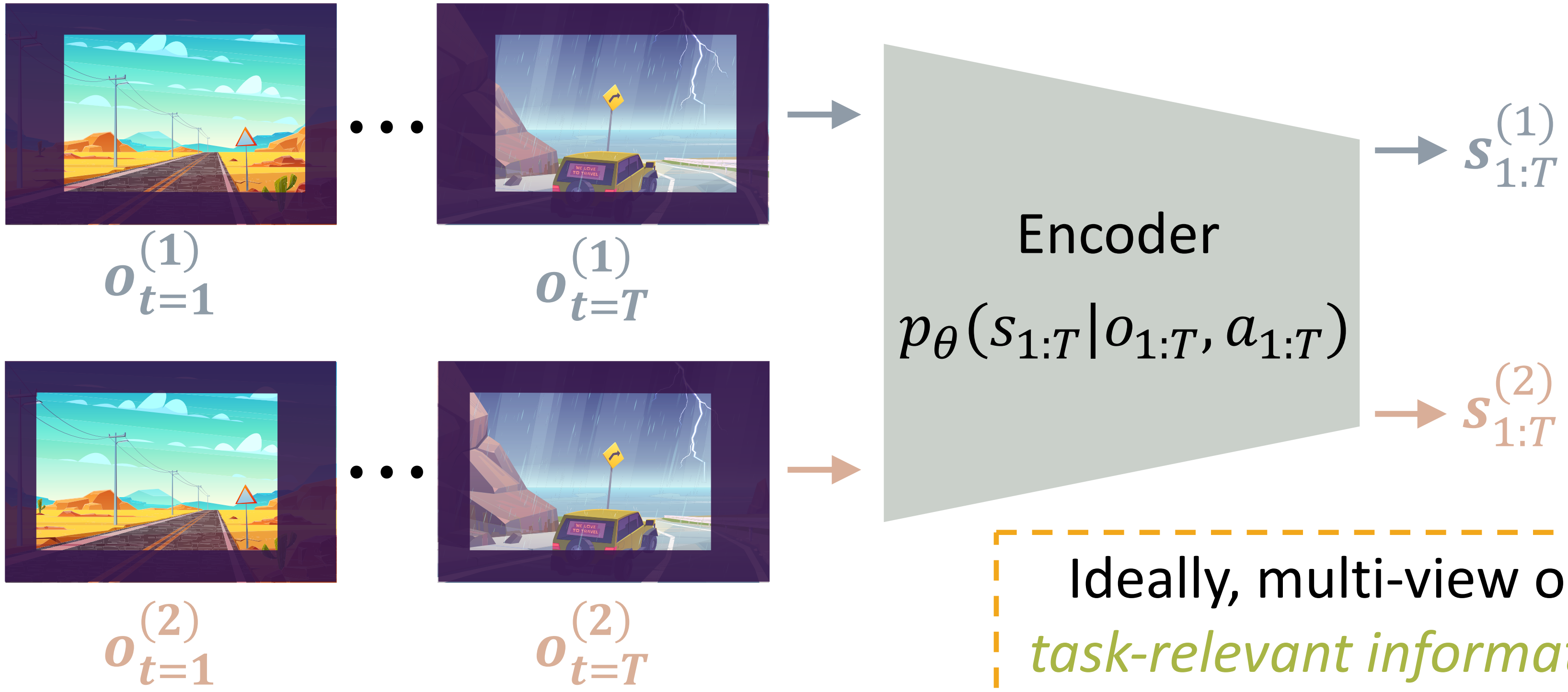
BU

Sequential multi-view observations in unsupervised settings



Ideally, multi-view observations share *the same task-relevant information* while all the information *not shared by them is task-irrelevant*

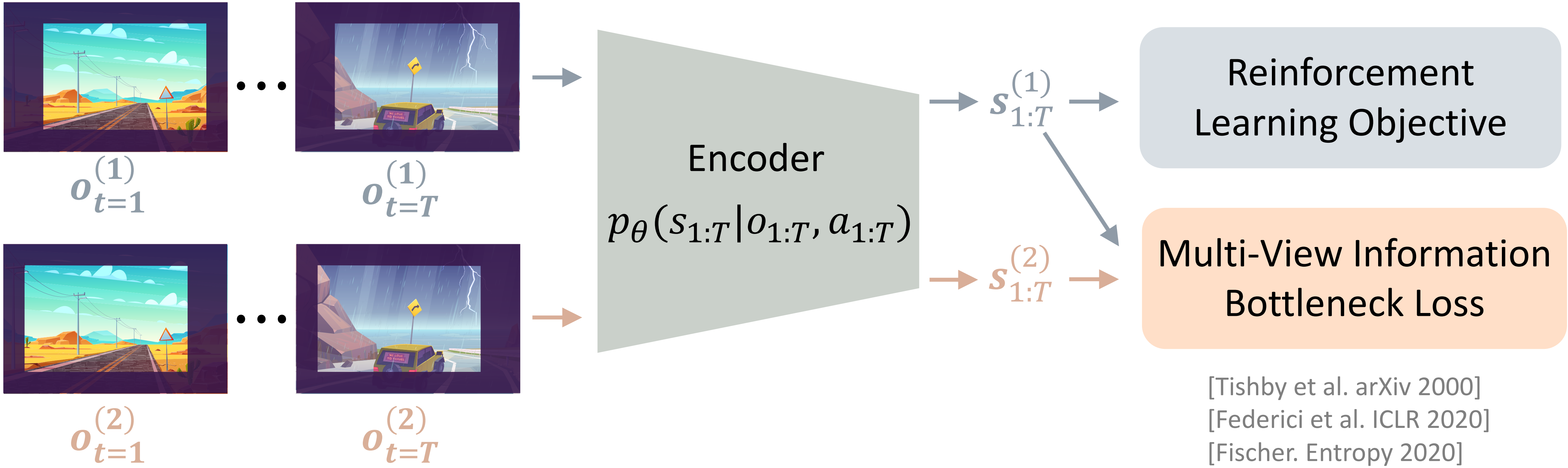
Sequential multi-view observations in unsupervised settings



