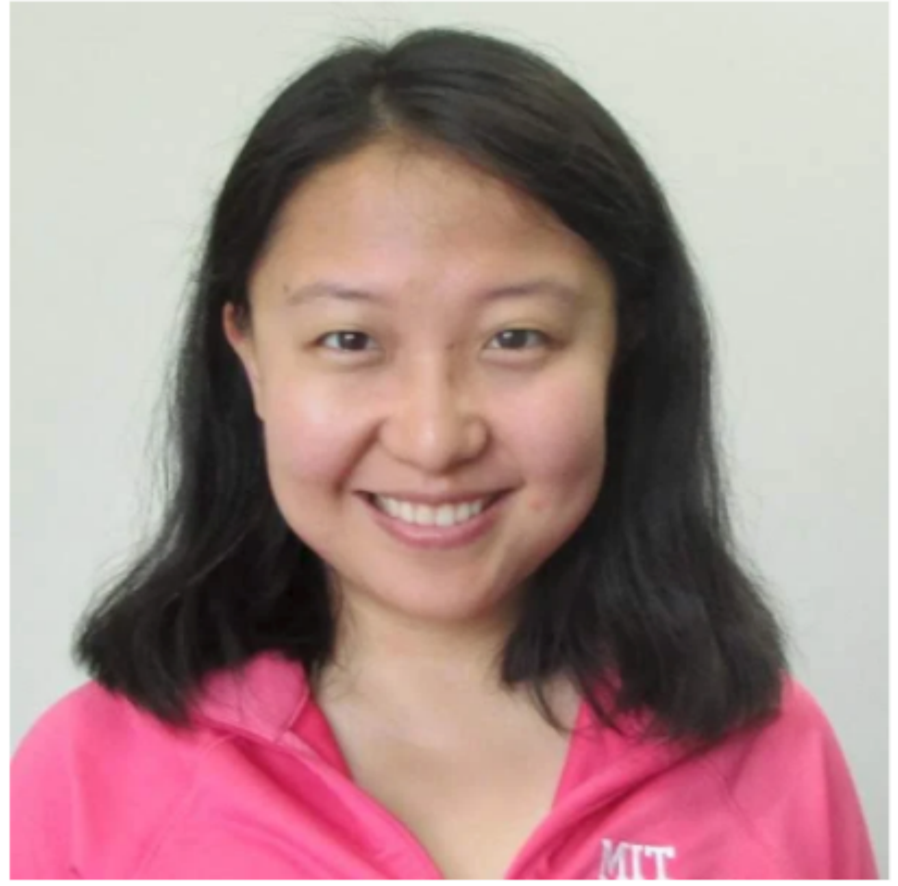
TRAIL: Near-Optimal Imitation Learning with Suboptimal Data

Sherry Yang
sherryy@

Sergey Levine
slevine@

Ofir Nachum
ofirnachum@

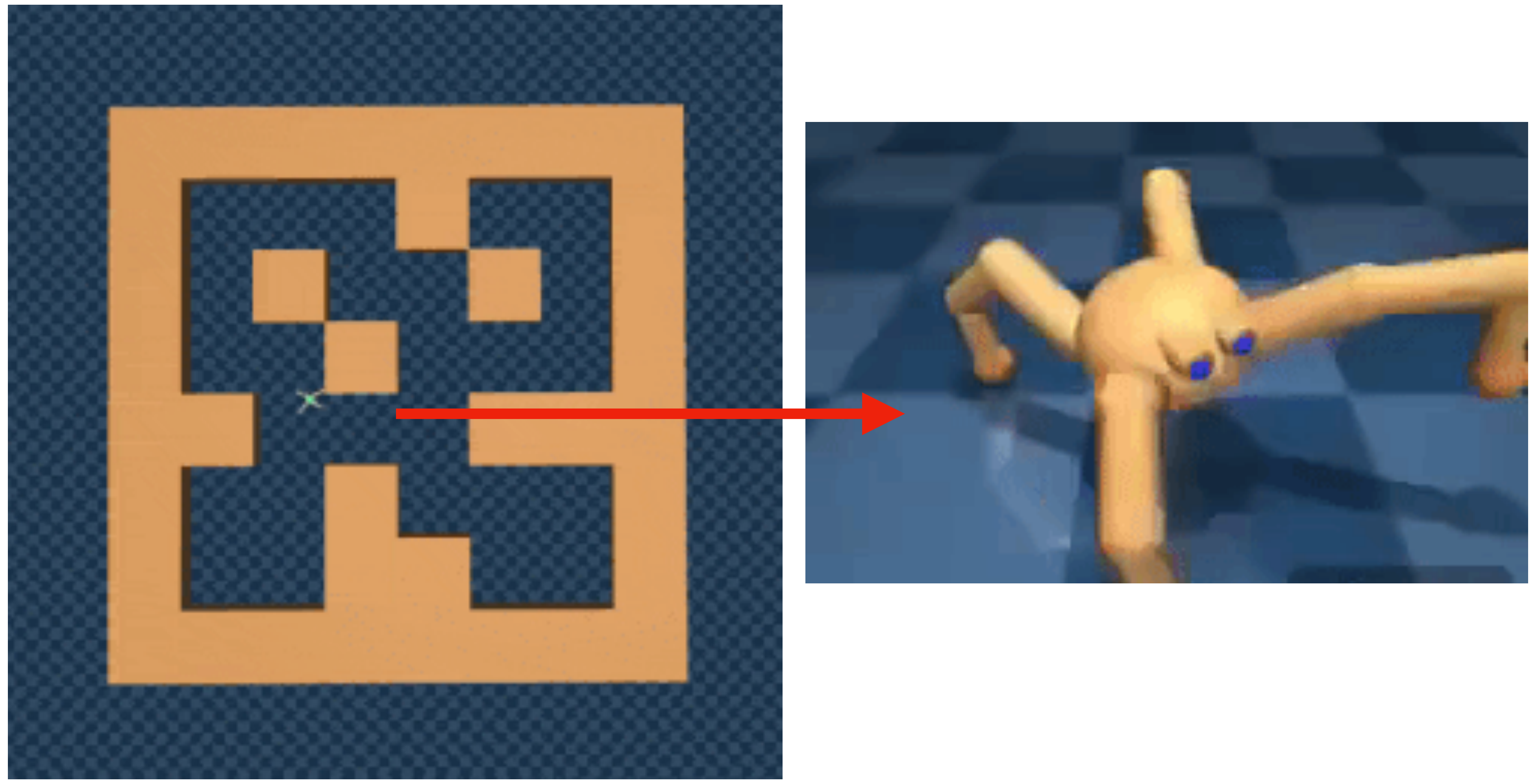


Paper: http://arxiv.org/abs/2110.14770
Code: https://github.com/google-research/google-research/tree/master/rl_repr

