Mutual Information State Intrinsic Control

<u>Rui Zhao</u>[1,2], Yang Gao[3], Pieter Abbeel[4], Volker Tresp[1,2], Wei Xu[5]
1. Ludwig Maximilian University of Munich, 2. Siemens AG, 3. Tsinghua University, 4. UC Berkeley, 5. Horizon Robotics

Mutual Information State Intrinsic Control

International Conference on Learning Representations (ICLR) 2021, Spotlight

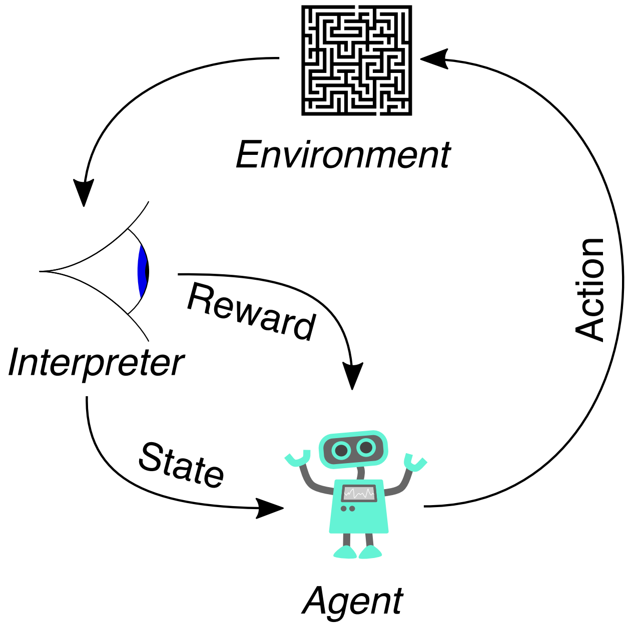
Motivation and Contribution

- Learning by interacting with the environment
  - Like the way humans learn

- Self-consciousness in psychology
  - the agent knows what constitutes itself

- Propose a new intrinsic objective
  - encourage the agent to have maximum control on the environment.

- Outperform previous methods
  - complete the pick-and-place task for the first time without using any task reward.



Mutual Information State Intrinsic Control

International Conference on Learning Representations (ICLR) 2021

Mutual Information Reward Function

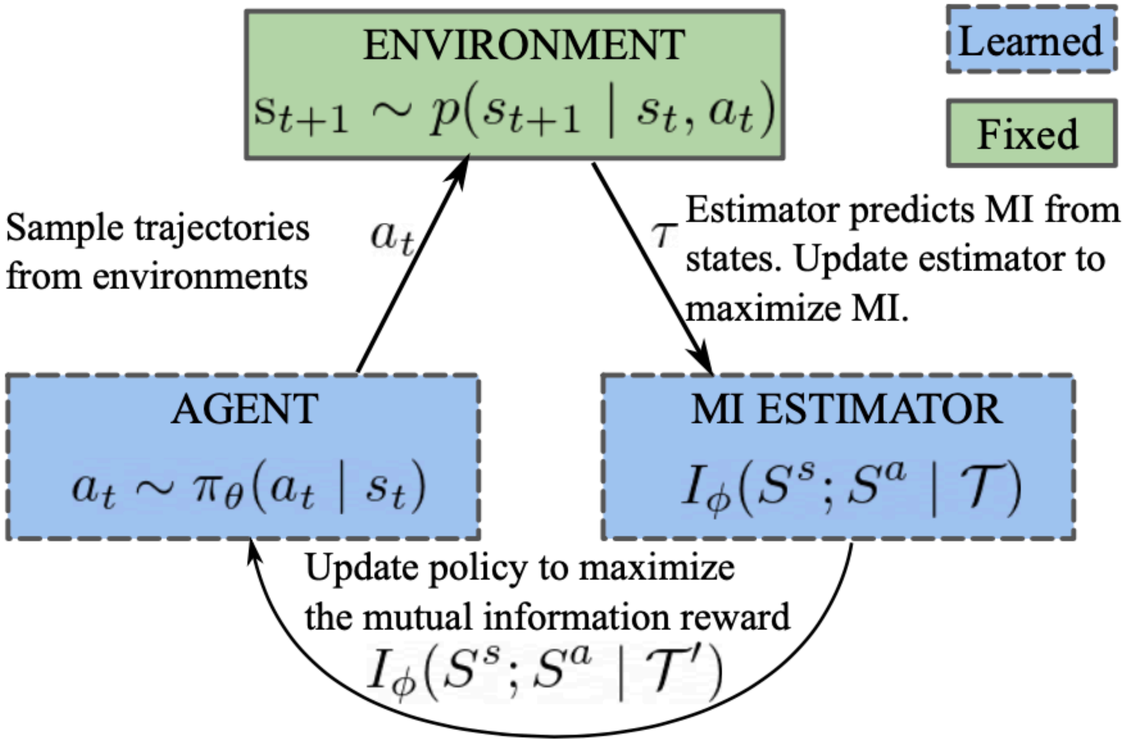
$$
\begin{aligned}
I(S^s; S^a) &= KL(\mathbb{P}_{S^s S^a} \,||\, \mathbb{P}_{S^s} \otimes \mathbb{P}_{S^a}) \\
&= \sup_{T:\Omega \to \mathbb{R}} \mathbb{E}_{\mathbb{P}_{S^s S^a}}[T] - \log(\mathbb{E}_{\mathbb{P}_{S^s} \otimes \mathbb{P}_{S^a}}[e^T]) \\
&\geq \sup_{\phi \in \Phi} \mathbb{E}_{\mathbb{P}_{S^s S^a}}[T_\phi] - \log(\mathbb{E}_{\mathbb{P}_{S^s} \otimes \mathbb{P}_{S^a}}[e^{T_\phi}]) := I_\Phi(S^s; S^a)
\end{aligned}
$$