Ideally, multi-view observations share *the same task-relevant information* while all the information *not shared by them is task-irrelevant*

$$I\left(S^{(1)}_{1:T}; O^{(1)}_{1:T} \middle| A_{1:T}\right) = \underbrace{I\left(S^{(1)}_{1:T}; O^{(1)}_{1:T} \middle| A_{1:T}, O^{(2)}_{1:T}\right)}_{\text{Task-irrelevant}} + \underbrace{I\left(O^{(2)}_{1:T}; S^{(1)}_{1:T} \middle| A_{1:T}\right)}_{\text{Task-relevant}}$$

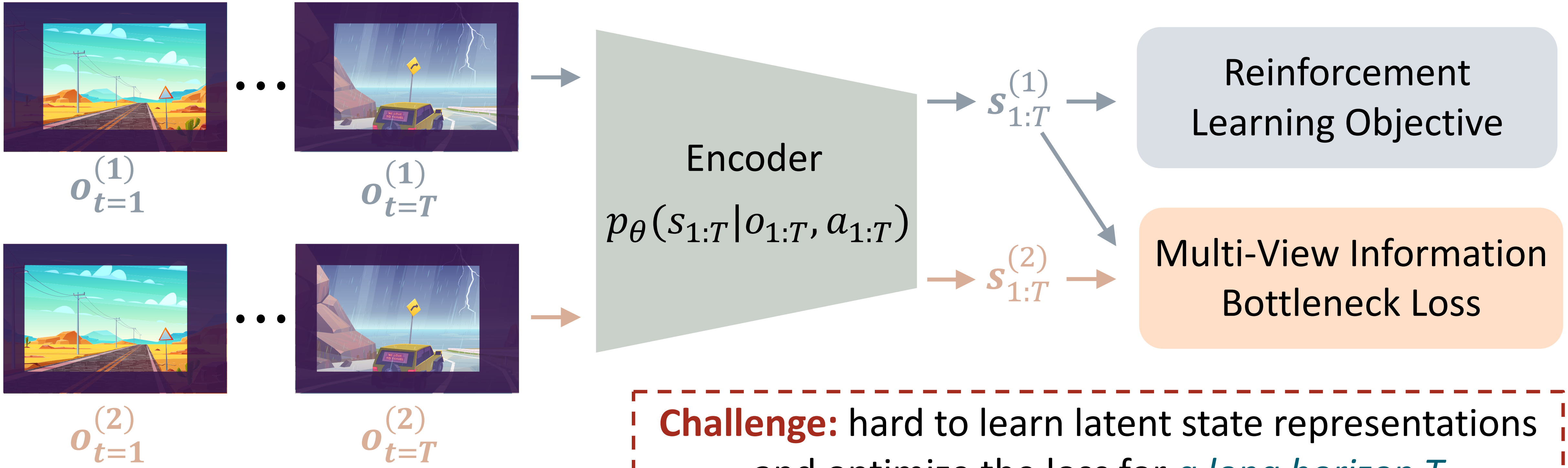
Sequential multi-view observations in unsupervised settings



[Tishby et al. arXiv 2000]
[Federici et al. ICLR 2020]
[Fischer. Entropy 2020]

$$I\left(S_{1:T}^{(1)}; O_{1:T}^{(1)} \middle| A_{1:T}\right) = \underbrace{I\left(S_{1:T}^{(1)}; O_{1:T}^{(1)} \middle| A_{1:T}, O_{1:T}^{(2)}\right)}_{\text{Task-irrelevant}} + \underbrace{I\left(O_{1:T}^{(2)}; S_{1:T}^{(1)} \middle| A_{1:T}\right)}_{\text{Task-relevant}}$$

Sequential multi-view observations in unsupervised settings



Challenge: hard to learn latent state representations and optimize the loss for *a long horizon T*

$$I\left(S^{(1)}_{1:T}; O^{(1)}_{1:T} \middle| A_{1:T}\right) = \underbrace{I\left(S^{(1)}_{1:T}; O^{(1)}_{1:T} \middle| A_{1:T}, O^{(2)}_{1:T}\right)}_{\text{Task-irrelevant}} + \underbrace{I\left(O^{(2)}_{1:T}; S^{(1)}_{1:T} \middle| A_{1:T}\right)}_{\text{Task-relevant}}$$

BU

DRIBO: robust deep reinforcement learning

- *Lower bound* of the sequential mutual information（**Theorem 1**）

$$I(S_{1:T}; O_{1:T} | A_{1:T}) \geq \sum_{t=1}^{T} I(S_t; O_t | S_{t-1}, A_{t-1})$$

- *Multi-view information bottleneck loss*

$$\mathcal{L}_{IB}^{(1)} = -\sum_{t} \left(I\left(S_t^{(1)}; O_t^{(1)} \middle| S_{t-1}^{(1)}, A_{t-1}, O_t^{(2)}\right) - \lambda_1 I\left(O_t^{(2)}; S_t^{(1)} \middle| S_{t-1}^{(1)}, A_{t-1}\right)\right)$$

$$\mathcal{L}_{IB}^{(2)} = -\sum_{t} \left(I\left(S_t^{(2)}; O_t^{(2)} \middle| S_{t-1}^{(2)}, A_{t-1}, O_t^{(1)}\right) - \lambda_2 I\left(O_t^{(1)}; S_t^{(2)} \middle| S_{t-1}^{(2)}, A_{t-1}\right)\right)$$

DRIBO: robust deep reinforcement learning

- *Lower bound* of the sequential mutual information（**Theorem 1**）

$$I(S_{1:T}; O_{1:T} | A_{1:T}) \geq \sum_{t=1}^{T} I(S_t; O_t | S_{t-1}, A_{t-1})$$

