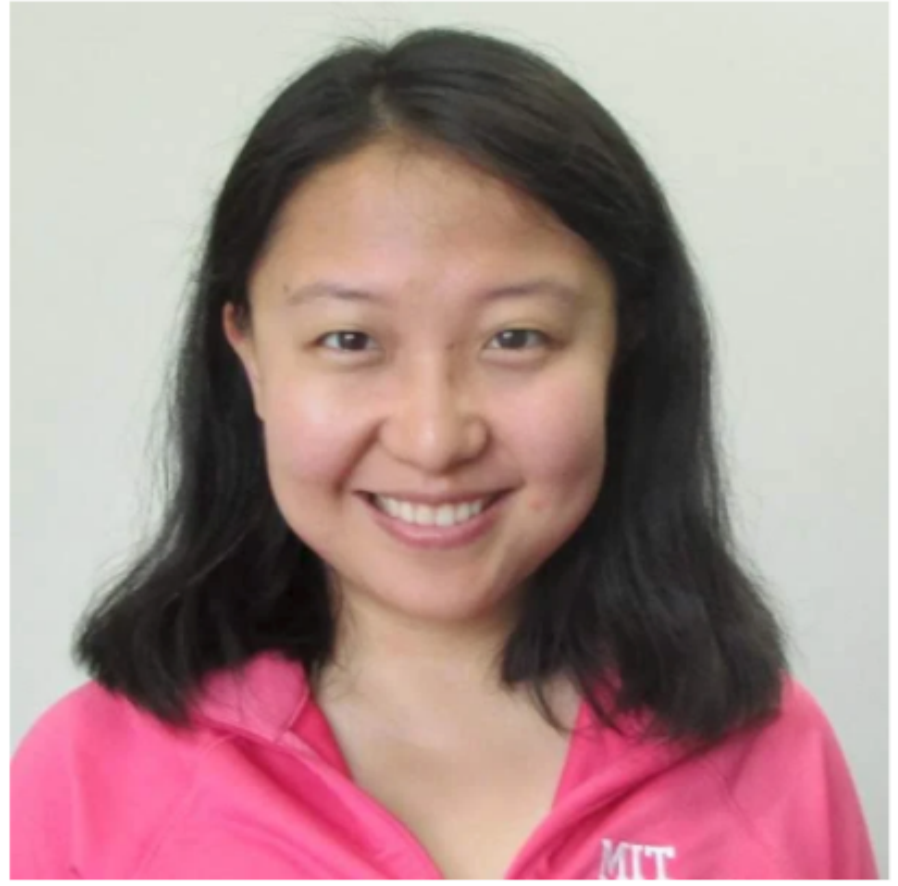
TRAIL: Near-Optimal Imitation Learning with Suboptimal Data