Imitation Learning

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

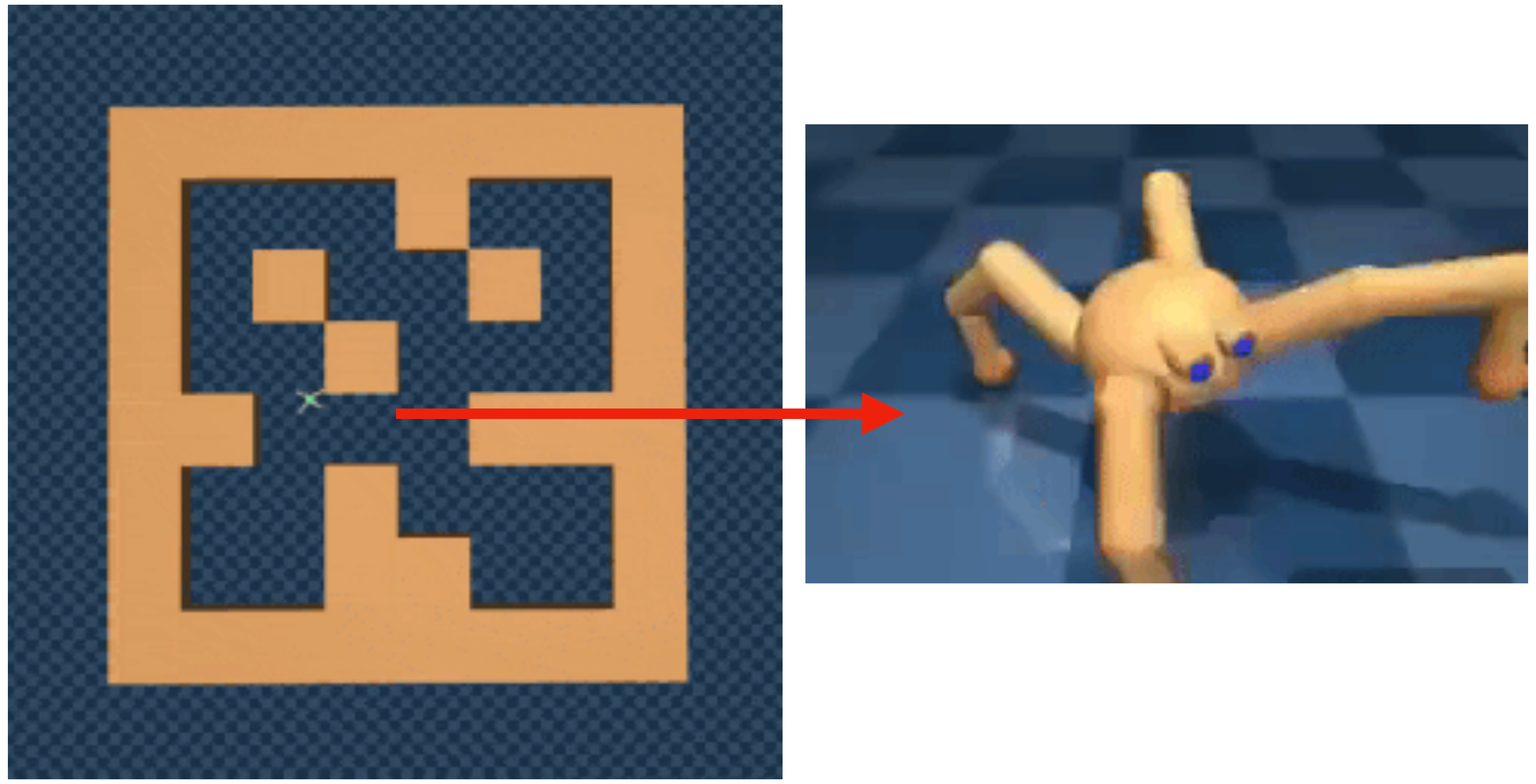
Learn π that recovers π^*: $\mathrm{Diff}(\pi, \pi_*) = D_{\mathrm{TV}}(d^{\pi} \| d^{\pi_*})$



Imitation Learning

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

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



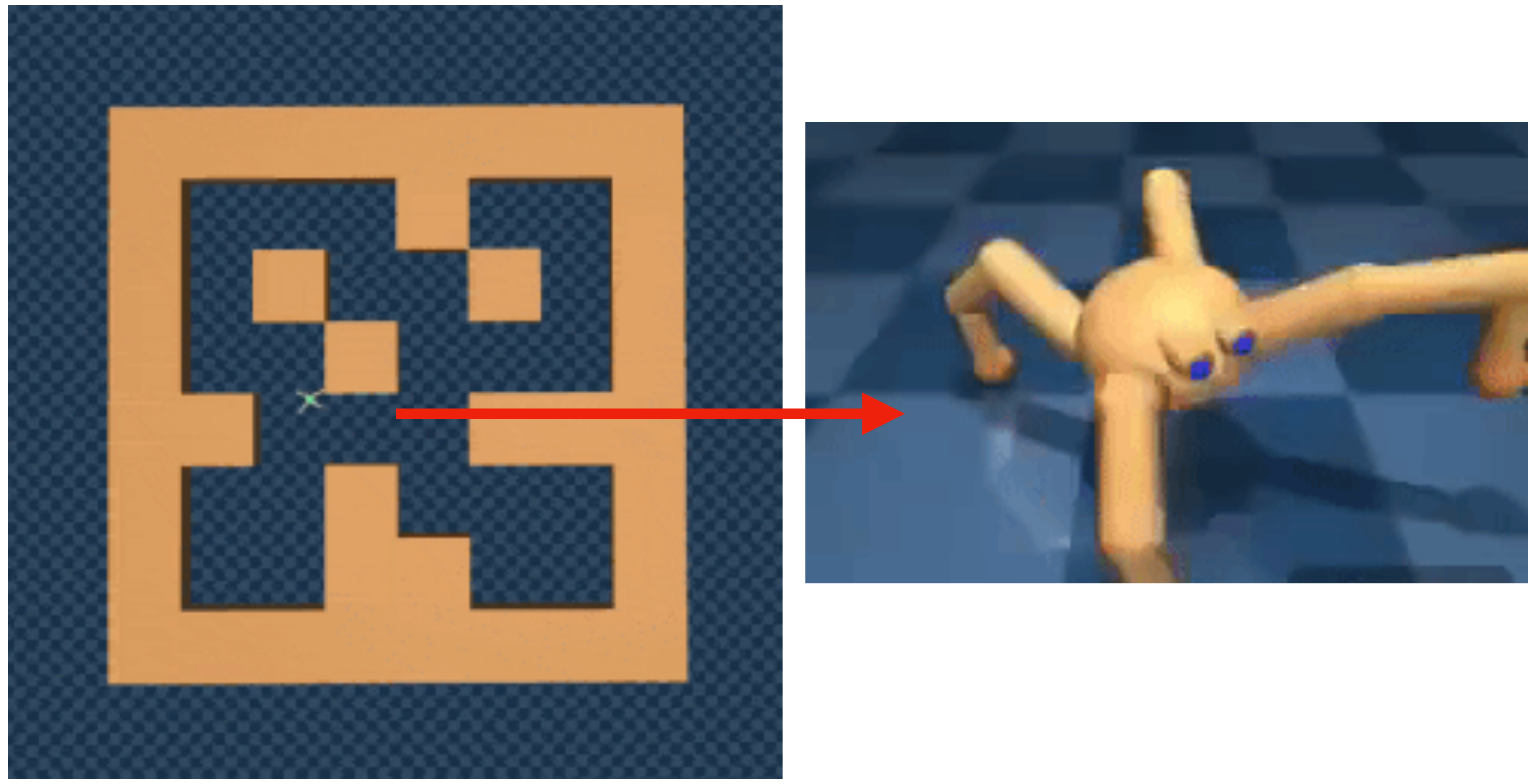
Behavioral cloning:

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

Imitation Learning

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

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



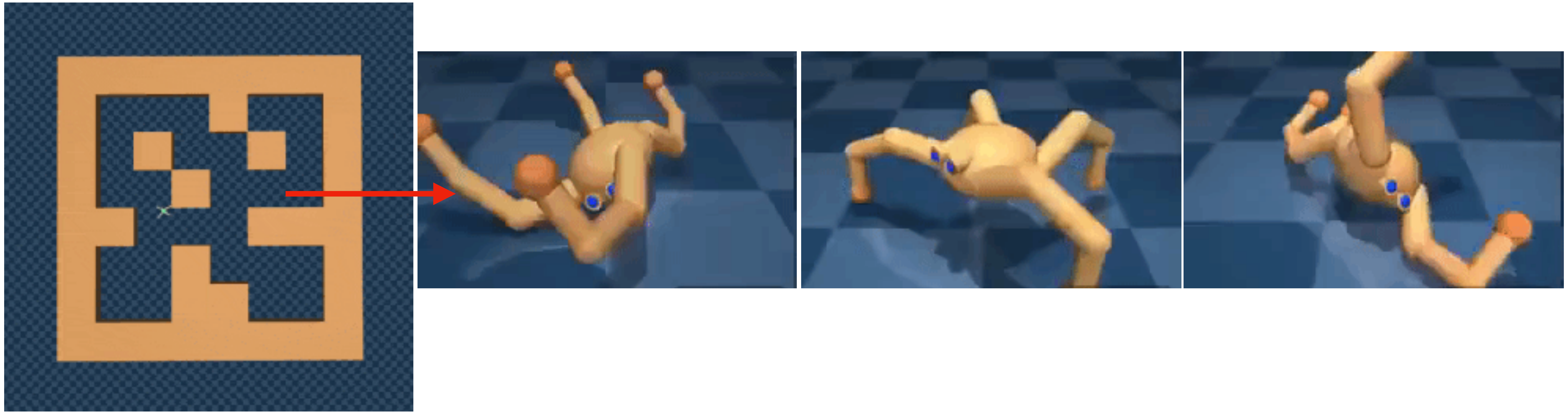
Behavioral cloning:

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

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

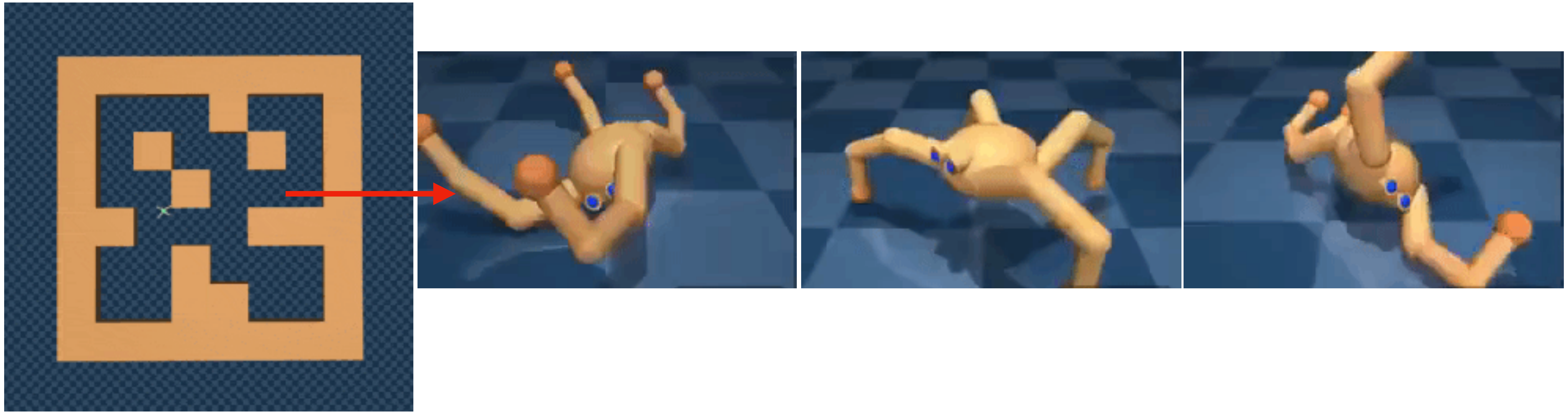
Suboptimal Offline Data

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



Suboptimal Offline Data

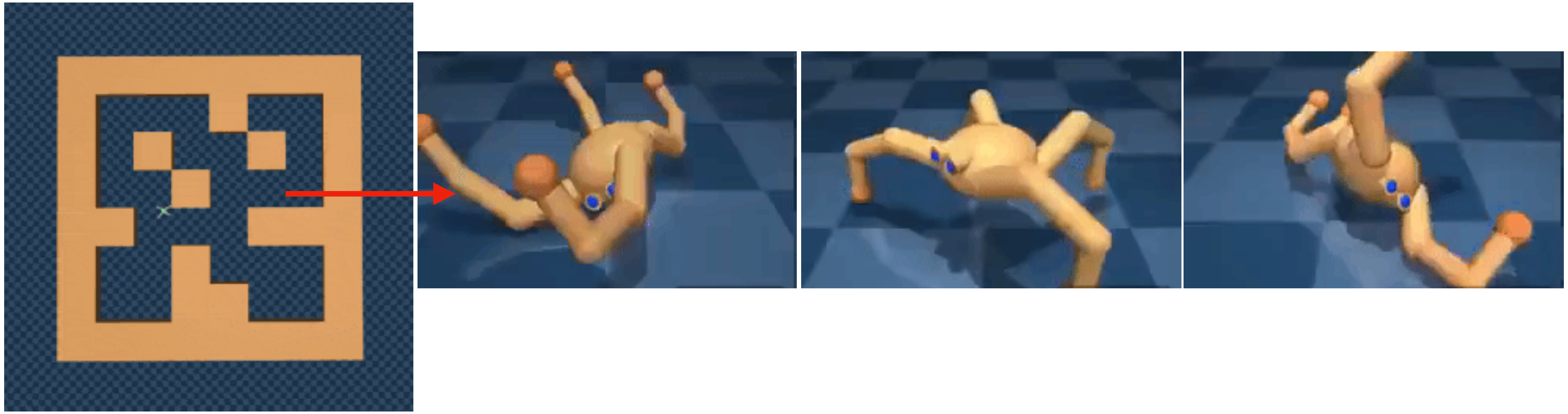
Large amounts of suboptimal offline data $\mathscr{D}^{off}$



How can $\mathscr{D}^{off}$ facilitate imitation learning?

Suboptimal Offline Data

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



How can $\mathscr{D}^{off}$ facilitate imitation learning?

- Directly imitate $\mathscr{D}^{off}$?

Suboptimal Offline Data

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

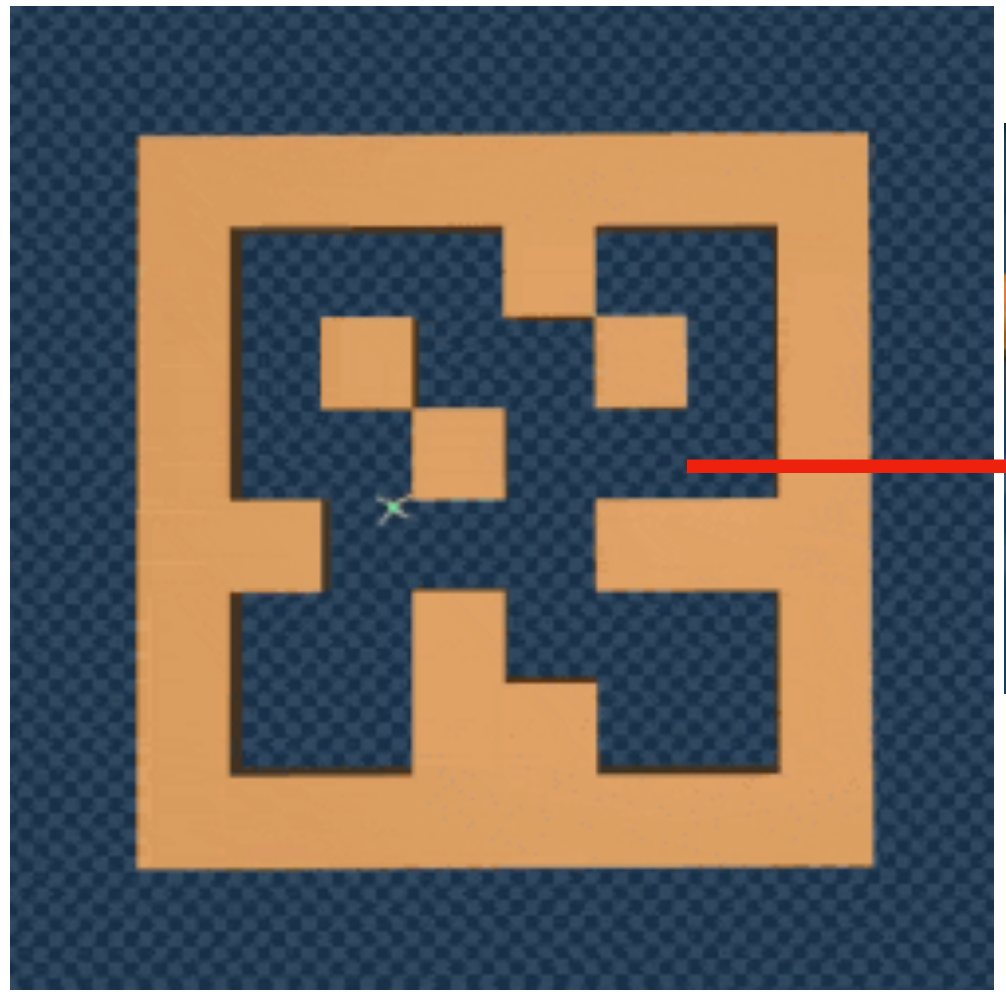
Highly suboptimal
(e.g., random policy)

How can $\mathscr{D}^{off}$ facilitate imitation learning?

- Directly imitate $\mathscr{D}^{off}$?

Suboptimal Offline Data

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



Highly suboptimal
(e.g., random policy)

How can $\mathscr{D}^{off}$ facilitate imitation learning?