- *Multi-view information bottleneck loss*

$$\mathcal{L}_{IB}^{(1)} = -\sum_t \left(I\left(S_t^{(1)}; O_t^{(1)} \middle| S_{t-1}^{(1)}, A_{t-1}, O_t^{(2)}\right) - \lambda_1 I\left(O_t^{(2)}; S_t^{(1)} \middle| S_{t-1}^{(1)}, A_{t-1}\right)\right)$$

$$\mathcal{L}_{IB}^{(2)} = -\sum_t \left(I\left(S_t^{(2)}; O_t^{(2)} \middle| S_{t-1}^{(2)}, A_{t-1}, O_t^{(1)}\right) - \lambda_2 I\left(O_t^{(1)}; S_t^{(2)} \middle| S_{t-1}^{(2)}, A_{t-1}\right)\right)$$

Compress away task-irrelevant information

DRIBO: robust deep reinforcement learning

DIRBO loss: $\mathcal{L}_t(\theta;\beta) = -I_\theta\left(S_t^{(1)}; S_t^{(2)} | S_{t-1}, A_{t-1}\right)$

$$+\beta D_{SKL}\left(p_\theta(s_t^{(1)}|o_t^{(1)}, s_{t-1}^{(1)}, a_{t-1}) || p_\theta(s_t^{(2)}|o_t^{(2)}, s_{t-1}^{(2)}, a_{t-1})\right)$$

BU

DRIBO: robust deep reinforcement learning

DIRBO loss: $\mathcal{L}_t(\theta;\beta) = -I_\theta\left(S_t^{(1)}; S_t^{(2)} | S_{t-1}, A_{t-1}\right)$

Shared information

Divergence between the representations

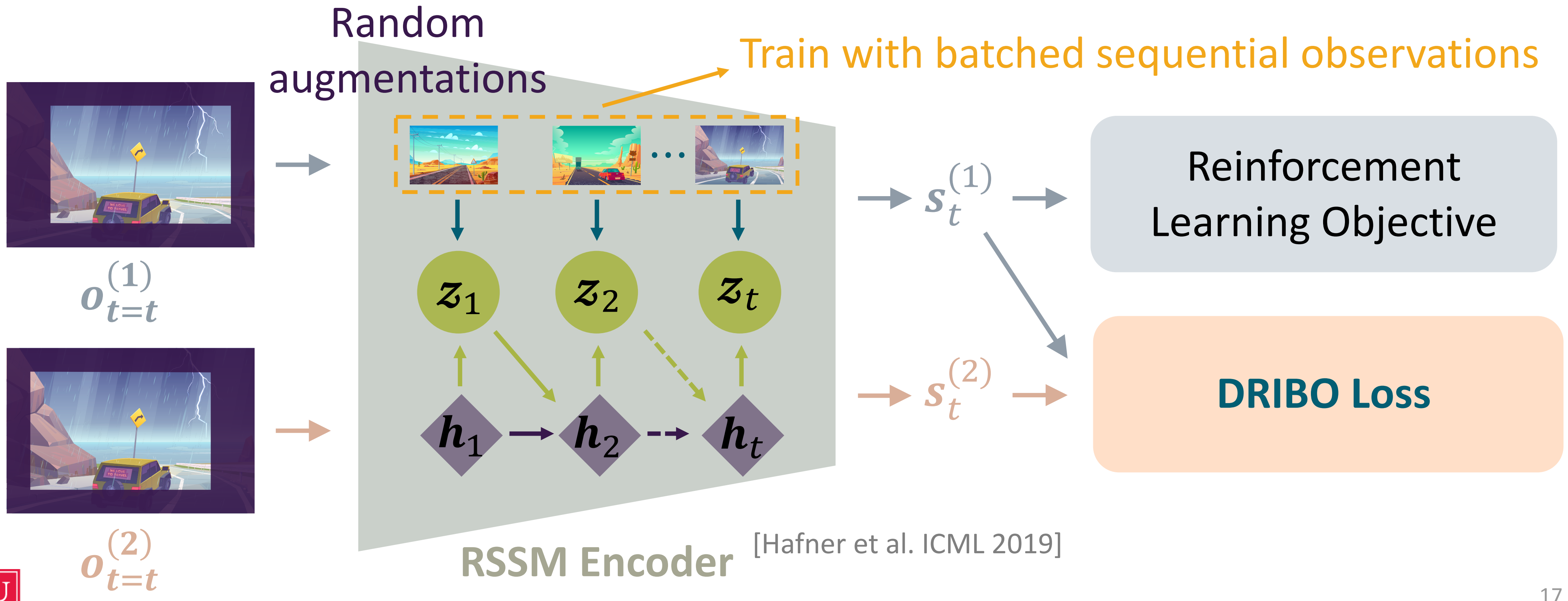
$$+\beta D_{SKL}\left(p_\theta(s_t^{(1)}|o_t^{(1)}, s_{t-1}^{(1)}, a_{t-1}) || p_\theta(s_t^{(2)}|o_t^{(2)}, s_{t-1}^{(2)}, a_{t-1})\right)$$