Sherry Yang Sergey Levine Ofir Nachum

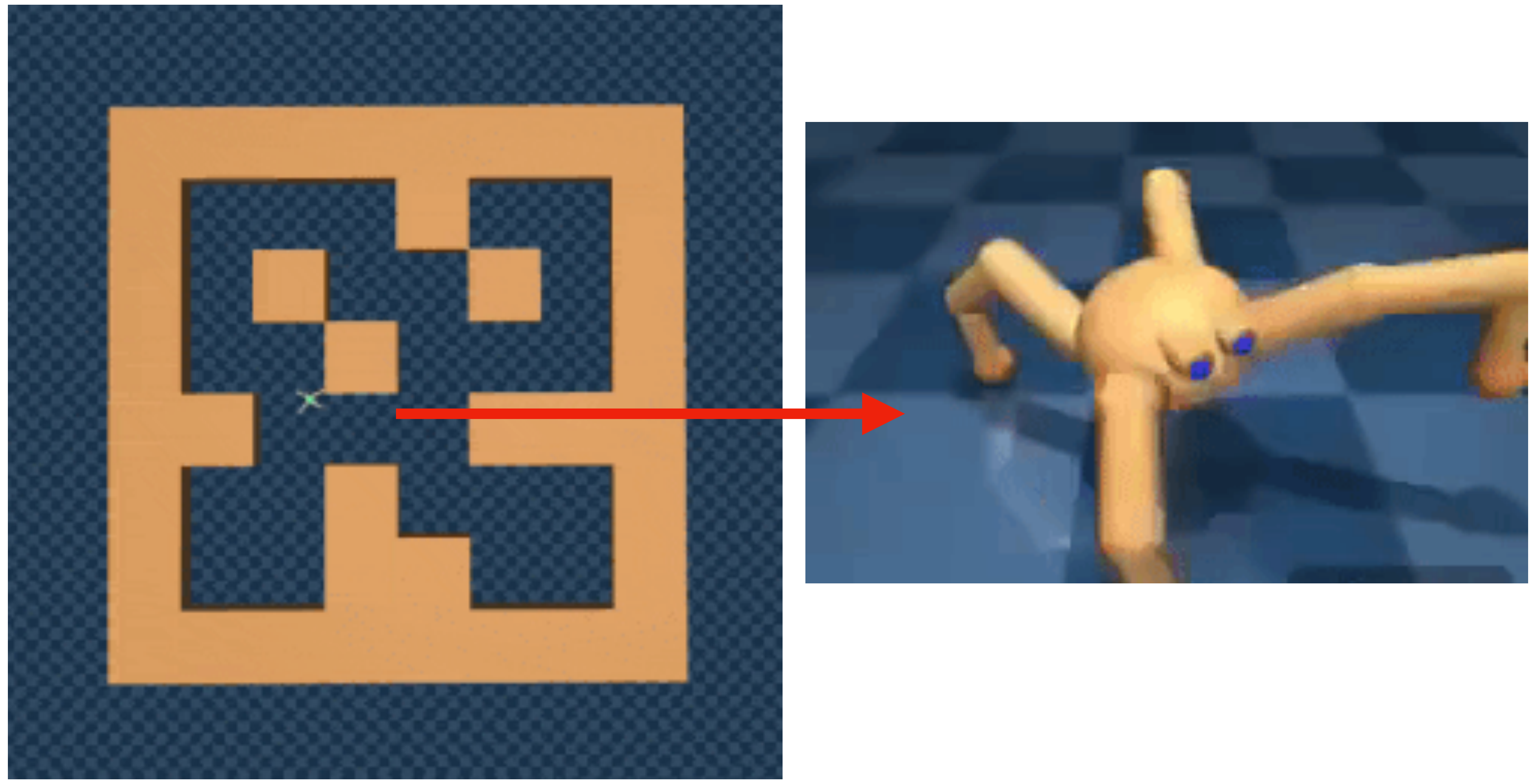


Paper: https://openreview.net/pdf?id=6q_2b6u0BnJ
Code: https://github.com/google-research/google-research/tree/master/rl_repr

Imitation Learning

Given expert demonstrations $\mathscr{D}^{\pi^*}$

Learn π that recovers π^*: $\mathrm{Diff}(\pi, \pi_*) = D_{\mathrm{TV}}(d^{\pi} \| d^{\pi}_*)$



Behavioral cloning:

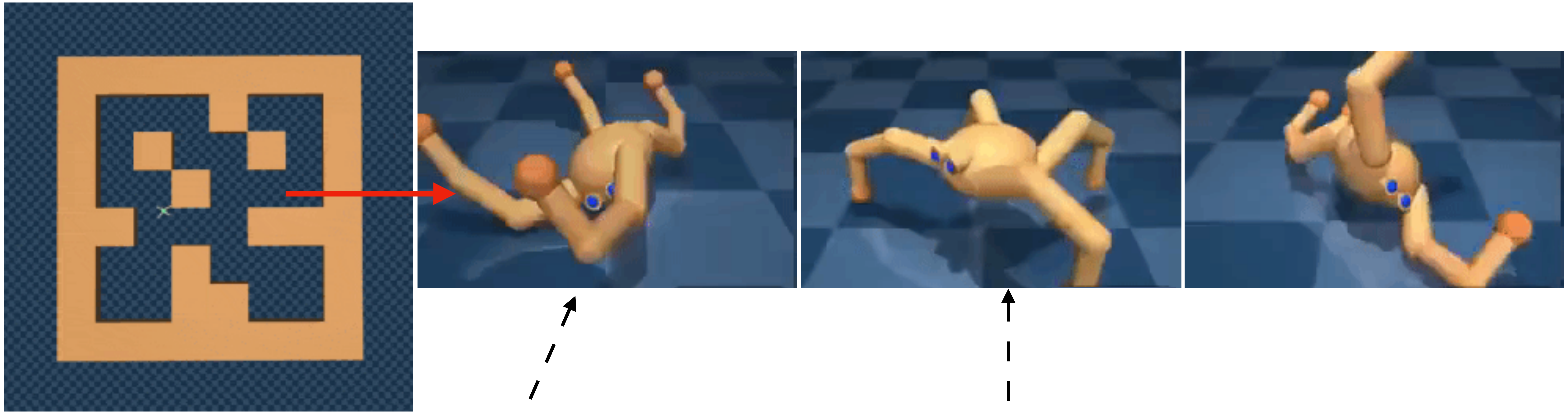
$$J_{\mathrm{BC}}(\pi) := \mathbb{E}_{(s,a)\sim(d^{\pi_*}, \pi_*)}[-\log \pi(a|s)]$$

Limited & Hard to obtain
(e.g., involves human expert)

Suboptimal Offline Data

Large amounts of suboptimal offline data $\mathscr{D}^{\mathit{off}}$

How can $\mathscr{D}^{\mathit{off}}$ <u>provably</u> facilitate imitation learning?

