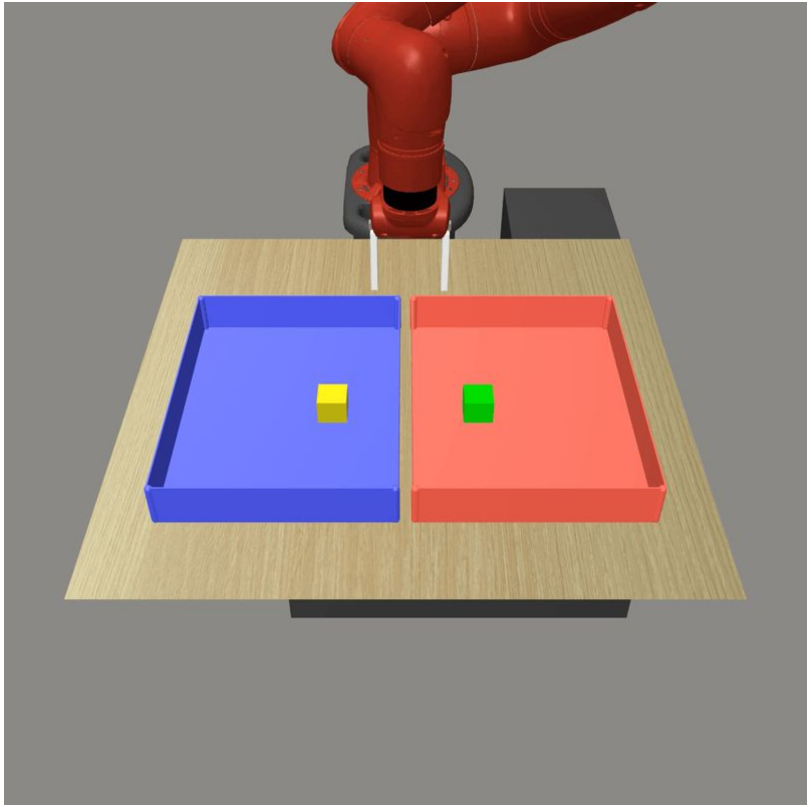
Can a MISL Fly?
Analysis and Ingredients for Mutual Information Skill Learning

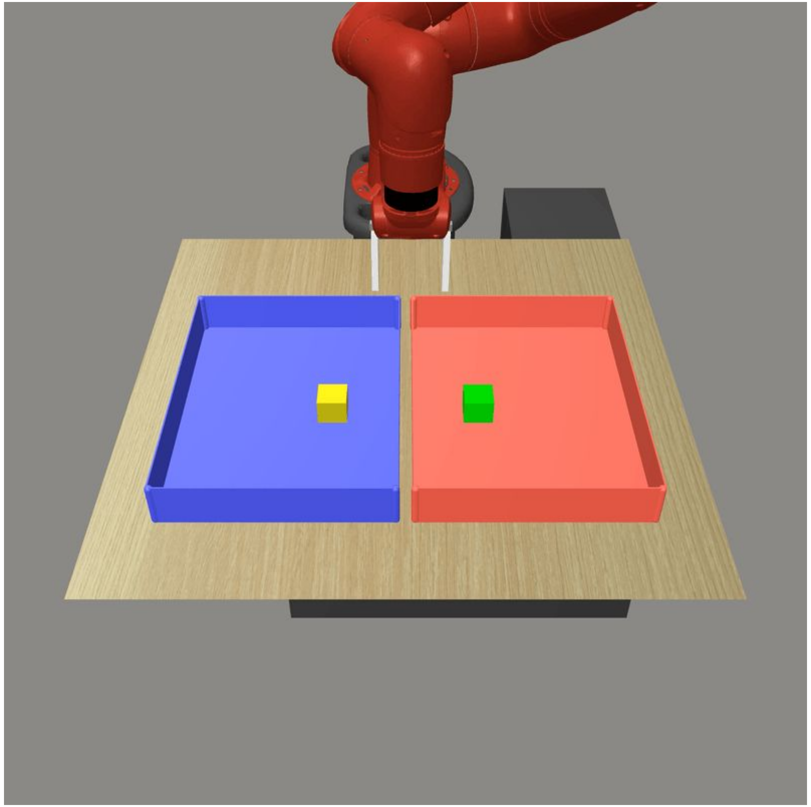
Chongyi Zheng*, **Jens Tuyls***, Joanne Peng,
Benjamin Eysenbach