BU

DRIBO: robust deep reinforcement learning

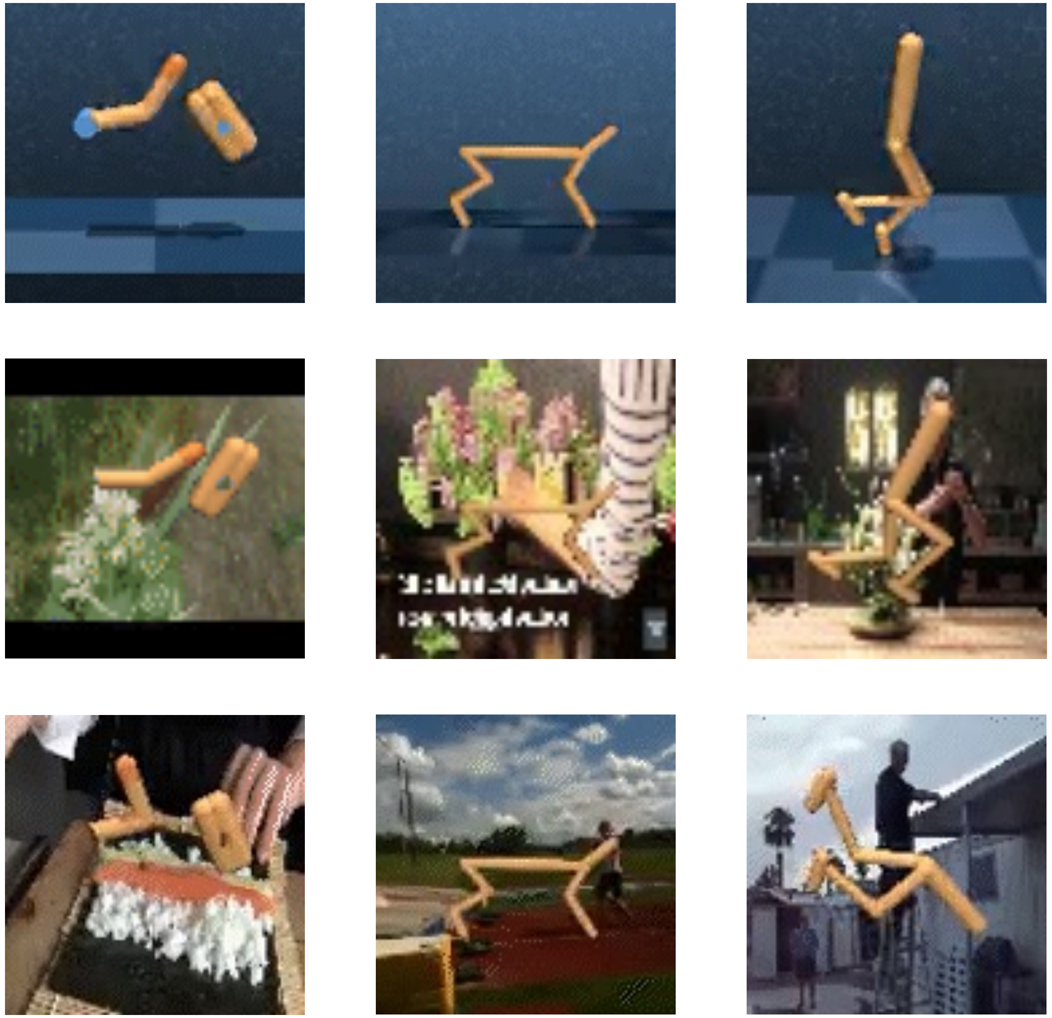
DIRBO loss: $\mathcal{L}_t(\theta;\beta) = -I_\theta\left(S_t^{(1)}; S_t^{(2)} | S_{t-1}, A_{t-1}\right)$ Shared information

$+\beta D_{SKL}\left(p_\theta(s_t^{(1)}|o_t^{(1)}, s_{t-1}^{(1)}, a_{t-1}) || p_\theta(s_t^{(2)}|o_t^{(2)}, s_{t-1}^{(2)}, a_{t-1})\right)$ Divergence between the representations



DRIBO results: robustness against visual distractions

[Tassa et al. arXiv 2018]
[Kay et al. arXiv 2017]
[Zhang et al. arXiv 2018]



Clean setting (no background change)

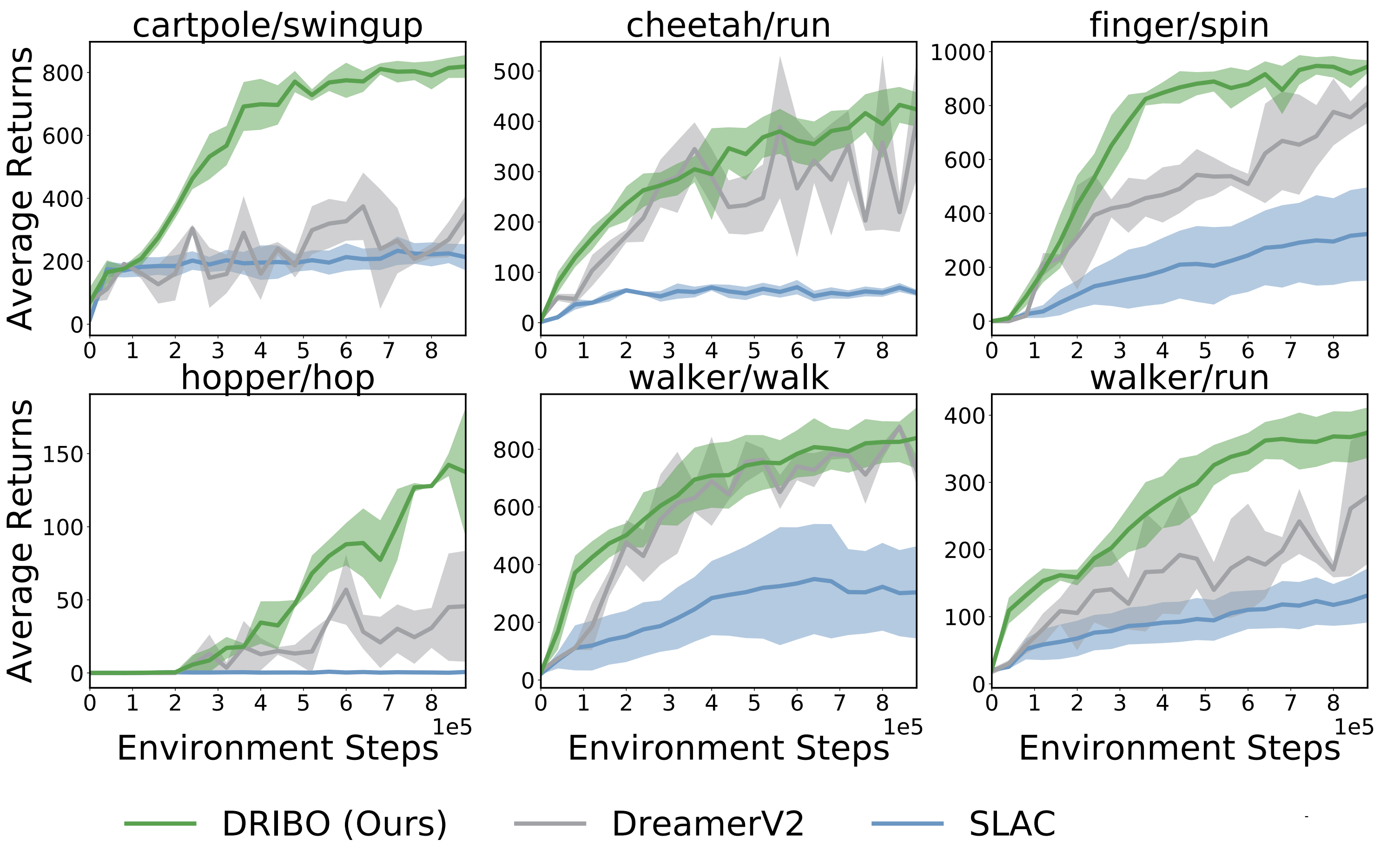
“Arranging flowers” natural video setting (training)

natural video setting (testing)