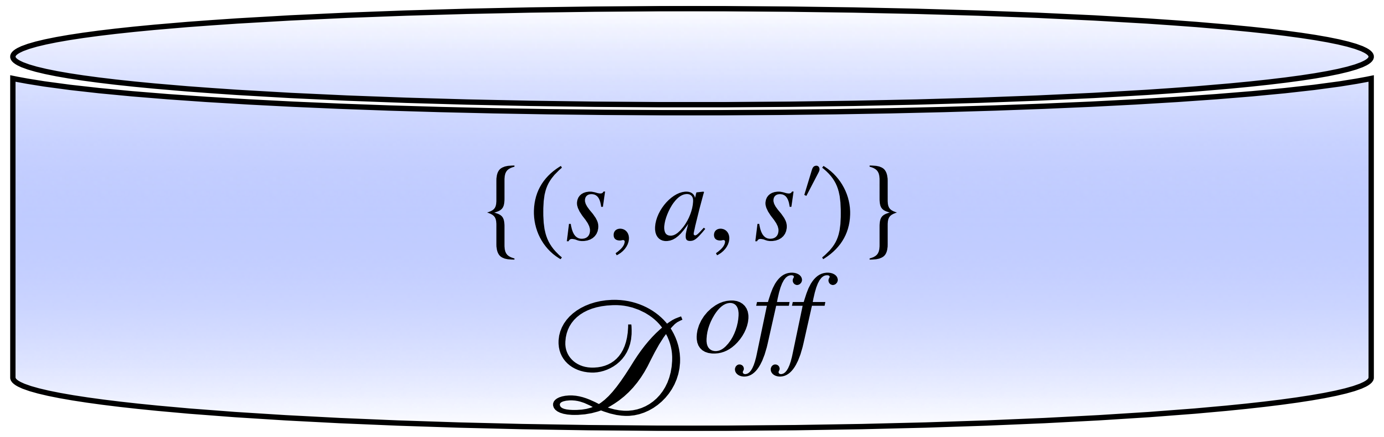


- Highly suboptimal (e.g., random policy)
- Single modal (e.g., collected by <u>one</u> stationary policy)