{chongyiz, jtuyls}@princeton.edu

PRINCETON
UNIVERSITY

Our work: unsupervised pre-training for RL (demo)

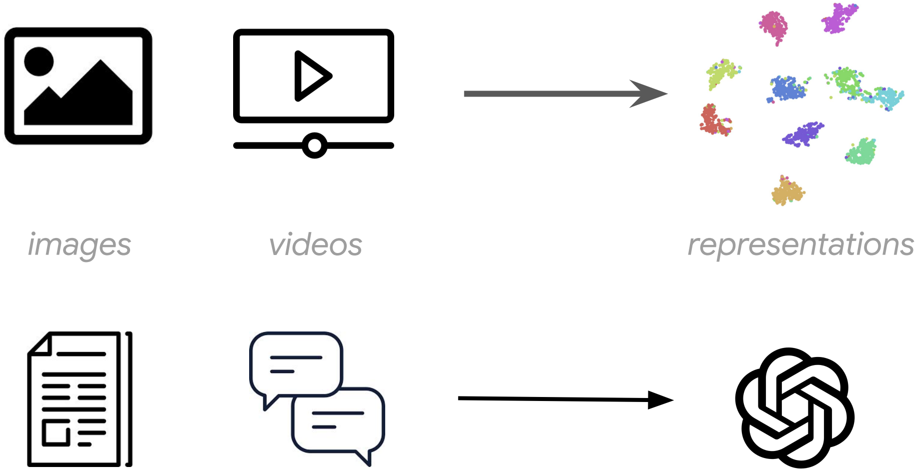




Unsupervised pre-training has proven successful in CV and NLP.

Unsupervised Pre-training

Can we do this for RL? ✓



images *videos* → *representations*

articles *dialogues* → *language models*

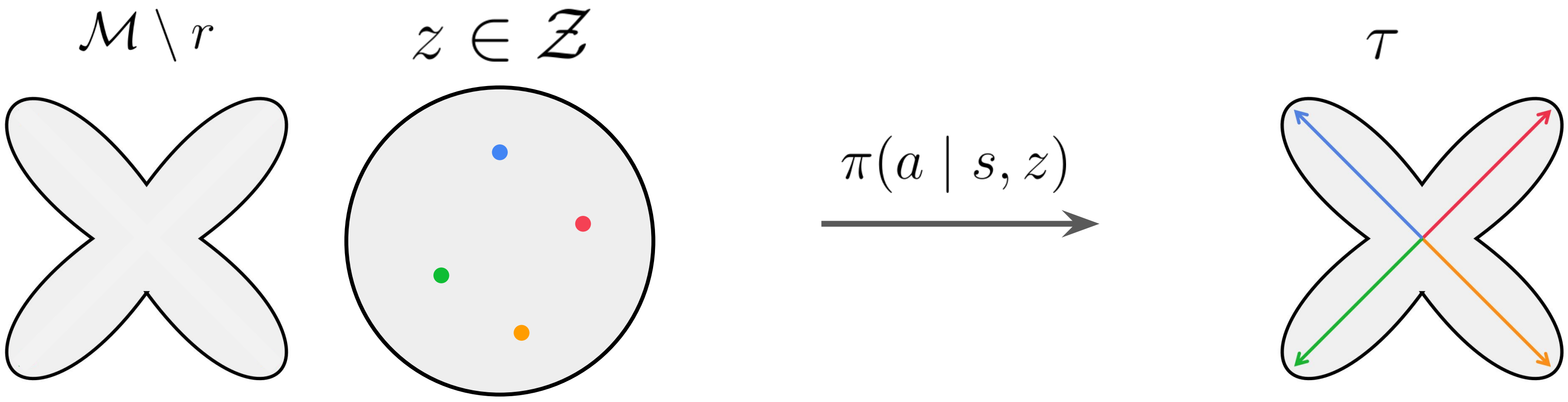
unsupervised skill learning

[1] Chuang et al. Debiased contrastive learning. 2020.
[2] He et al. Masked Autoencoders Are Scalable Vision Learners. 2021.
[3] Radford et al. Improving language understanding by generative pre-training. 2018.
[4] Sharma et al. Dynamics-Aware Unsupervised Discovery of Skills. 2020.