- Directly imitate $\mathscr{D}^{off}$?
- Run offline RL on $\mathscr{D}^{off}$?

Suboptimal Offline Data

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

Highly suboptimal
(e.g., random policy)

How can $\mathscr{D}^{off}$ facilitate imitation learning?

- Directly imitate $\mathscr{D}^{off}$?
- Run offline RL on $\mathscr{D}^{off}$? Requires reward signal

Suboptimal Offline Data

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

Highly suboptimal
(e.g., random policy)

How can $\mathscr{D}^{off}$ facilitate imitation learning?

- Directly imitate $\mathscr{D}^{off}$?
- Run offline RL on $\mathscr{D}^{off}$? Requires reward signal
- Extract latent skills from $\mathscr{D}^{off}$ showing what could be done.

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D}, z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log\pi_\theta(a_t|s_t,z)\right]$$

Ajay et. al. 2021

Hakhamaneshi et. al. 2021

Pertsch et. al. 2020

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

Ajay et. al. 2021
Hakhamaneshi et. al. 2021
Pertsch et. al. 2020

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D}, z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log \pi_\theta(a_t|s_t, z)\right]$$

$$\text{s.t.} \;\; \hat{\mathbb{E}}_{\tau\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(q_\phi(z|\tau)||\rho_\omega(z|s_0))] \leq \epsilon_{\mathrm{KL}}$$

with some regularizer over skill prior $p(z)$

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

Ajay et. al. 2021
Hakhamaneshi et. al. 2021
Pertsch et. al. 2020

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D}, z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log \pi_\theta(a_t|s_t,z)\right]$$

$$\text{s.t. } \hat{\mathbb{E}}_{\tau\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(q_\phi(z|\tau)||\rho_\omega(z|s_0))] \le \epsilon_{\mathrm{KL}}$$

with some regularizer over skill prior $p(z)$

- Relies on $\mathscr{D}^{off}$ already have good / diverse behavior

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

Ajay et. al. 2021
Hakhamaneshi et. al. 2021
Pertsch et. al. 2020

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D},z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log \pi_\theta(a_t|s_t,z)\right]$$

$$\text{s.t. } \hat{\mathbb{E}}_{\tau\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(q_\phi(z|\tau)||\rho_\omega(z|s_0))] \le \epsilon_{\mathrm{KL}}$$

with some regularizer over skill prior $p(z)$

- Relies on $\mathscr{D}^{off}$ already have good / diverse behavior

Degenerate latent mode

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D}, z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log \pi_\theta(a_t|s_t,z)\right]$$

$$\text{s.t. } \hat{\mathbb{E}}_{\tau\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(q_\phi(z|\tau)||\rho_\omega(z|s_0))] \le \epsilon_{\mathrm{KL}}$$

Ajay et. al. 2021

Hakhamaneshi et. al. 2021

Pertsch et. al. 2020

with some regularizer over skill prior $p(z)$

- Relies on $\mathscr{D}^{off}$ already have good / diverse behavior

  Degenerate latent mode

- Benefit attributed to increased temporal abstraction.

Previously: Latent Skill Extraction

Max-likelihood learning of latent skills z (e.g. OPAL, SPiRL)

Ajay et. al. 2021
Hakhamaneshi et. al. 2021
Pertsch et. al. 2020

$$\min_{\theta,\phi,\omega} J(\theta,\phi,\omega) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D},z\sim q_\phi(z|\tau)}\left[-\sum_{t=0}^{c-1}\log\pi_\theta(a_t|s_t,z)\right]$$

$$\text{s.t. } \hat{\mathbb{E}}_{\tau\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(q_\phi(z|\tau)||\rho_\omega(z|s_0))] \leq \epsilon_{\mathrm{KL}}$$

with some regularizer over skill prior $p(z)$

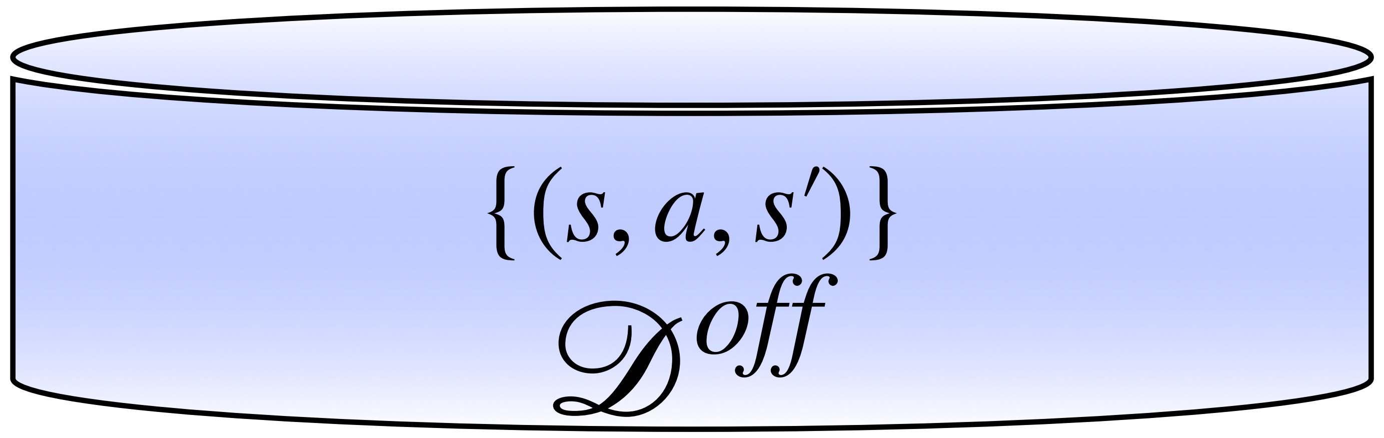
- Relies on $\mathscr{D}^{off}$ already have good / diverse behavior

  Degenerate latent mode

- Benefit attributed to increased temporal abstraction.

  Can we benefit from a “simpler” action space (even for a single step model)?

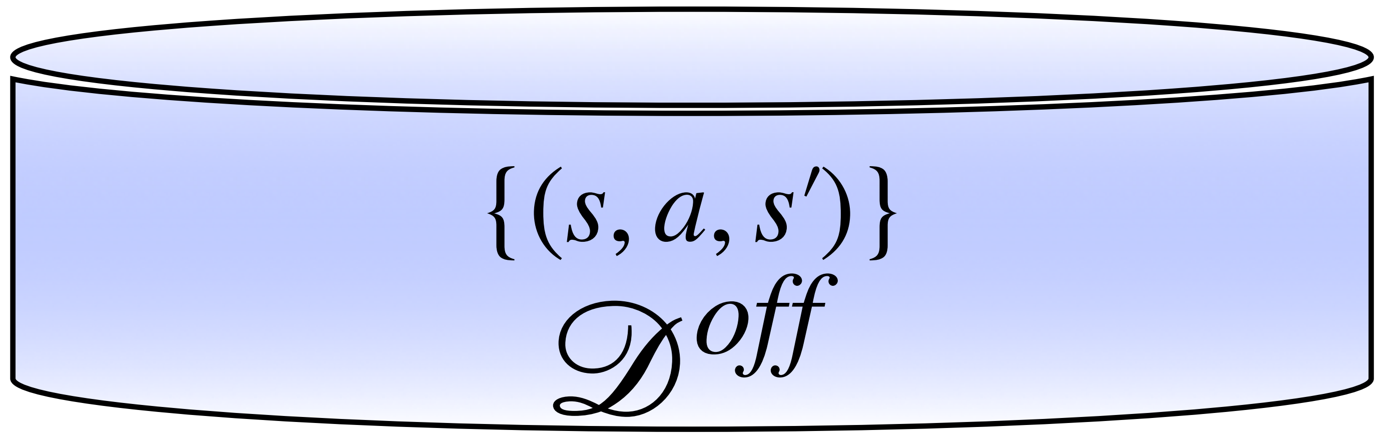
TRAIL: Transition Reparametrized Actions