TRAIL: Transition Reparametrized Actions

Factored transition model

(1) $T_z \circ \phi(s, a)$

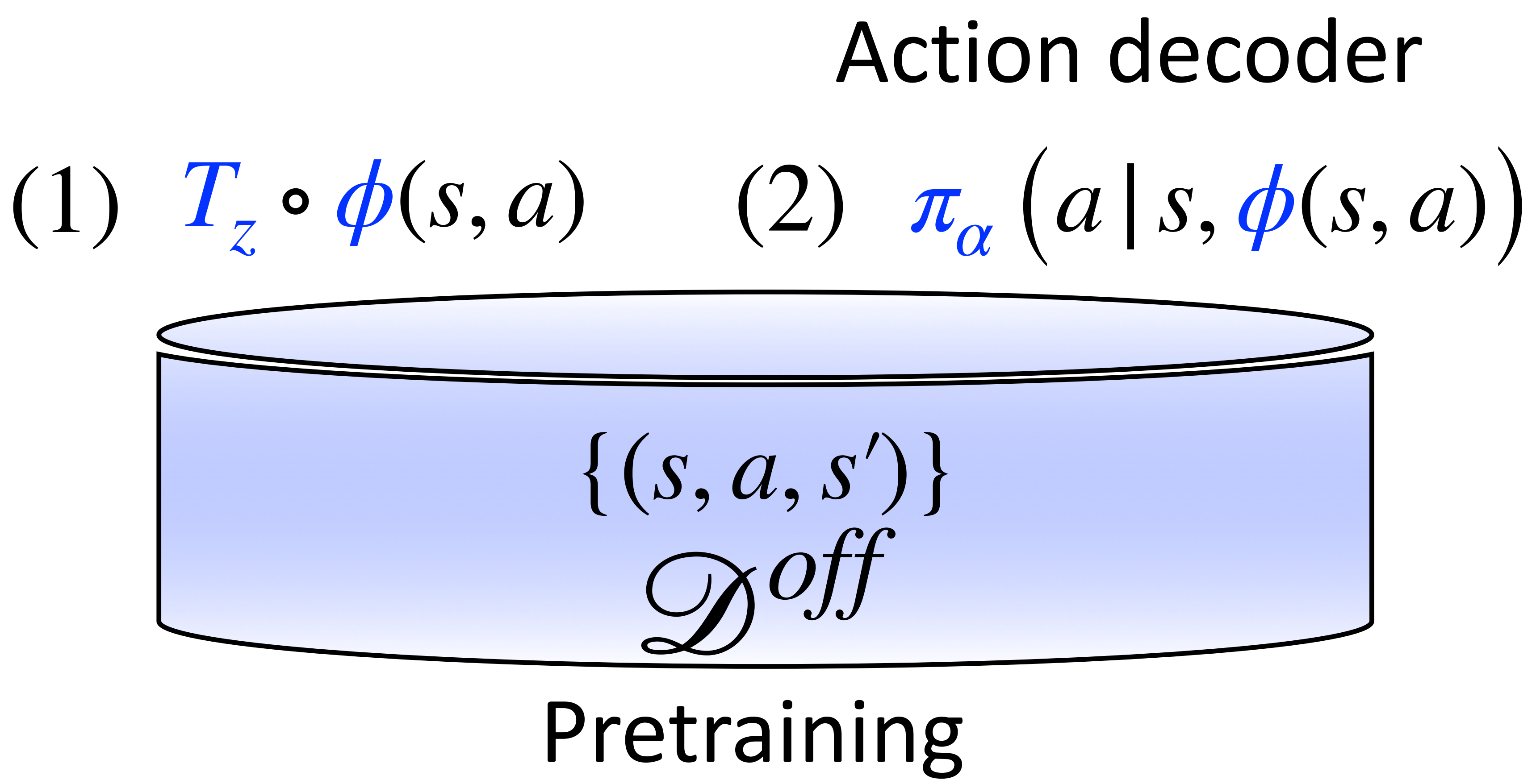


Pretraining

Pretraining

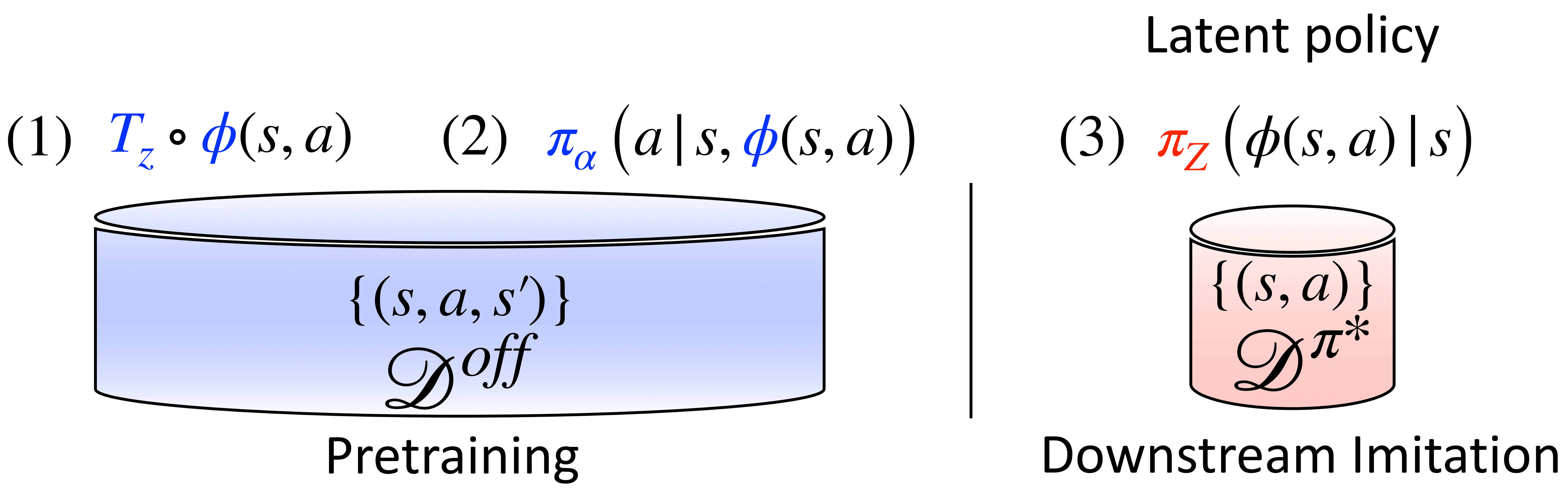
$$\underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\text{KL}}(\mathcal{T}(s,a)\|\mathcal{T}_Z(s,\phi(s,a)))\right]}_{=J_{\text{T}}(\mathcal{T}_Z,\phi)} \quad (1)$$

TRAIL: Transition Reparametrized Actions



$$\textit{Pretraining} \begin{cases} \underbrace{\mathbb{E}_{(s,a)\sim d^{\mathrm{off}}}\left[D_{\mathrm{KL}}(\mathcal{T}(s,a) \| \mathcal{T}_Z(s,\phi(s,a)))\right]}_{= J_{\mathrm{T}}(\mathcal{T}_Z,\phi)} & (1) \\ \underbrace{\mathbb{E}_{s\sim d^{\mathrm{off}}}[\max_{z\in Z} D_{\mathrm{KL}}(\pi_{\alpha^*}(s,z)\|\pi_\alpha(s,z))]}_{\approx \ \mathrm{const}(d^{\mathrm{off}},\phi) + J_{\mathrm{DE}}(\pi_\alpha,\phi)} & (2) \end{cases}$$

TRAIL: Transition Reparametrized Actions



Pretraining

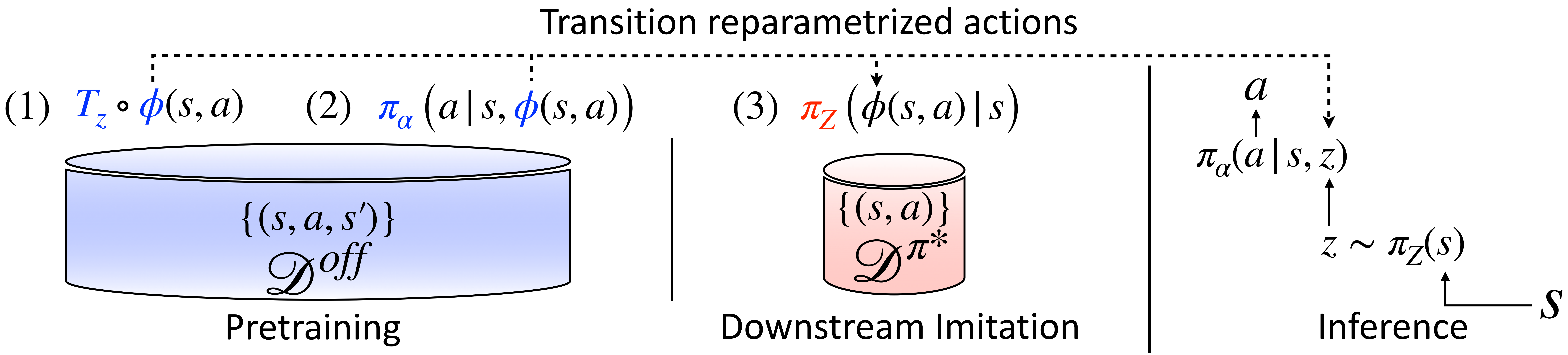
$$\underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\mathrm{KL}}(\mathcal{T}(s,a)\|\mathcal{T}_Z(s,\phi(s,a)))\right]}_{=\,J_{\mathrm{T}}(\mathcal{T}_Z,\phi)} \tag{1}$$