BU

DeepMind Control Suite

DRIBO results: robustness against visual distractions

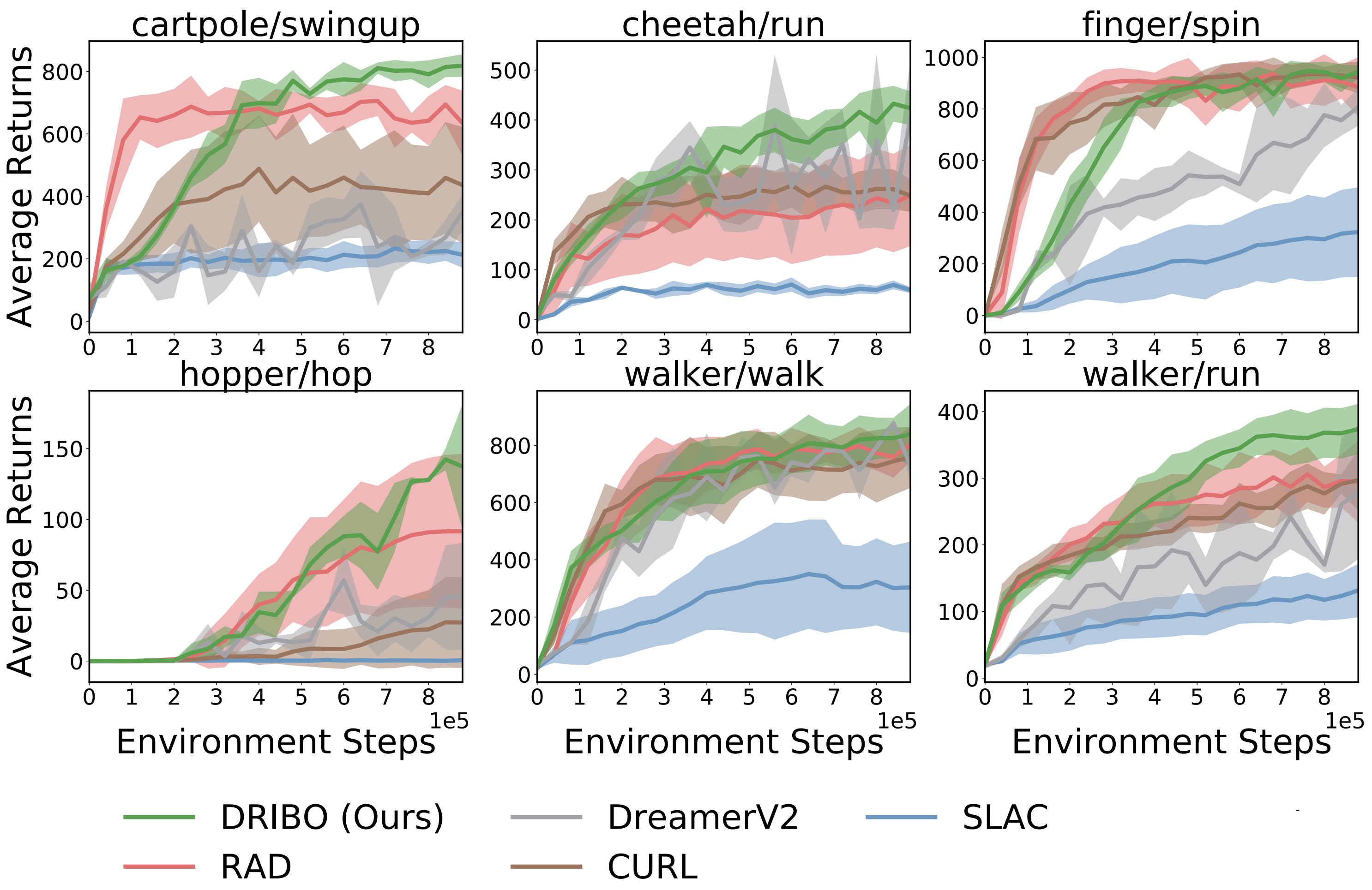


- Averaged 68% improvement compared with reconstruction-based methods (DreamerV2, SLAC)

[Hafner et al. ICLR 2021] [Lee et al. NeurIPS 2020]