TRAIL: Transition Reparametrized Actions

Factored transition model

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



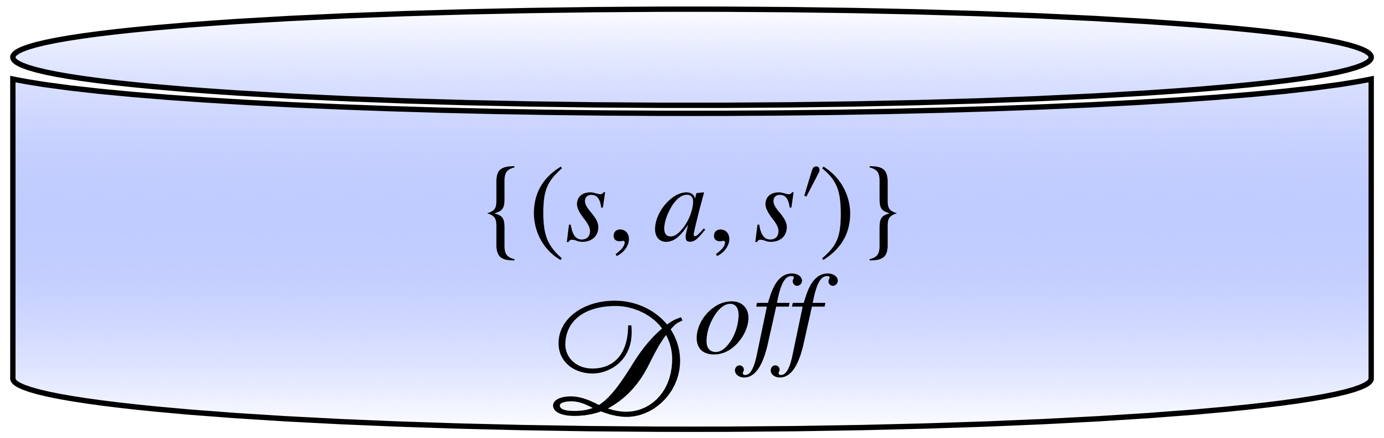
Pretraining

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

TRAIL: Transition Reparametrized Actions

Factored transition model

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



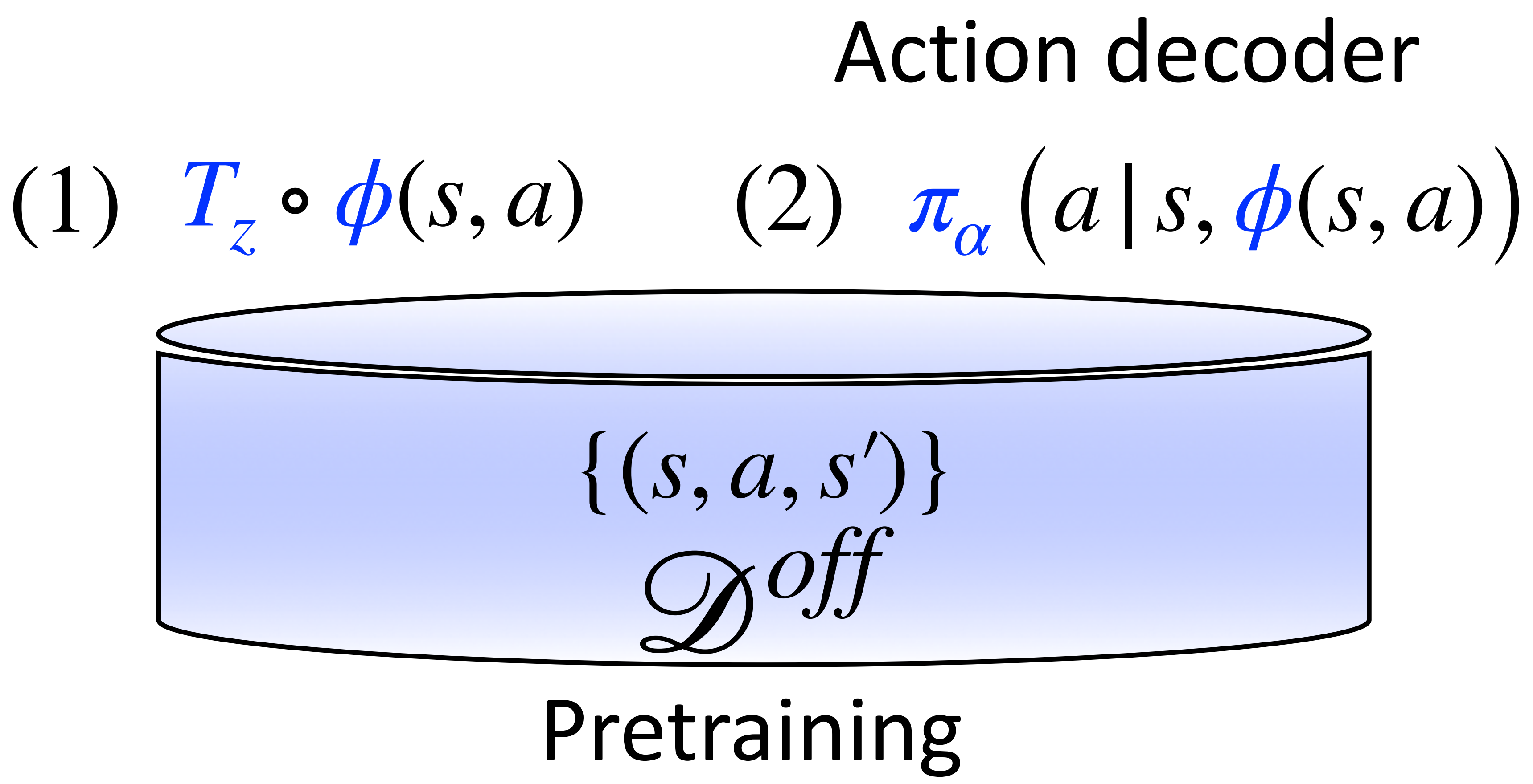
Pretraining

$S \times Z \to \Delta(S)$

Pretraining

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

TRAIL: Transition Reparametrized Actions

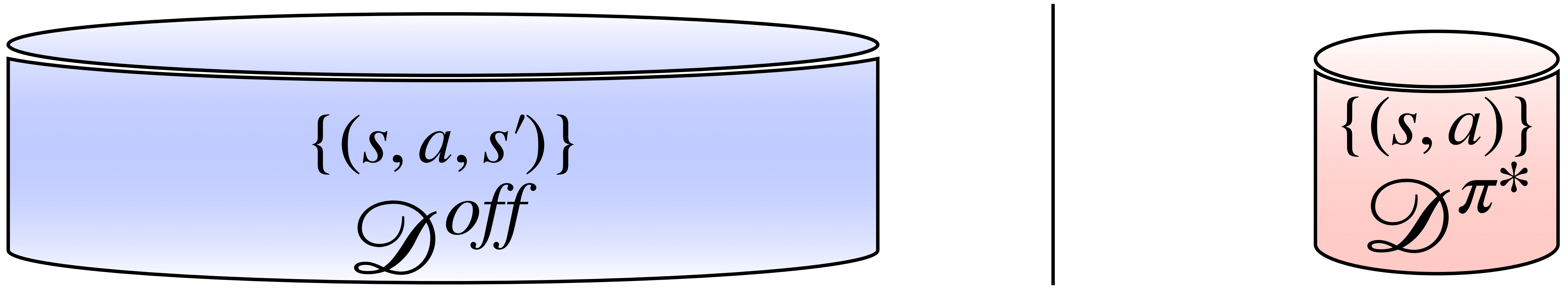


$$
\textit{Pretraining}\;
\begin{cases}
\underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\text{KL}}(\mathcal{T}(s,a)\|\mathcal{T}_Z(s,\phi(s,a)))\right]}_{=\,J_{\text{T}}(\mathcal{T}_Z,\phi)} & (1) \\[2ex]
\underbrace{\mathbb{E}_{s\sim d^{\text{off}}}[\max_{z\in Z} D_{\text{KL}}(\pi_{\alpha^*}(s,z)\|\pi_\alpha(s,z))]}_{\approx\;\text{const}(d^{\text{off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