$$\underbrace{\mathbb{E}_{s\sim d^{\text{off}}}\left[\max_{z\in Z} D_{\mathrm{KL}}(\pi_{\alpha^*}(s,z)\|\pi_\alpha(s,z))\right]}_{\approx\ \mathrm{const}(d^{\text{off}},\phi)\,+\,J_{\mathrm{DE}}(\pi_\alpha,\phi)} \tag{2}$$

Downstream Imitation

$$\underbrace{\mathbb{E}_{s\sim d^{\pi_*}}\left[D_{\mathrm{KL}}(\pi_{*,Z}(s)\|\pi_Z(s))\right]}_{=\ \mathrm{const}(\pi_*,\phi)\,+\,J_{\mathrm{BC},\phi}(\pi_Z)}, \tag{3}$$

TRAIL: Transition Reparametrized Actions



Pretraining

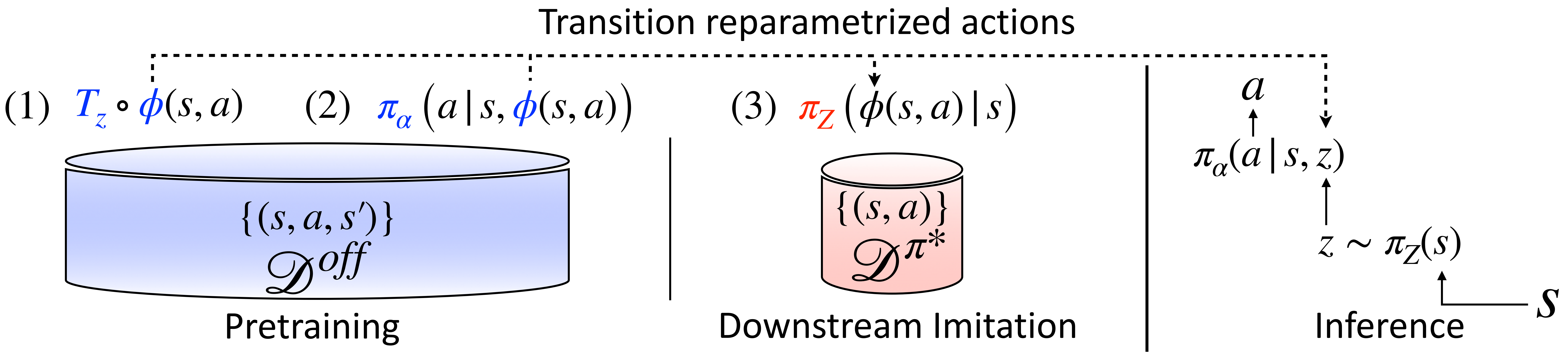
$$\underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\text{KL}}(\mathcal{T}(s,a) \| \mathcal{T}_Z(s, \phi(s,a)))\right]}_{= J_{\text{T}}(\mathcal{T}_Z, \phi)} \tag{1}$$

$$\underbrace{\mathbb{E}_{s\sim d^{\text{off}}}\left[\max_{z\in Z} D_{\text{KL}}(\pi_{\alpha^*}(s,z) \| \pi_\alpha(s,z))\right]}_{\approx\ \text{const}(d^{\text{off}}, \phi) + J_{\text{DE}}(\pi_\alpha, \phi)} \tag{2}$$

Downstream Imitation

$$\underbrace{\mathbb{E}_{s\sim d^{\pi_*}}\left[D_{\text{KL}}(\pi_{*,Z}(s) \| \pi_Z(s))\right]}_{=\ \text{const}(\pi_*, \phi) + J_{\text{BC},\phi}(\pi_Z)}, \tag{3}$$