Mutual Information Skill Learning (MISL)

diverse *distinguishable*

$$I^\pi(S, S'; Z) = H^\pi(S, S') - H^\pi(S, S' \mid Z)$$



[1] Park et al. *METRA: Scalable Unsupervised RL with Metric-Aware Abstraction*. 2024.

[2] Eysenbach et al. *Diversity is all you need: Learning skills without a reward function*. 2018.

[3] Sharma et al. *Dynamics-Aware Unsupervised Discovery of Skills*. 2020.

[4] Laskin et al. *CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery*. 2022.

Alternative ideas to unsupervised skill learning

Can we build effective skill learning algorithms within the

***MISL** framework?*

MISL

DIYAN [2]

VISR [1]

CIC [3]

VIC [4]

DADS [5]

EDL [6]

METRA [7] $\max_{\pi} I^{\pi}_{\mathcal{W}}(S;Z)$

Wasserstein dependency measure

Learning state representations $\phi(s)$ to approximate the WDM

Updating the policy $\pi(a \mid s, z)$ using the successor features

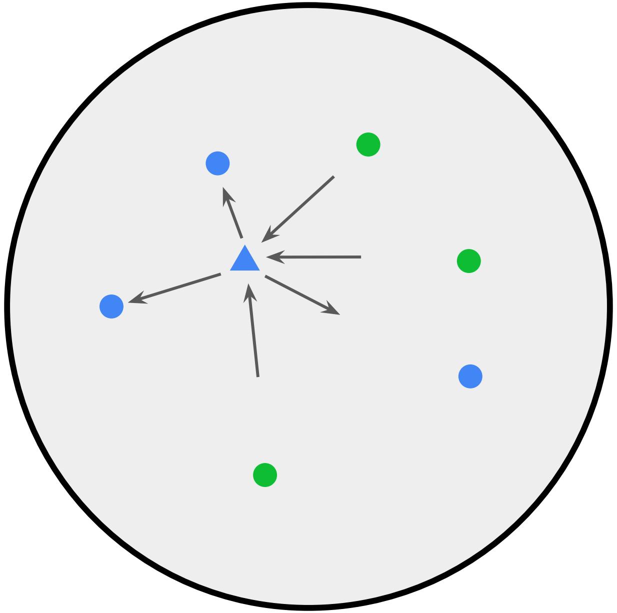
[1] Hansen et al. *Fast task inference with variational intrinsic successor features.* 2020.
[2] Eysenbach et al. *Diversity is all you need: Learning skills without a reward function.* 2018.
[3] Laskin et al. *CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery.* 2022.
[4] Gregor et al. *Variational intrinsic control.* 2016.
[5] Sharma et al. *Dynamics-Aware Unsupervised Discovery of Skills.* 2020.
[6] Campos et al. *Explore, discover and learn: Unsupervised discovery of state-covering skills.* 2020.
[7] Park et al. *METRA: Scalable Unsupervised RL with Metric-Aware Abstraction.* 2024.

Intuitions of METRA's representation objective

$$\min_{\lambda\geq 0}\max_{\phi} \underline{\mathbb{E}_{p^\beta(s,s',z)}\left[(\phi(s')-\phi(s))^\top z\right]} - \lambda \left(\underline{\mathbb{E}_{p^\beta(s,s')}\left[\|\phi(s')-\phi(s)\|_2^2\right] - 1}\right)$$

$\phi(s') - \phi(s)$: ▲

z : ● ●



Contrastive learning? ✓

[1] Park et al. *METRA: Scalable Unsupervised RL with Metric-Aware Abstraction.* 2024.

Relating METRA's representation objective to contrastive learning

Ours

InfoNCE loss

$$I^\beta(S, S'; Z) \geq \underline{\mathbb{E}_{p^\beta(s,s',z)}\left[(\phi(s') - \phi(s))^\top z\right]} - \underline{\mathbb{E}_{p^\beta(s,s')}\left[\log \mathbb{E}_{p(z')}\left[e^{(\phi(s')-\phi(s))^\top z}\right]\right]}$$

+: push together representations and skills from the same trajectory.

Second-order Taylor approximation

- : push away representations from other trajectories.