Effectively Computing the Mutual Information Reward in Practice

$$
I_\phi(S^s; S^a \mid \mathcal{T}) \propto \mathbb{E}_{\mathbb{P}_{\mathcal{T}'}}[I_\phi(S^s; S^a \mid \mathcal{T}')]
$$

$$
r_\phi(a_t, s_t) := I_\phi(S^s; S^a | \mathcal{T}') = 0.5 \sum_{i=t}^{t+1} T_\phi(s_i^s, s_i^a) - \log(0.5 \sum_{i=t}^{t+1} e^{T_\phi(s_i^s, \bar{s}_i^a)})
$$

Mutual Information State Intrinsic Control
MUSIC Algorithm



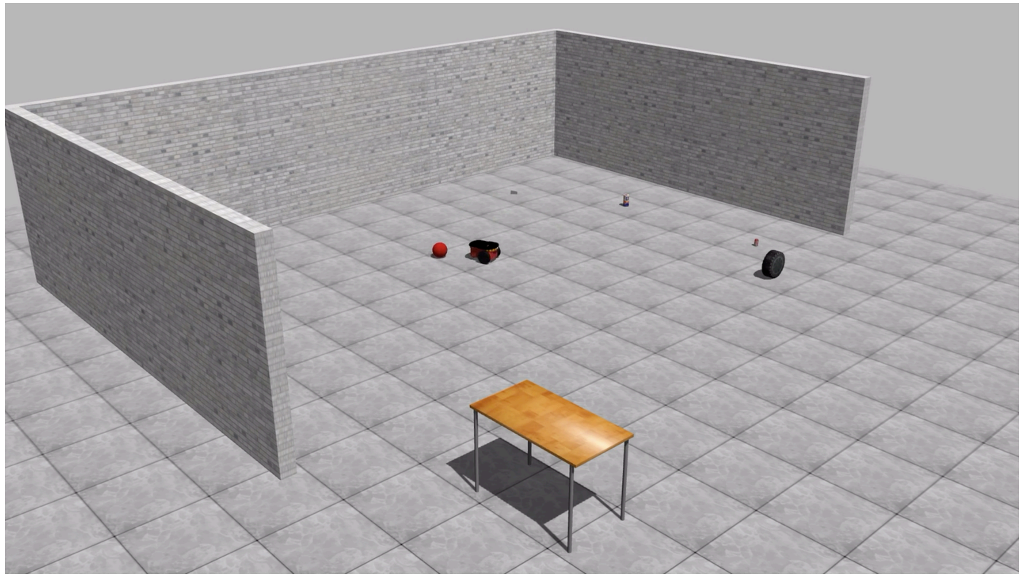
Algorithm 1: MUSIC

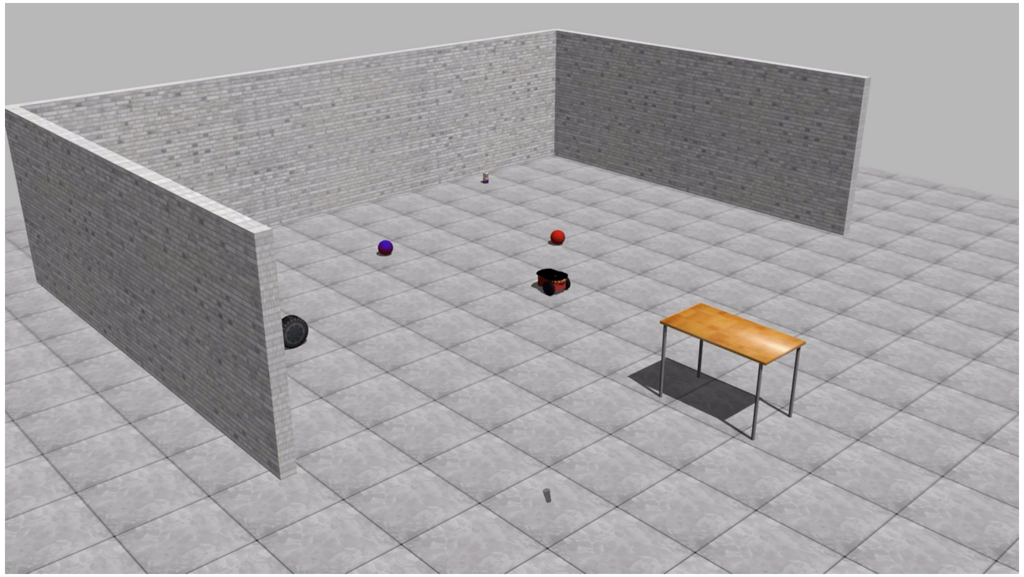
while *not converged* **do**
- Sample an initial state $s_0 \sim p(s_0)$.
- **for** $t \leftarrow 1$ **to** *steps_per_episode* **do**