TRAIL: Transition Reparametrized Actions

Latent policy

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



Pretraining Downstream Imitation

Pretraining

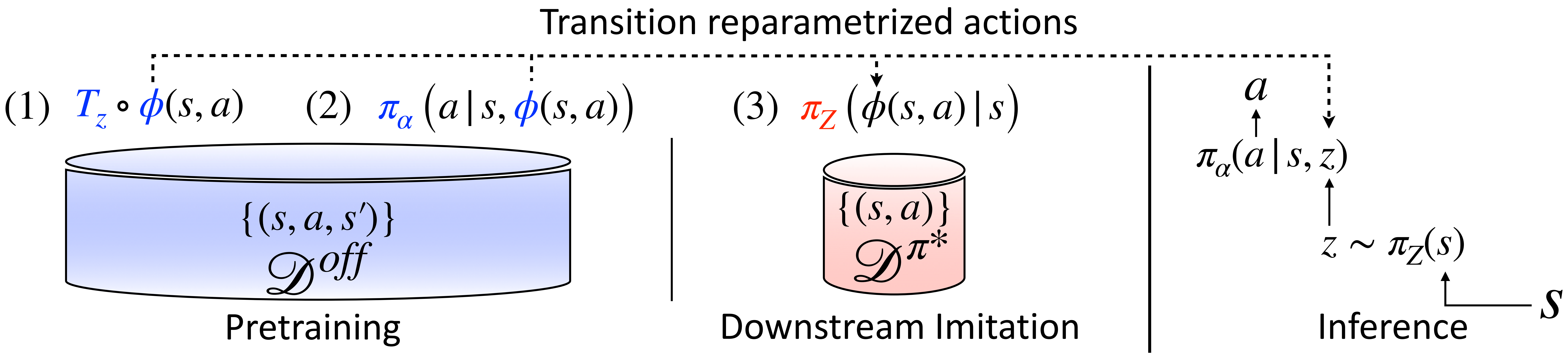
$$\underbrace{\mathbb{E}_{(s,a)\sim d^{\mathrm{off}}}\left[D_{\mathrm{KL}}(\mathcal{T}(s,a)\|\mathcal{T}_Z(s,\phi(s,a)))\right]}_{=\,J_{\mathrm{T}}(\mathcal{T}_Z,\phi)} \tag{1}$$

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

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

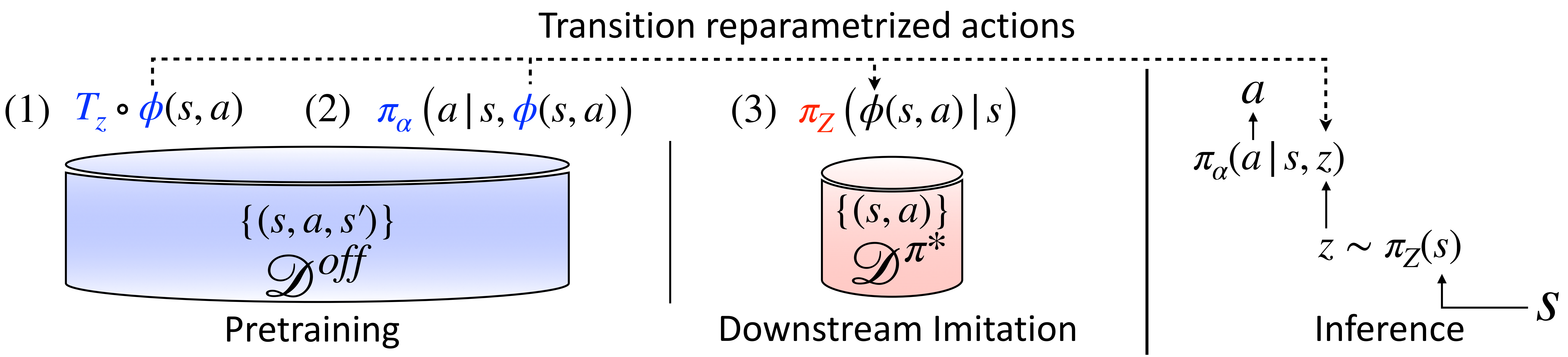
$$\underbrace{\mathbb{E}_{(s,a)\sim d^{\text{off}}}\left[D_{\text{KL}}(\mathcal{T}(s,a)\|\mathcal{T}_Z(s,\phi(s,a)))\right]}_{=J_{\text{T}}(\mathcal{T}_Z,\phi)} \tag{1}$$

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

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



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

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

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

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

TRAIL Derivation Overview

$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$

TRAIL Derivation Overview

$$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$$
$$\leq \frac{\gamma}{1-\gamma} \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \mathcal{T})$$

Near-optimal representation learning, Nachum et. al.

$$\frac{1}{2} \sum_{s' \in S} \left| \mathbb{E}_{s \sim d^{\pi_1}, a_1 \sim \pi_1(s), a_2 \sim \pi_2(s)} [\mathcal{T}(s'|s, a_1) - \mathcal{T}(s'|s, a_2)] \right|$$

$$D_{\mathrm{TV}}(\mathcal{T} \circ \pi_1 \circ d^{\pi_1} \| \mathcal{T} \circ \pi_2 \circ d^{\pi_1})$$

TRAIL Derivation Overview

$$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$$
$$\leq \frac{\gamma}{1-\gamma} \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \mathcal{T})$$

Near-optimal representation learning, Nachum et. al.

$$\frac{1}{2} \sum_{s' \in S} \left| \mathbb{E}_{s \sim d^{\pi_1}, a_1 \sim \pi_1(s), a_2 \sim \pi_2(s)} [\mathcal{T}(s'|s, a_1) - \mathcal{T}(s'|s, a_2)] \right|$$

$$D_{\mathrm{TV}}(\mathcal{T} \circ \pi_1 \circ d^{\pi_1} \| \mathcal{T} \circ \pi_2 \circ d^{\pi_1})$$

$$\leq |A| \mathbb{E}_{(s,a) \sim (d^{\pi_1}, \mathrm{Unif}_A)} [\boxed{D_{\mathrm{TV}}(\mathcal{T}(s,a) \| \overline{\mathcal{T}}(s,a))}] + \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \overline{\mathcal{T}})$$

TRAIL Derivation Overview

Near-optimal representation learning, Nachum et. al.

$$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$$
$$\leq \frac{\gamma}{1-\gamma} \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \mathcal{T})$$

$$\frac{1}{2} \sum_{s' \in S} \left| \mathbb{E}_{s \sim d^{\pi_1}, a_1 \sim \pi_1(s), a_2 \sim \pi_2(s)} [\mathcal{T}(s'|s, a_1) - \mathcal{T}(s'|s, a_2)] \right|$$

$$D_{\mathrm{TV}}(\mathcal{T} \circ \pi_1 \circ d^{\pi_1} \| \mathcal{T} \circ \pi_2 \circ d^{\pi_1})$$

$$\leq |A| \mathbb{E}_{(s,a) \sim (d^{\pi_1}, \mathrm{Unif}_A)} [\boxed{D_{\mathrm{TV}}(\mathcal{T}(s,a) \| \overline{\mathcal{T}}(s,a))}] + \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \overline{\mathcal{T}})$$

$$\pi_{k,Z}(z|s) := \sum_{a \in A, z = \phi(s,a)} \pi_k(a|s)$$

$$\leq \mathbb{E}_{s \sim d^{\pi_1}} [D_{\mathrm{TV}}(\pi_{1,Z} \| \pi_{2,Z})]$$

TRAIL Derivation Overview

[Near-optimal representation learning](), Nachum et. al.

$$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$$
$$\leq \frac{\gamma}{1-\gamma} \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \mathcal{T})$$

$$\frac{1}{2} \sum_{s' \in S} \left| \mathbb{E}_{s \sim d^{\pi_1}, a_1 \sim \pi_1(s), a_2 \sim \pi_2(s)} [\mathcal{T}(s'|s, a_1) - \mathcal{T}(s'|s, a_2)] \right|$$

$$D_{\mathrm{TV}}(\mathcal{T} \circ \pi_1 \circ d^{\pi_1} \| \mathcal{T} \circ \pi_2 \circ d^{\pi_1})$$

$$\leq |A| \mathbb{E}_{(s,a) \sim (d^{\pi_1}, \mathrm{Unif}_A)} [\boxed{D_{\mathrm{TV}}(\mathcal{T}(s,a) \| \overline{\mathcal{T}}(s,a))}] + \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \overline{\mathcal{T}})$$

$$\pi_{k,Z}(z|s) := \sum_{a \in A, z = \phi(s,a)} \pi_k(a|s)$$

$$\leq \mathbb{E}_{s \sim d^{\pi_1}} [D_{\mathrm{TV}}(\pi_{1,Z} \| \pi_{2,Z})]$$

$$D_{\mathrm{TV}}(\pi_{1,Z}(s) \| \pi_{\alpha,Z}(s)) \leq \max_{z \in Z} D_{\mathrm{TV}}(\pi_\alpha(s,z) \| \pi_{\alpha^*}(s,z)) + D_{\mathrm{TV}}(\pi_{1,Z}(s) \| \pi_Z(s))$$

$$\pi_{\alpha,Z}(z|s) := \sum_{a \in A, z = \phi(s,a)} (\pi_\alpha \circ \pi_Z)(a|s)$$