METRA

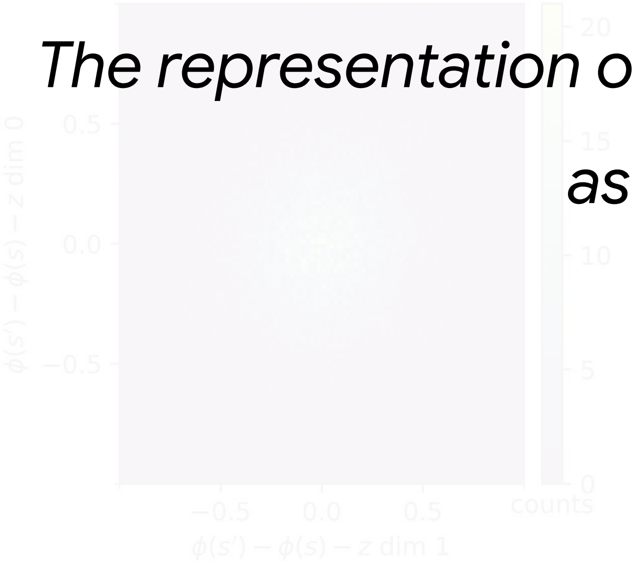
$$\underline{\mathbb{E}_{p^\beta(s,s',z)}\left[(\phi(s') - \phi(s))^\top z\right]} - \lambda\left(\underline{\mathbb{E}_{p^\beta(s,s')}\left[\|\phi(s') - \phi(s)\|_2^2\right] - 1}\right)$$

[1] Poole et al. *On variational bounds of mutual information*. 2019.

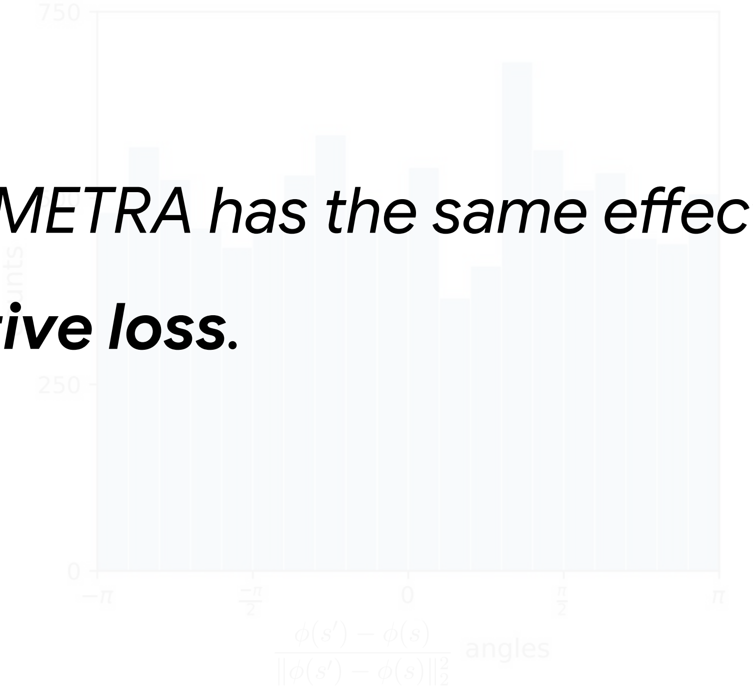
[2] Park et al. *METRA: Scalable Unsupervised RL with Metric-Aware Abstraction*. 2024.

METRA learns contrastive representations

Gaussianity ⇐ *contrastive representations [1]* ⇒ *Uniformity*



The representation objective in METRA has the same effect as a ***contrastive loss****.*



[1] Wang & Isola. *Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere.* 2020.