  - Sample action $a_t \sim \pi_\theta(a_t \mid s_t)$.
  - Step environment $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$.
  - Sample transitions $\mathcal{T}'$ from the buffer.
  - Set intrinsic reward $r = I_\phi(S^s; S^a \mid \mathcal{T}')$.
  - Update policy (θ) via DDPG or SAC.
  - Update the MI estimator (ϕ) with SGD.

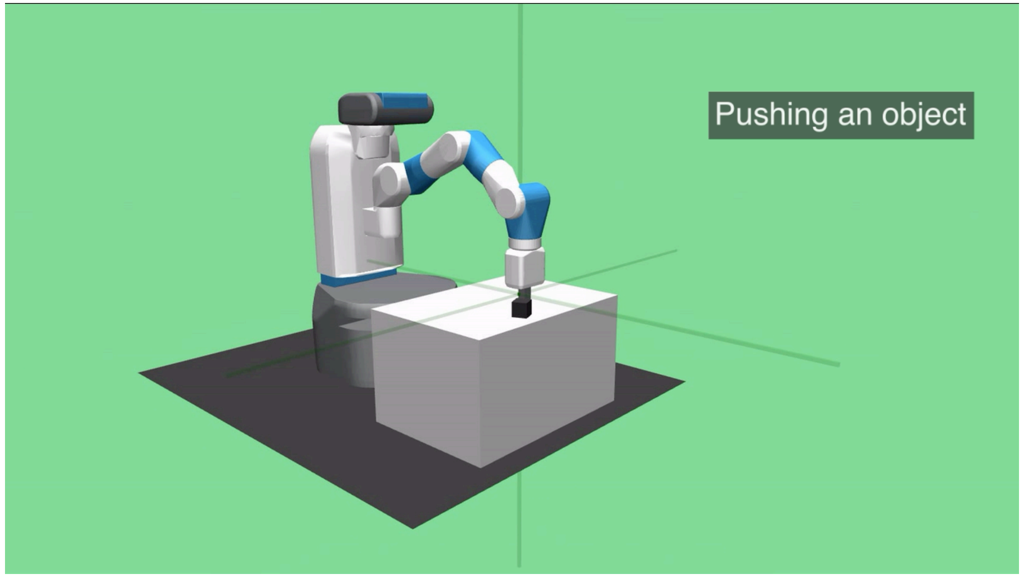
MUSIC Algorithm: We update the estimator to better predict the MI, and update the agent to control the surrounding state to have higher MI with the agent state.

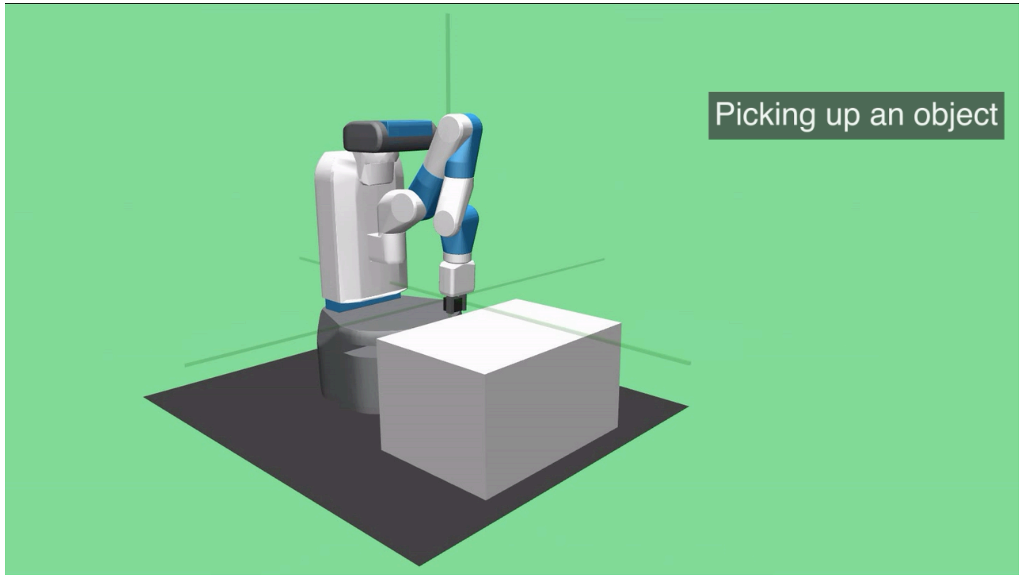
Mutual Information State Intrinsic Control