TRAIL: Transition Reparametrized Actions



$$\mathrm{Diff}(\pi_\alpha \circ \pi_Z, \pi_*) \leq$$

$$\textit{Pretraining} \begin{cases} C_1 \cdot \sqrt{\dfrac{1}{2} \underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\mathrm{KL}}(\mathcal{T}(s,a) \| \mathcal{T}_Z(s, \phi(s,a)))\right]}_{= J_{\mathrm{T}}(\mathcal{T}_Z, \phi)}} & (1) \\ +C_2 \cdot \sqrt{\dfrac{1}{2} \underbrace{\mathbb{E}_{s\sim d^{\text{off}}}\left[\max_{z \in Z} D_{\mathrm{KL}}(\pi_{\alpha^*}(s,z) \| \pi_\alpha(s,z))\right]}_{\approx\ \mathrm{const}(d^{\text{off}}, \phi) + J_{\mathrm{DE}}(\pi_\alpha, \phi)}} & (2) \end{cases}$$

$$\begin{matrix}\textit{Downstream} \\ \textit{Imitation}\end{matrix} \begin{cases} +C_3 \cdot \sqrt{\dfrac{1}{2} \underbrace{\mathbb{E}_{s\sim d^{\pi_*}}\left[D_{\mathrm{KL}}(\pi_{*,Z}(s) \| \pi_Z(s))\right]}_{=\ \mathrm{const}(\pi_*, \phi) + J_{\mathrm{BC},\phi}(\pi_Z)}}, \end{cases}$$

$$\begin{aligned} C_1 &= \gamma |A| (1-\gamma)^{-1} (1 + D_{\chi^2}(d^{\pi_*} \| d^{\text{off}})^{\frac{1}{2}}) \\ C_2 &= \gamma (1-\gamma)^{-1} (1 + D_{\chi^2}(d^{\pi_*} \| d^{\text{off}})^{\frac{1}{2}}) \\ C_3 &= \gamma (1-\gamma)^{-1} \end{aligned}$$

Sample Complexity of TRAIL

$$\mathbb{E}_{\mathcal{D}^{\pi_*}}[\mathrm{Diff}(\pi_{opt,Z}, \pi_*)] \leq (1)(\phi_{opt}) + (2)(\phi_{opt}) + C_3 \cdot \sqrt{\frac{|Z||S|}{n}}$$

So far, our analysis is based on tabular latent actions.

What about continuous latent actions and stochastic expert policy?

TRAIL with Linear Transition Dynamics

deterministic linear: $T_z = w(s')^\top \phi(s,a)$

$$\mathrm{Diff}(\pi_\alpha \circ \boxed{\pi_\theta}, \pi_*) \leq (1)(\boxed{\mathcal{T}_Z}, \phi) + (2)(\pi_\alpha, \phi)$$

Downstream Imitation
$$\left\{ + C_4 \cdot \boxed{\left\| \frac{\partial}{\partial \theta} \mathbb{E}_{s \sim d^{\pi_*}, a \sim \pi_*(s)} [(\theta_s - \phi(s,a))^2] \right\|_1} \right.$$

TRAIL with Linear Transition Dynamics

deterministic

linear: $T_z = w(s')^\top \phi(s, a)$

$$\mathrm{Diff}(\pi_\alpha \circ \boxed{\pi_\theta}, \pi_*) \leq (1)(\boxed{\mathcal{T}_Z}, \phi) + (2)(\pi_\alpha, \phi)$$

$$\textit{Downstream Imitation} \left\{ + C_4 \cdot \boxed{\left\| \frac{\partial}{\partial \theta} \mathbb{E}_{s \sim d^{\pi_*}, a \sim \pi_*(s)} [(\theta_s - \phi(s,a))^2] \right\|_1} \right.$$

easier to optimize compared to:

$$\cdot C_3 \cdot \sqrt{\underbrace{\frac{1}{2} \mathbb{E}_{s \sim d^{\pi_*}} [D_{\mathrm{KL}}(\pi_{*,Z}(s) \| \pi_Z(s))]}_{= \ \mathrm{const}(\pi_*, \phi) + J_{\mathrm{BC},\phi}(\pi_Z)}}$$

Learning TRAIL in Practice

$$(1)\quad T_z \circ \phi(s, a)$$

TRAIL EBM: $\mathcal{T}_Z(s'|s, \phi(s,a)) \propto \rho(s')\exp(-\|\phi(s,a) - \psi(s')\|^2)$.

$$\mathbb{E}_{d^{\text{off}}}[-\log \mathcal{T}_Z(s'|s, \phi(s,a)))] = \text{const}(d^{\text{off}}) + \frac{1}{2}\mathbb{E}_{d^{\text{off}}}[||\phi(s,a) - \psi(s')||^2] + \log \mathbb{E}_{\tilde{s}' \sim \rho}[\exp\{-\frac{1}{2}||\phi(s,a) - \psi(\tilde{s}')||^2\}]$$

contrastive learning

Learning TRAIL in Practice

(1) $T_z \circ \phi(s,a)$

TRAIL EBM: $\mathcal{T}_Z(s'|s,\phi(s,a)) \propto \rho(s')\exp(-\|\phi(s,a)-\psi(s')\|^2)$

$$\mathbb{E}_{d^{\text{off}}}[-\log \mathcal{T}_Z(s'|s,\phi(s,a)))] = \text{const}(d^{\text{off}}) + \frac{1}{2}\mathbb{E}_{d^{\text{off}}}[||\phi(s,a)-\psi(s')||^2] \\ + \log \mathbb{E}_{\tilde{s}'\sim\rho}[\exp\{-\frac{1}{2}||\phi(s,a)-\psi(\tilde{s}')||^2\}]$$

contrastive learning

TRAIL linear: $\overline{\mathcal{T}}(s'|s,a) \propto \rho(s')\exp\{-||f(s,a)-g(s')||^2/2\} \propto \bar{\psi}(s')^\top \bar{\phi}(s,a)$

recover $\bar{\phi}$ with random Fourier features: $\bar{\phi}(s,a) = \cos(Wf(s,a)+b)$