The intrinsic reward of METRA

✓

positive pairs have similar representations

$$r(s, s', z) = (\phi(s') - \phi(s))^\top z \qquad \nleq I^\pi(S, S'; Z)$$

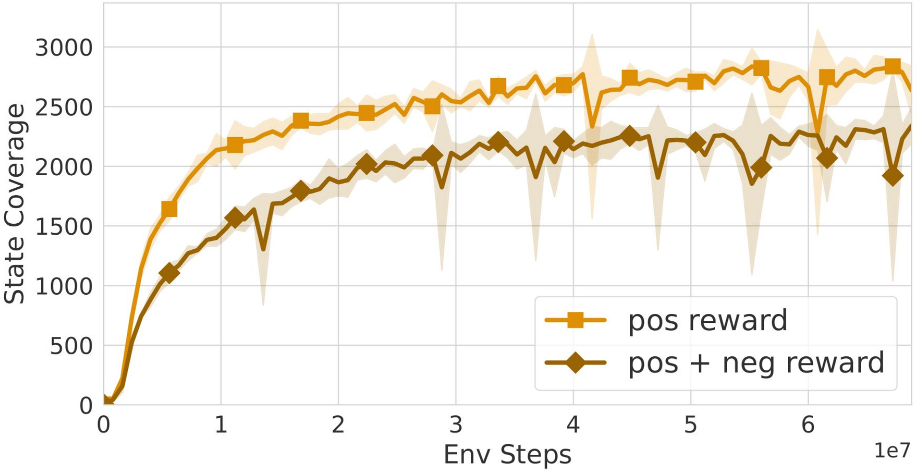
✗

negative pairs (ignored)

[1] Poole et al. *On variational bounds of mutual information*. 2019.

Relating METRA's policy objective to an information bottleneck

Maximizing MI is not enough

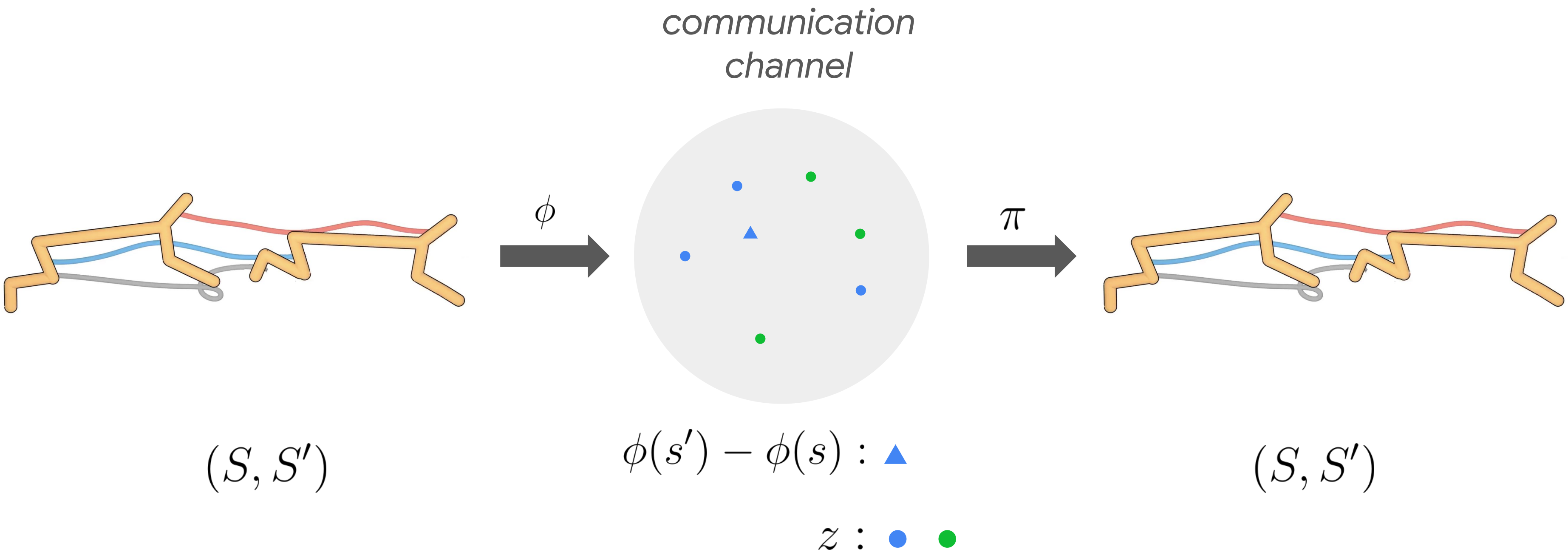


$$r(s, s', z) = (\phi(s') - \phi(s))^\top z \quad ?$$

Lower bound on a Information bottleneck

[1] Tishby et al. *The information bottleneck method*. 2000.
[2] Alemi et al. *Deep variational information bottleneck*. 2016.