BU

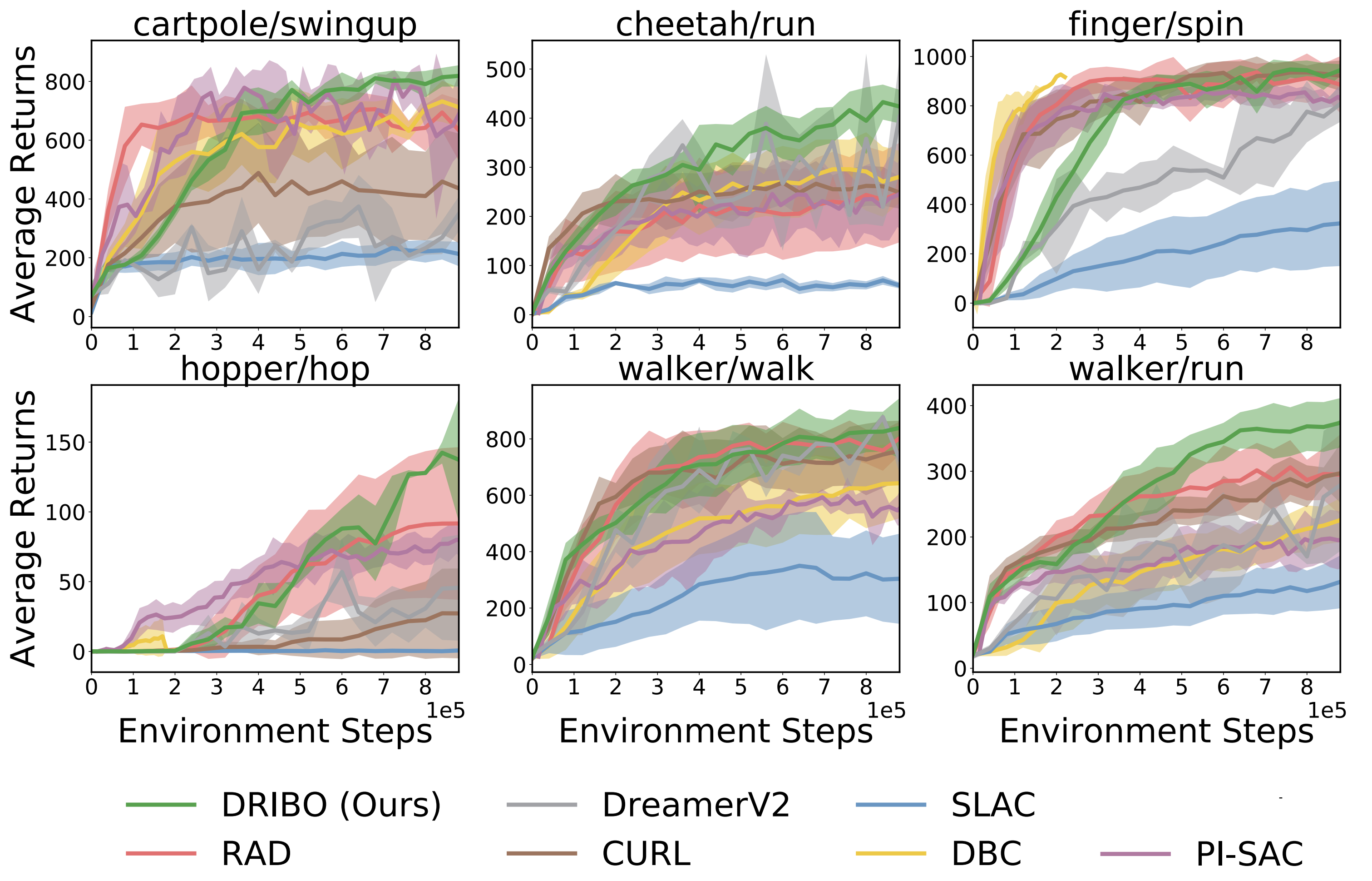
DRIBO results: robustness against visual distractions



- Averaged 68% improvement compared with reconstruction-based methods (DreamerV2, SLAC)
- Averaged 31% improvement compared with RAD and CURL

[Laskin et al. NeurIPS 2020] [Laskin et al. ICML 2020]

DRIBO results: robustness against visual distractions



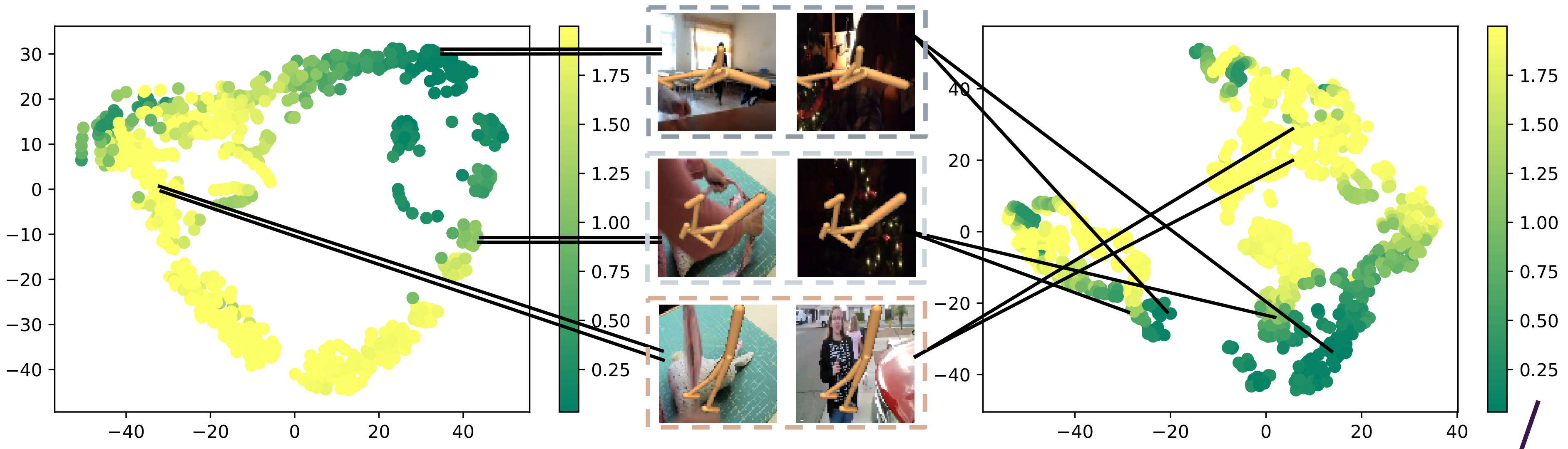
- Averaged 68% improvement compared with reconstruction-based methods (SLAC, DreamerV2)
- Averaged 31% improvement compared with RAD and CURL
- Averaged 41% improvement compared with DBC and PI-SAC which also explicitly compress away task-irrelevant information

[Zhang et al. ICLR 2021]

[Lee et al. NeurIPS 2020]