TRAIL Derivation Overview

Near-optimal representation learning, Nachum et. al.

$$\mathrm{Diff}(\pi_2, \pi_1) = D_{\mathrm{TV}}(d^{\pi_2} \| d^{\pi_1})$$

$$\leq \frac{\gamma}{1-\gamma} \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \mathcal{T})$$

$$\frac{1}{2} \sum_{s' \in S} \left| \mathbb{E}_{s \sim d^{\pi_1}, a_1 \sim \pi_1(s), a_2 \sim \pi_2(s)} [\mathcal{T}(s'|s, a_1) - \mathcal{T}(s'|s, a_2)] \right|$$

$$D_{\mathrm{TV}}(\mathcal{T} \circ \pi_1 \circ d^{\pi_1} \| \mathcal{T} \circ \pi_2 \circ d^{\pi_1})$$

$$\leq |A| \mathbb{E}_{(s,a) \sim (d^{\pi_1}, \mathrm{Unif}_A)} [\boxed{D_{\mathrm{TV}}(\mathcal{T}(s,a) \| \overline{\mathcal{T}}(s,a))}] + \mathrm{Err}_{d^{\pi_1}}(\pi_1, \pi_2, \overline{\mathcal{T}})$$

$$\pi_{k,Z}(z|s) := \sum_{a \in A, z = \phi(s,a)} \pi_k(a|s)$$

$$\leq \mathbb{E}_{s \sim d^{\pi_1}} [D_{\mathrm{TV}}(\pi_{1,Z} \| \pi_{2,Z})]$$

$$D_{\mathrm{TV}}(\pi_{1,Z}(s) \| \pi_{\alpha,Z}(s)) \leq \max_{z \in Z} \boxed{D_{\mathrm{TV}}(\pi_\alpha(s,z) \| \pi_{\alpha^*}(s,z))} + \boxed{D_{\mathrm{TV}}(\pi_{1,Z}(s) \| \pi_Z(s))}$$

$$\pi_{\alpha,Z}(z|s) := \sum_{a \in A, z = \phi(s,a)} (\pi_\alpha \circ \pi_Z)(a|s)$$

TRAIL Derivation Overview

Near-optimal representation learning, Nachum et. al.

$$\mathrm{Diff}(\pi_2,\pi_1) = D_{\mathrm{TV}}(d^{\pi_2}\|d^{\pi_1}) \le \frac{\gamma}{1-\gamma}\mathrm{Err}_{d^{\pi_1}}(\pi_1,\pi_2,\mathcal{T})$$

$$\frac{1}{2}\sum_{s'\in S}\left|\mathbb{E}_{s\sim d^{\pi_1},a_1\sim\pi_1(s),a_2\sim\pi_2(s)}[\mathcal{T}(s'|s,a_1)-\mathcal{T}(s'|s,a_2)]\right|$$

$$D_{\mathrm{TV}}(\mathcal{T}\circ\pi_1\circ d^{\pi_1}\|\mathcal{T}\circ\pi_2\circ d^{\pi_1})$$

$$\le |A|\mathbb{E}_{(s,a)\sim(d^{\pi_1},\mathrm{Unif}_A)}[\boxed{D_{\mathrm{TV}}(\mathcal{T}(s,a)\|\overline{\mathcal{T}}(s,a))}] + \mathrm{Err}_{d^{\pi_1}}(\pi_1,\pi_2,\overline{\mathcal{T}})$$

$$\pi_{k,Z}(z|s) := \sum_{a\in A,z=\phi(s,a)}\pi_k(a|s)$$

$$\le \mathbb{E}_{s\sim d^{\pi_1}}[D_{\mathrm{TV}}(\pi_{1,Z}\|\pi_{2,Z})]$$

$$D_{\mathrm{TV}}(\pi_{1,Z}(s)\|\pi_{\alpha,Z}(s)) \le \max_{z\in Z}\boxed{D_{\mathrm{TV}}(\pi_\alpha(s,z)\|\pi_{\alpha^*}(s,z))} + \boxed{D_{\mathrm{TV}}(\pi_{1,Z}(s)\|\pi_Z(s))}$$

$$\pi_{\alpha,Z}(z|s) := \sum_{a\in A,z=\phi(s,a)}(\pi_\alpha\circ\pi_Z)(a|s)$$

Lastly, the on-policy to off-policy translation: $\mathbb{E}_{\rho_1}[h(s)] \le (1+D_{\chi^2}(\rho_1\|\rho_2)^{\frac{1}{2}})\sqrt{\mathbb{E}_{\rho_2}[h(s)^2]}$

TRAIL's Sample complexity

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

TRAIL's Sample complexity

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

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

TRAIL's Sample complexity

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

Can be further reduced by state representation learning

Pretraining

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

TRAIL's Sample complexity

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

Can be further reduced
by state representation learning

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

So far, our analysis is based on tabular actions.

What about continuous actions and stochastic expert policy?

TRAIL with Linear Transition Dynamics

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

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

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

TRAIL with Linear Transition Dynamics

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

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

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

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

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

Recall tabular:

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

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

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

easier to optimize

Recall tabular:

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

Learning TRAIL in Practice

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

Learning TRAIL in Practice

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

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

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

Learning TRAIL in Practice

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

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

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

Random features for large-scale kernel machines Rahimi et al., 2007)

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

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

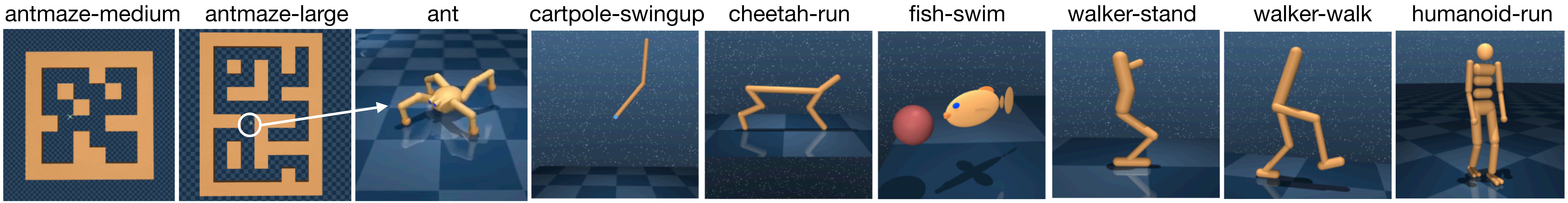
TRAIL linear: $\overline{\mathcal{T}}(s'|s,a) \propto \rho(s')\exp\{-||f(s,a) - g(s')||^2/2\} \propto \bar{\psi}(s')^\top \bar{\phi}(s,a)$

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

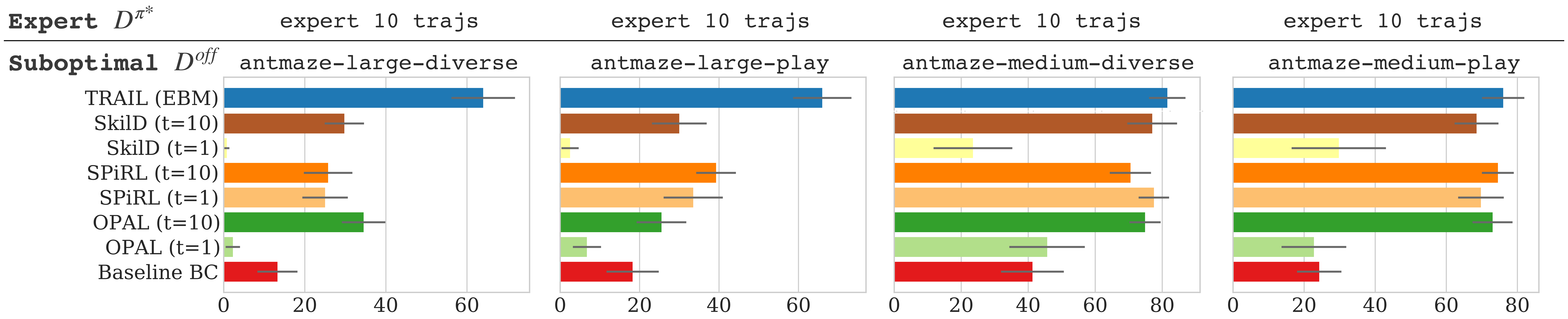
Random features for large-scale kernel machines Rahimi et al., 2007)

π_α and π_Z are neural-network parametrized Guassian policies.