Intuition for the information bottleneck



communication channel

ϕ

π

(S, S')

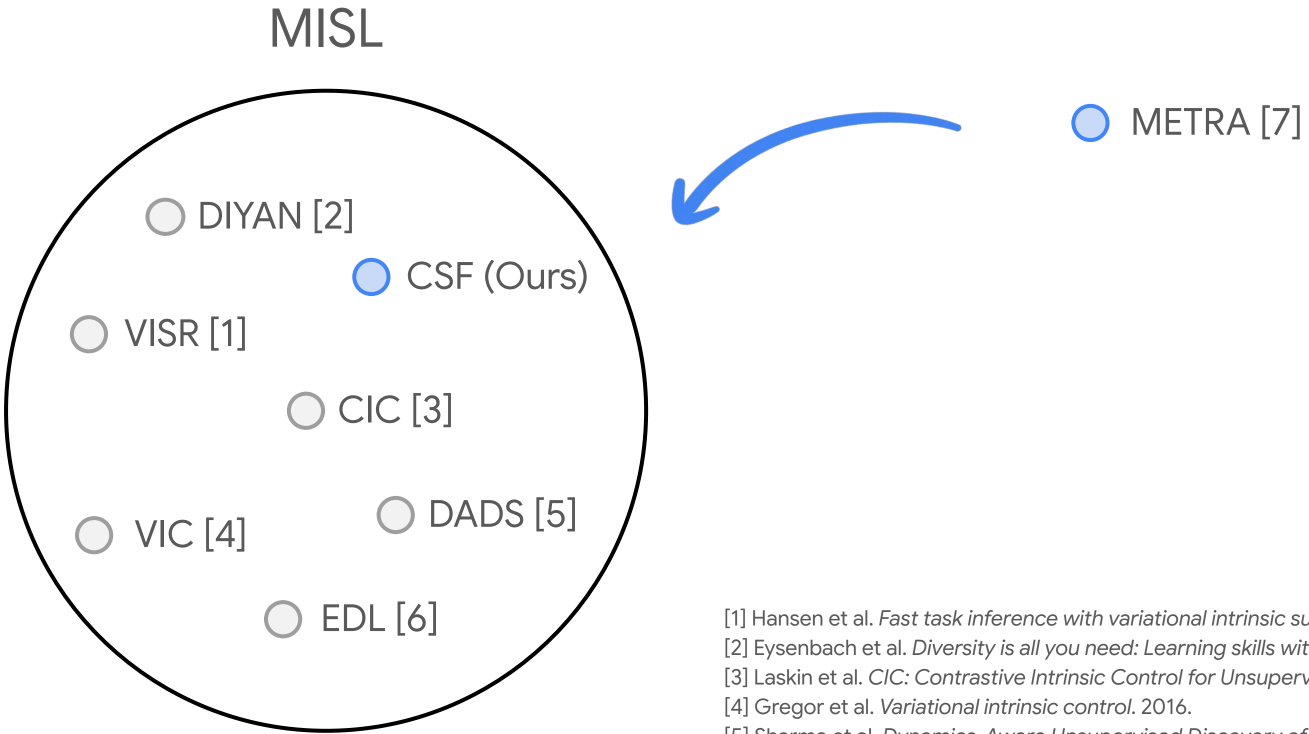
$\phi(s') - \phi(s)$: ▲

z : ● ●

(S, S')

[1] Janner & Du et al. *Planning with Diffusion for Flexible Behavior Synthesis*. 2022.
[2] Tishby et al. *The information bottleneck method*. 2000.
[3] Alemi et al. *Deep variational information bottleneck*. 2016.
[4] Kingma. *Auto-Encoding Variational Bayes*. 2016.

Why do we need new interpretations?