Learning TRAIL in Practice

(1) $T_z \circ \phi(s,a)$ (2) $\pi_\alpha\left(a \mid s, \phi(s,a)\right)$ (3) $\pi_Z\left(\phi(s,a) \mid s\right)$

TRAIL EBM: $\mathcal{T}_Z(s'|s,\phi(s,a)) \propto \rho(s')\exp(-\|\phi(s,a)-\psi(s')\|^2)$

$$\mathbb{E}_{d^{\text{off}}}[-\log \mathcal{T}_Z(s'|s,\phi(s,a)))] = \text{const}(d^{\text{off}}) + \frac{1}{2}\mathbb{E}_{d^{\text{off}}}[||\phi(s,a)-\psi(s')||^2]$$

contrastive learning

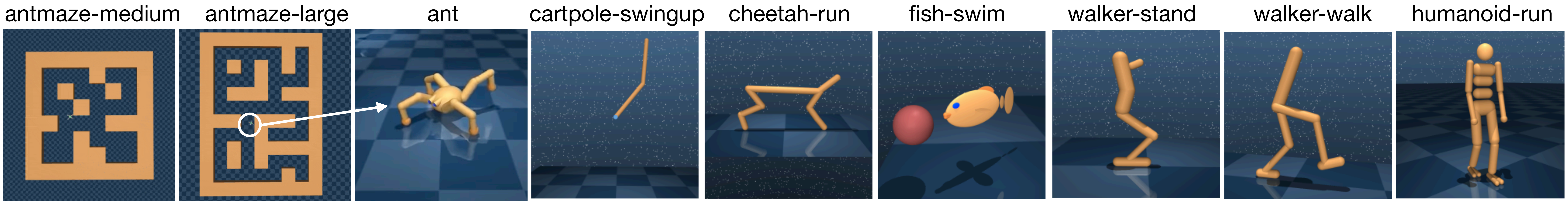
$$+\log \mathbb{E}_{\tilde{s}'\sim\rho}[\exp\{-\frac{1}{2}||\phi(s,a)-\psi(\tilde{s}')||^2\}]$$

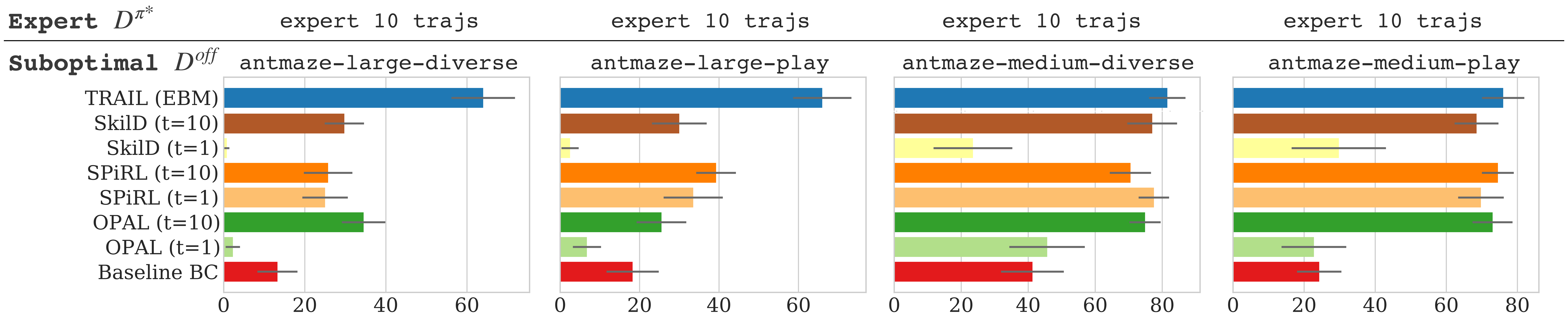
TRAIL linear: $\overline{\mathcal{T}}(s'|s,a) \propto \rho(s')\exp\{-||f(s,a)-g(s')||^2/2\} \propto \bar{\psi}(s')^\top\bar{\phi}(s,a)$

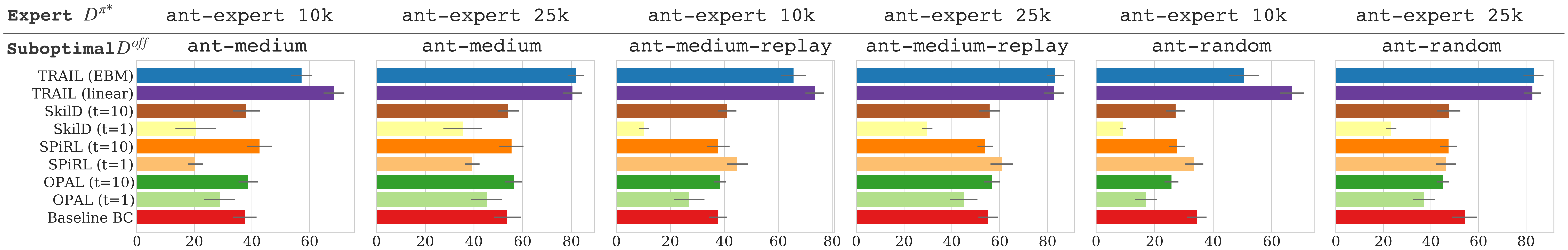
recover $\bar{\phi}$ with random Fourier features: $\bar{\phi}(s,a) = \cos(Wf(s,a)+b)$

π_α and π_Z are neural-network parametrized Guassian policies.