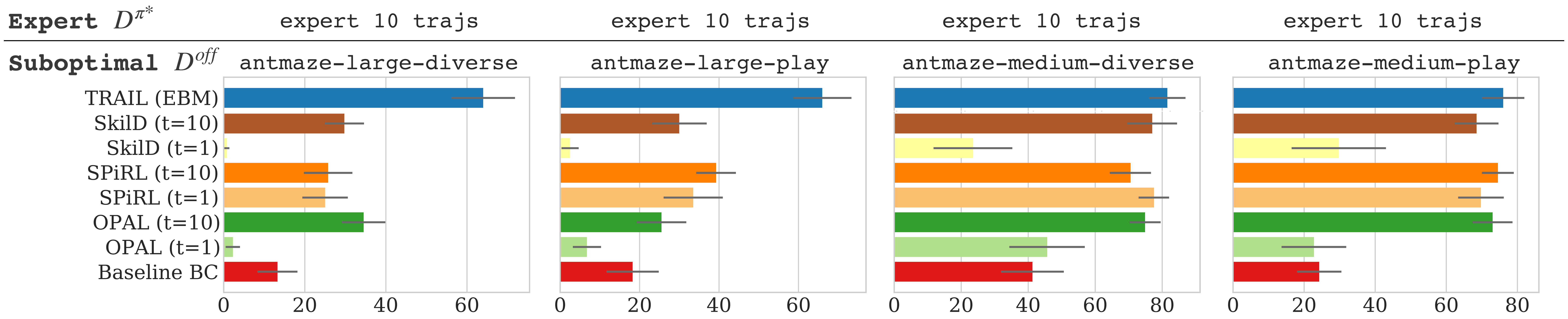
Experiments

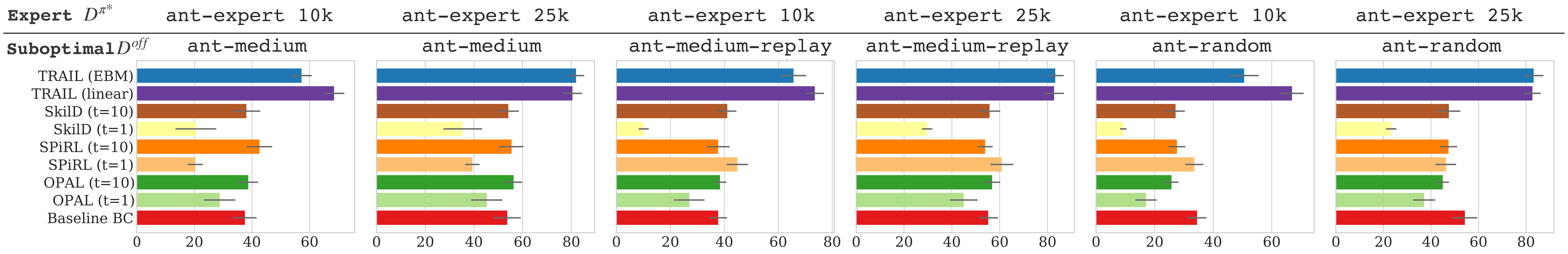


Experiments

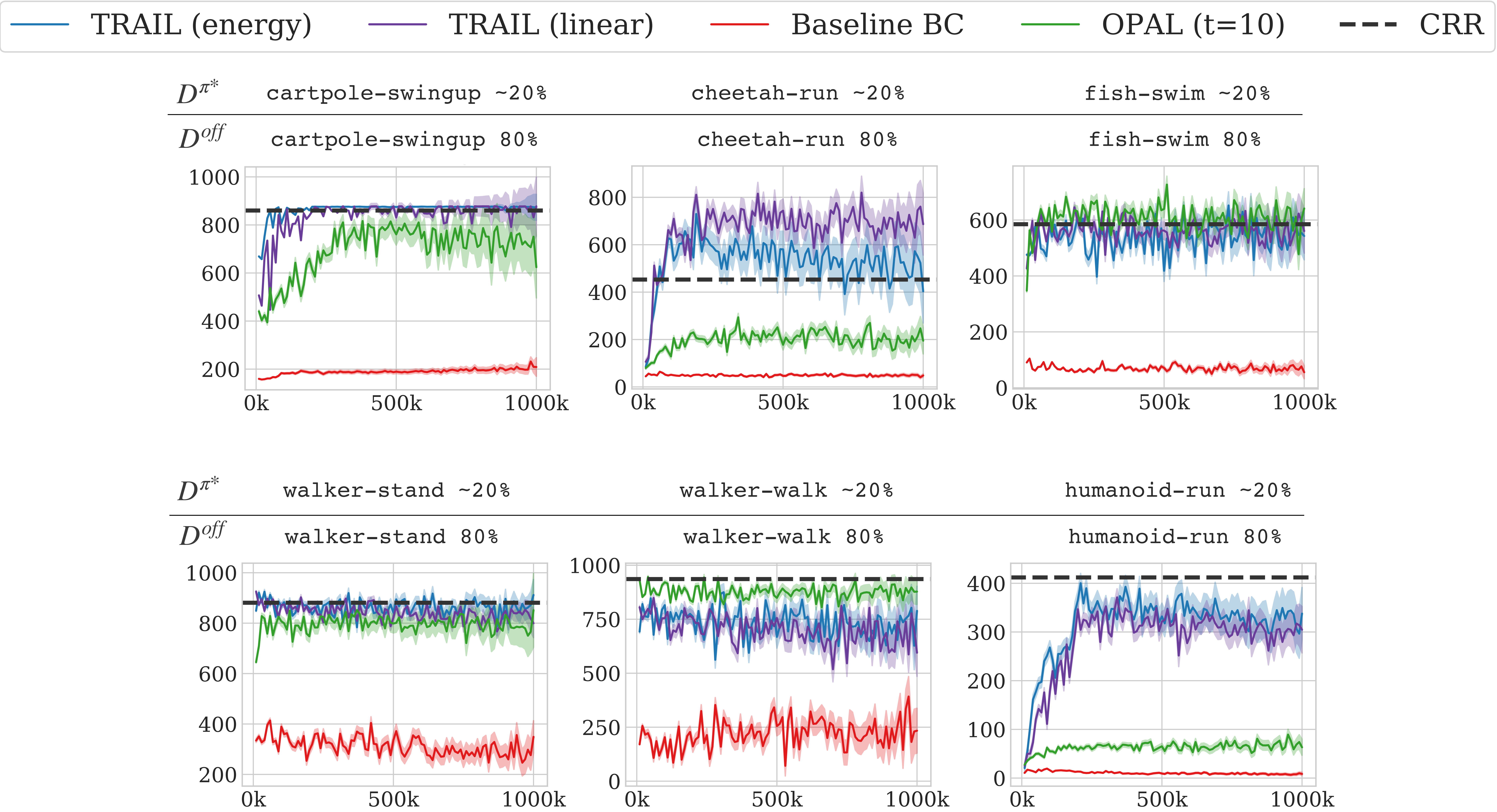


Experiments





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.

More on representation learning for RL / IL

- Representation Matters: Offline Pretraining for Sequential Decision Making
  - Empirical study where this started from
- Provable Representation Learning for Imitation with Contrastive Fourier Features
  - Provable state representation learning

More on representation learning for RL / IL

- Representation Matters: Offline Pretraining for Sequential Decision Making
  - Empirical study where this started from
- Provable Representation Learning for Imitation with Contrastive Fourier Features
  - Provable state representation learning

Thank you. Checkout

Paper: http://arxiv.org/abs/2110.14770
Code: https://github.com/google-research/google-research/tree/master/rl_repr
Website: https://sites.google.com/corp/view/trail-repr