BU

DRIBO results: robustness against visual distractions



t-SNE of **DRIBO**

t-SNE of **CURL**

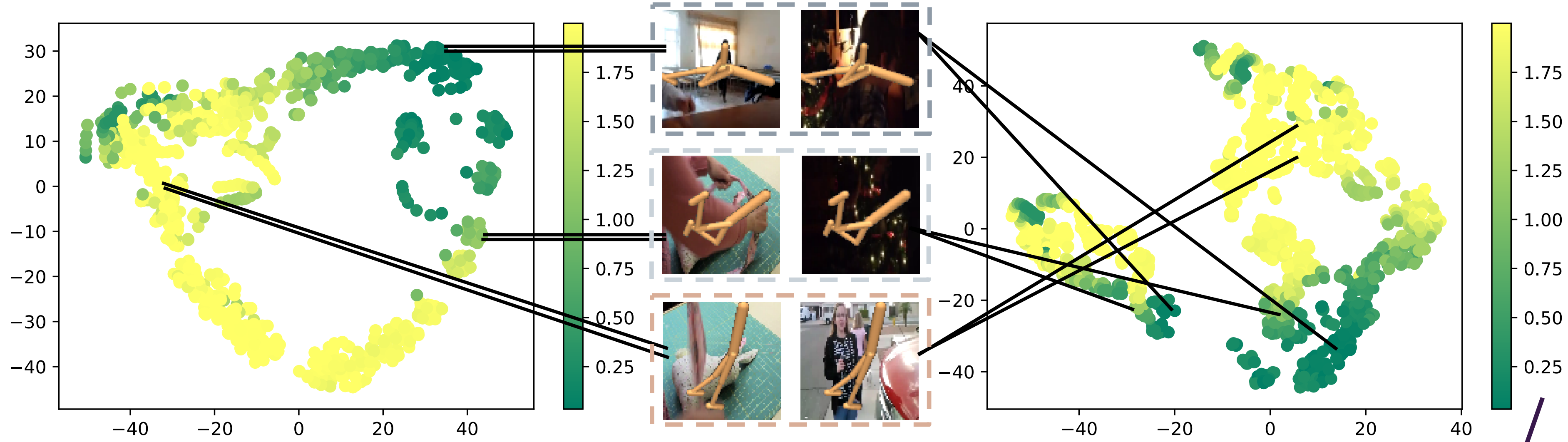
Reward value

[Van der Maaten et al. JML 2008]

Identical foregrounds but different backgrounds

DRIBO learns latent states that *are neighboring in the embedding space with similar reward values*.

DRIBO results: robustness against visual distractions



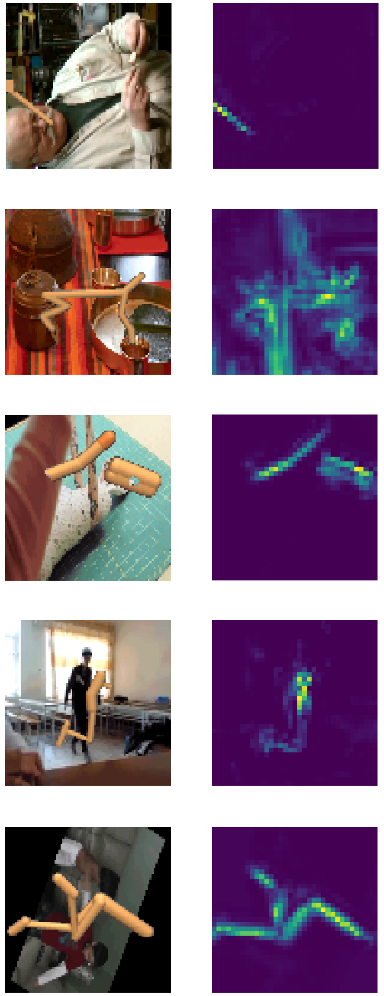
t-SNE of DRIBO

t-SNE of CURL

Reward value

[Van der Maaten et al. JML 2008]

Identical foregrounds but different backgrounds



Spatial attention maps

[Zagoruyko et al. ICLR 2017]

DRIBO learns encoders that *focus on the robots' body* and *ignore irrelevant visual details in the background*.

DRIBO results: generalization to unseen environments

Procgen: agents are *trained on the first 200 levels* and *evaluated on unseen levels during testing*.

[Cobbe et al. ICML 2020]

BU

DRIBO results: generalization to unseen environments

Env	PPO	RAD	DrAC	UCB-DrAC	DAAC	IDAAC	DRIBO
BigFish	4.0 ± 1.2	9.9 ± 1.7	8.7 ± 1.4	9.7 ± 1.0	17.8 ± 1.4	**18.5 ± 1.2**	10.9 ± 1.6
StarPilot	24.7 ± 3.4	33.4 ± 5.1	29.5 ± 5.4	30.2 ± 2.8	36.4 ± 2.8	**37.0 ± 2.3**	36.5 ± 3.0
FruitBot	26.7 ± 0.8	27.3 ± 1.8	28.2 ± 0.8	28.3 ± 0.9	28.6 ± 0.6	27.9 ± 0.5	**30.8 ± 0.8**
BossFight	7.7 ± 1.0	7.9 ± 0.6	7.5 ± 0.8	8.3 ± 0.8	9.6 ± 0.5	9.8 ± 0.6	**12.0 ± 0.5**
Ninja	5.9 ± 0.7	6.9 ± 0.8	7.0 ± 0.4	6.9 ± 0.6	6.8 ± 0.4	6.8 ± 0.4	**9.7 ± 0.7**
Plunder	5.0 ± 0.5	8.5 ± 1.2	9.5 ± 1.0	8.9 ± 1.0	20.7 ± 3.3	**23.3 ± 1.4**	5.8 ± 1.0
CaveFlyer	5.1 ± 0.9	5.1 ± 0.6	6.3 ± 0.8	5.3 ± 0.9	4.6 ± 0.2	5.0 ± 0.6	**7.5 ± 1.0**
CoinRun	8.5 ± 0.5	9.0 ± 0.8	8.8 ± 0.2	8.5 ± 0.6	9.2 ± 0.2	**9.4 ± 0.1**	9.2 ± 0.7
Jumper	5.8 ± 0.5	6.5 ± 0.6	6.6 ± 0.4	6.4 ± 0.6	6.5 ± 0.4	6.3 ± 0.2	**8.4 ± 1.6**
Chaser	5.0 ± 0.8	5.9 ± 1.0	5.7 ± 0.6	6.7 ± 0.6	6.6 ± 1.2	**6.8 ± 1.0**	4.8 ± 0.8
Climber	5.7 ± 0.8	6.9 ± 0.8	7.1 ± 0.7	6.5 ± 0.8	7.8 ± 0.2	**8.3 ± 0.4**	8.1 ± 1.6
DodgeBall	**11.7 ± 0.3**	2.8 ± 0.7	4.3 ± 0.8	4.7 ± 0.7	3.3 ± 0.5	3.3 ± 0.3	3.8 ± 0.9
Heist	2.4 ± 0.5	4.1 ± 1.0	4.0 ± 0.8	4.0 ± 0.7	3.3 ± 0.2	3.5 ± 0.2	**7.7 ± 1.6**
Leaper	4.9 ± 0.7	4.3 ± 1.0	5.3 ± 1.1	5.0 ± 0.3	7.3 ± 1.1	**7.7 ± 1.0**	5.3 ± 1.5
Maze	5.7 ± 0.6	6.1 ± 1.0	6.6 ± 0.8	6.3 ± 0.6	5.5 ± 0.2	5.6 ± 0.3	**8.5 ± 1.6**
Miner	8.5 ± 0.5	9.4 ± 1.2	**9.8 ± 0.6**	9.7 ± 0.7	8.6 ± 0.9	9.5 ± 0.4	**9.8 ± 0.9**