MISL

DIYAN [2]

CSF (Ours)

VISR [1]

CIC [3]

DADS [5]

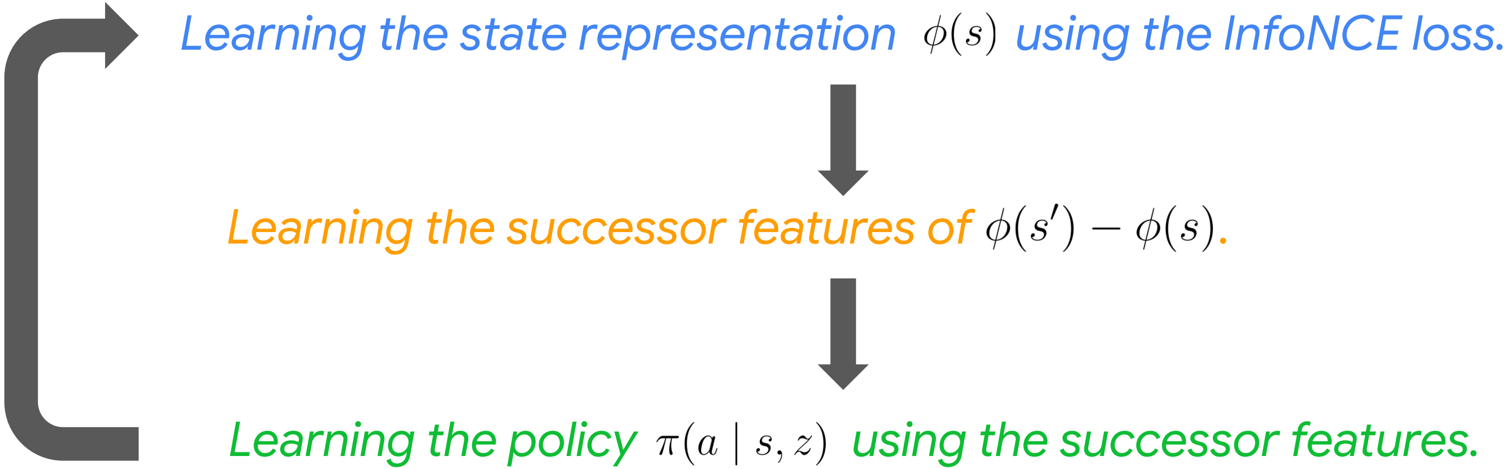
VIC [4]

EDL [6]

METRA [7]

[1] Hansen et al. *Fast task inference with variational intrinsic successor features*. 2020.
[2] Eysenbach et al. *Diversity is all you need: Learning skills without a reward function*. 2018.
[3] Laskin et al. *CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery*. 2022.
[4] Gregor et al. *Variational intrinsic control*. 2016.
[5] Sharma et al. *Dynamics-Aware Unsupervised Discovery of Skills*. 2020.
[6] Campos et al. *Explore, discover and learn: Unsupervised discovery of state-covering skills*. 2020.
[7] Park et al. *METRA: Scalable Unsupervised RL with Metric-Aware Abstraction*. 2024.

Ideas of contrastive successor features (CSF)



Learning the state representation $\phi(s)$ *using the InfoNCE loss.*

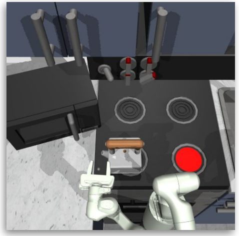
Learning the successor features of $\phi(s') - \phi(s)$.

Learning the policy $\pi(a \mid s, z)$ *using the successor features.*

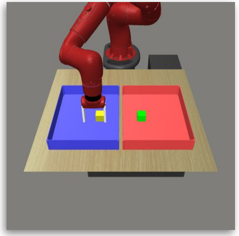
Learning skills to explore the state space from pixels