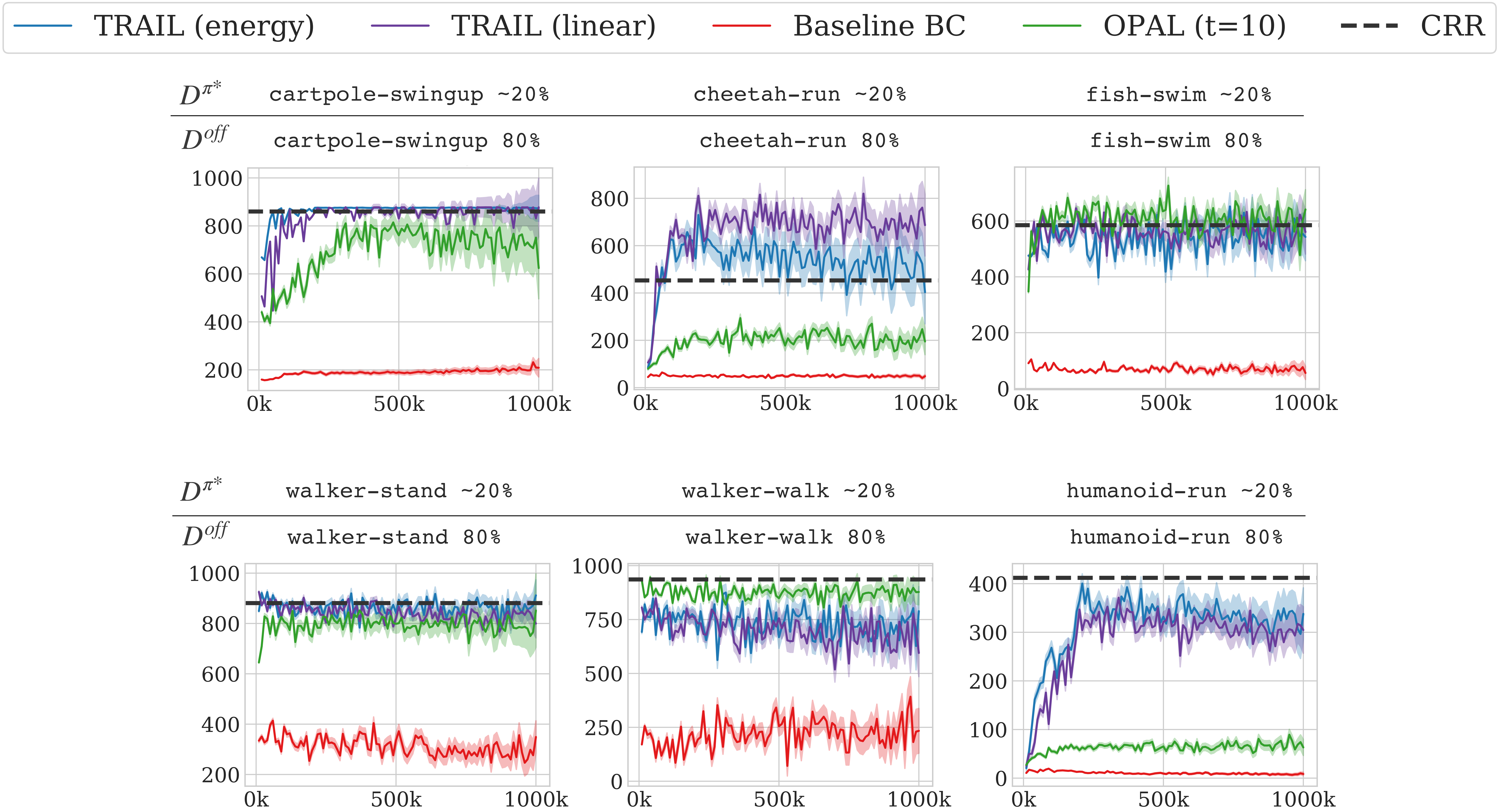
Experiments







Experiments - DM Control Suite



Recap & Conclusion

- How to utilize additional offline data for imitation learning?
  - Learn action representations.
- What if the offline data is highly suboptimal or unimodal?
  - Learn transition model as opposed to temporal skills.
- Representation learning + imitation learning as an alternative to offline RL?
  - Beneficial especially in the absence of reward labels.

Thank you. Checkout

Paper: https://openreview.net/pdf?id=6q_2b6u0BnJ
Code: https://github.com/google-research/google-research/tree/master/rl_repr