Procgen: agents are *trained on the first 200 levels* and *evaluated on unseen levels during testing*.

DRIBO achieves better performance in *13 of the 16 games compared with augmentation-based methods*.

[Cobbe et al. ICML 2020]

[Laskin et al. NeurIPS 2020] [Raileanu et al. NeurIPS 2021]

DRIBO results: generalization to unseen environments

Env	PPO	RAD	DrAC	UCB-DrAC	DAAC	IDAAC	DRIBO
BigFish	4.0 ± 1.2	9.9 ± 1.7	8.7 ± 1.4	9.7 ± 1.0	17.8 ± 1.4	**18.5 ± 1.2**	10.9 ± 1.6
StarPilot	24.7 ± 3.4	33.4 ± 5.1	29.5 ± 5.4	30.2 ± 2.8	36.4 ± 2.8	**37.0 ± 2.3**	36.5 ± 3.0
FruitBot	26.7 ± 0.8	27.3 ± 1.8	28.2 ± 0.8	28.3 ± 0.9	28.6 ± 0.6	27.9 ± 0.5	**30.8 ± 0.8**
BossFight	7.7 ± 1.0	7.9 ± 0.6	7.5 ± 0.8	8.3 ± 0.8	9.6 ± 0.5	9.8 ± 0.6	**12.0 ± 0.5**
Ninja	5.9 ± 0.7	6.9 ± 0.8	7.0 ± 0.4	6.9 ± 0.6	6.8 ± 0.4	6.8 ± 0.4	**9.7 ± 0.7**
Plunder	5.0 ± 0.5	8.5 ± 1.2	9.5 ± 1.0	8.9 ± 1.0	20.7 ± 3.3	**23.3 ± 1.4**	5.8 ± 1.0
CaveFlyer	5.1 ± 0.9	5.1 ± 0.6	6.3 ± 0.8	5.3 ± 0.9	4.6 ± 0.2	5.0 ± 0.6	**7.5 ± 1.0**
CoinRun	8.5 ± 0.5	9.0 ± 0.8	8.8 ± 0.2	8.5 ± 0.6	9.2 ± 0.2	**9.4 ± 0.1**	9.2 ± 0.7
Jumper	5.8 ± 0.5	6.5 ± 0.6	6.6 ± 0.4	6.4 ± 0.6	6.5 ± 0.4	6.3 ± 0.2	**8.4 ± 1.6**
Chaser	5.0 ± 0.8	5.9 ± 1.0	5.7 ± 0.6	6.7 ± 0.6	6.6 ± 1.2	**6.8 ± 1.0**	4.8 ± 0.8
Climber	5.7 ± 0.8	6.9 ± 0.8	7.1 ± 0.7	6.5 ± 0.8	7.8 ± 0.2	**8.3 ± 0.4**	8.1 ± 1.6
DodgeBall	**11.7 ± 0.3**	2.8 ± 0.7	4.3 ± 0.8	4.7 ± 0.7	3.3 ± 0.5	3.3 ± 0.3	3.8 ± 0.9
Heist	2.4 ± 0.5	4.1 ± 1.0	4.0 ± 0.8	4.0 ± 0.7	3.3 ± 0.2	3.5 ± 0.2	**7.7 ± 1.6**
Leaper	4.9 ± 0.7	4.3 ± 1.0	5.3 ± 1.1	5.0 ± 0.3	7.3 ± 1.1	**7.7 ± 1.0**	5.3 ± 1.5
Maze	5.7 ± 0.6	6.1 ± 1.0	6.6 ± 0.8	6.3 ± 0.6	5.5 ± 0.2	5.6 ± 0.3	**8.5 ± 1.6**
Miner	8.5 ± 0.5	9.4 ± 1.2	**9.8 ± 0.6**	9.7 ± 0.7	8.6 ± 0.9	9.5 ± 0.4	**9.8 ± 0.9**

Procgen: agents are *trained on the first 200 levels* and *evaluated on unseen levels during testing*.

DRIBO achieves better performance in *9 of the 16 games* *compared with the SOTA method IDAAC*.

[Cobbe et al. ICML 2020]

[Raileanu et al. ICML 2021]

BU

Contributions

- We propose DRIBO, a new representation learning method that improves DRL agents' robustness to task-irrelevant visual distractions.
- State-of-the-art empirical results on *robustness against visual distractions* and *generalization performance*.

Thank you!

github.com/BU-DEPEND-Lab/DRIBO

BU