Humanoid

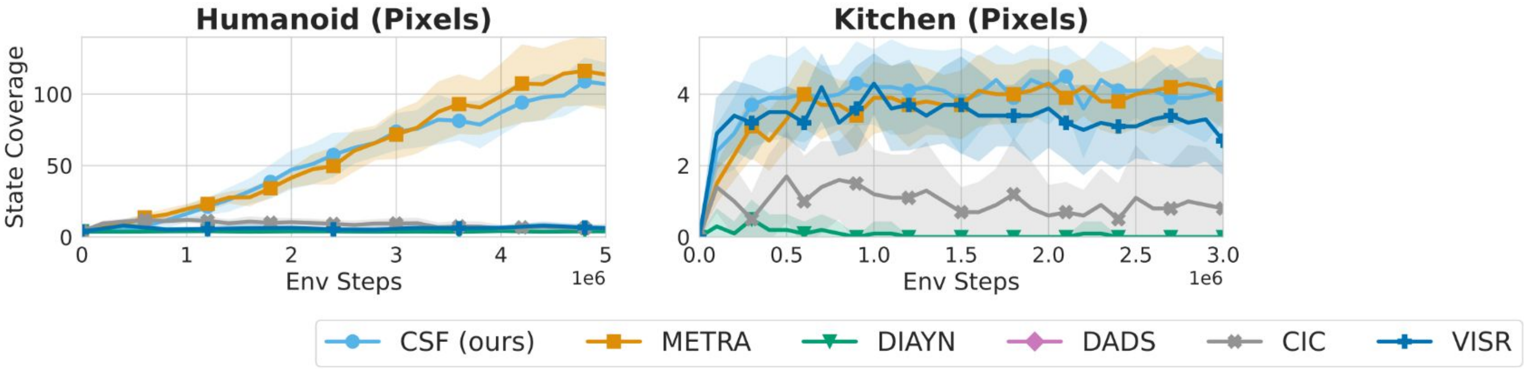
Kitchen



Robobin

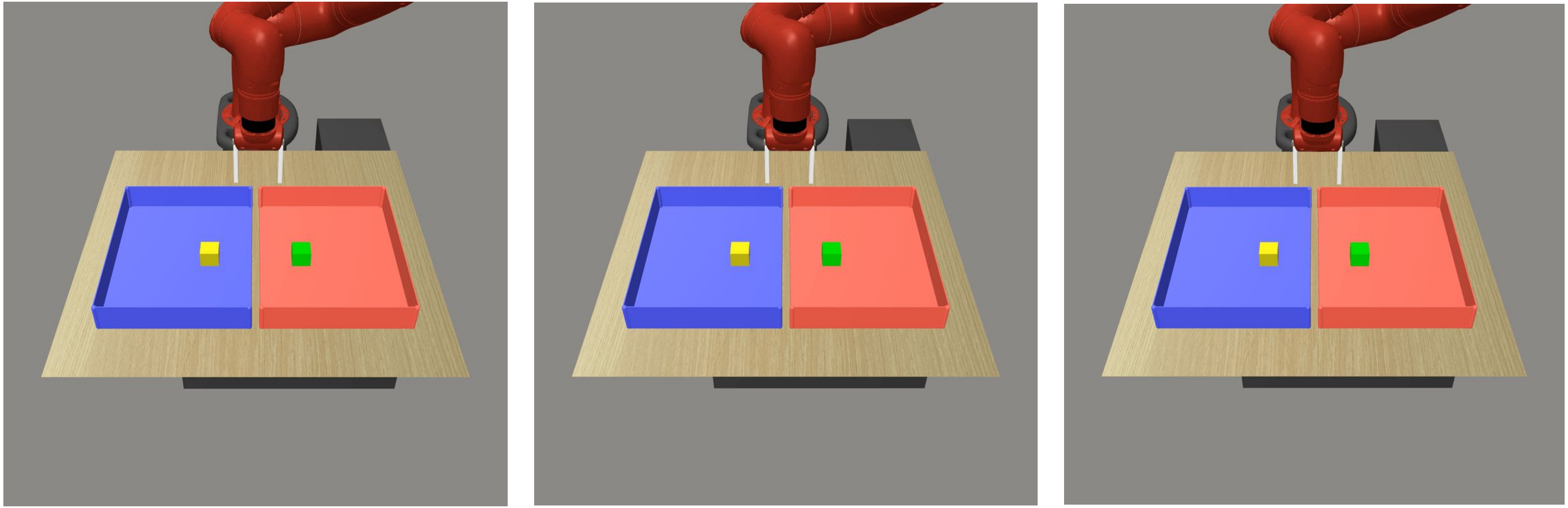


1.5x higher coverage!



[1] Park et al. METRA: Scalable *Unsupervised RL with Metric-Aware Abstraction*. 2024.
[2] Mendonca & Rybkin. *Discovering and Achieving Goals via World Model*. 2021.

CSF can learn manipulation skills without a reward function.



Summary and connections

tldr: explain and simplify METRA within MISL

- Representation learning - contrastive learning
- Policy learning - information bottleneck
- Simplified version: contrastive successor features
- Connections with many other areas:
  - Forward backward representations
  - Goal-conditioned RL
  - Zero-shot adaption

Video, code, and paper!

https://princeton-rl.github.io/contrastive-successor-features/