Unsupervised learned behaviour



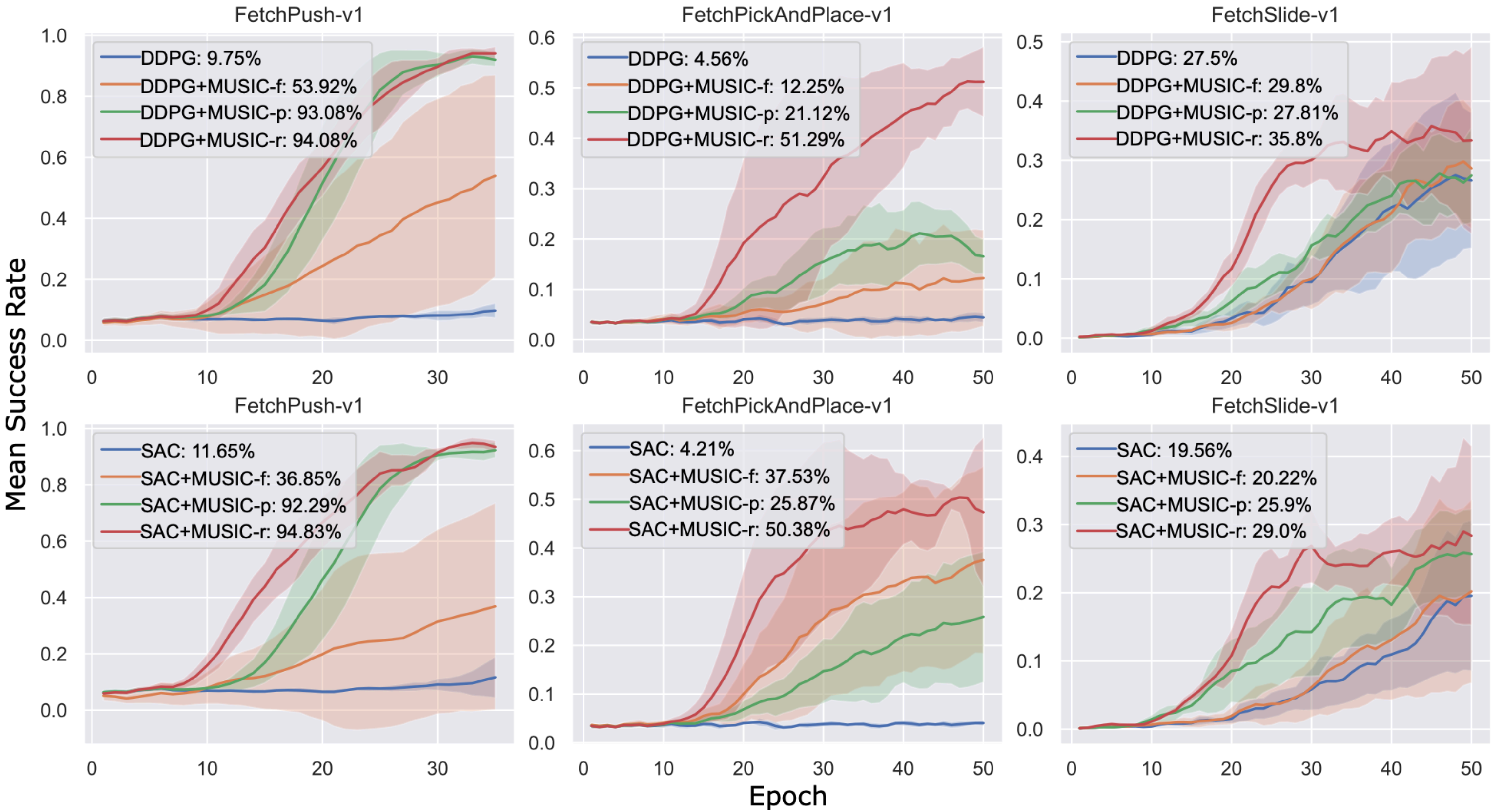


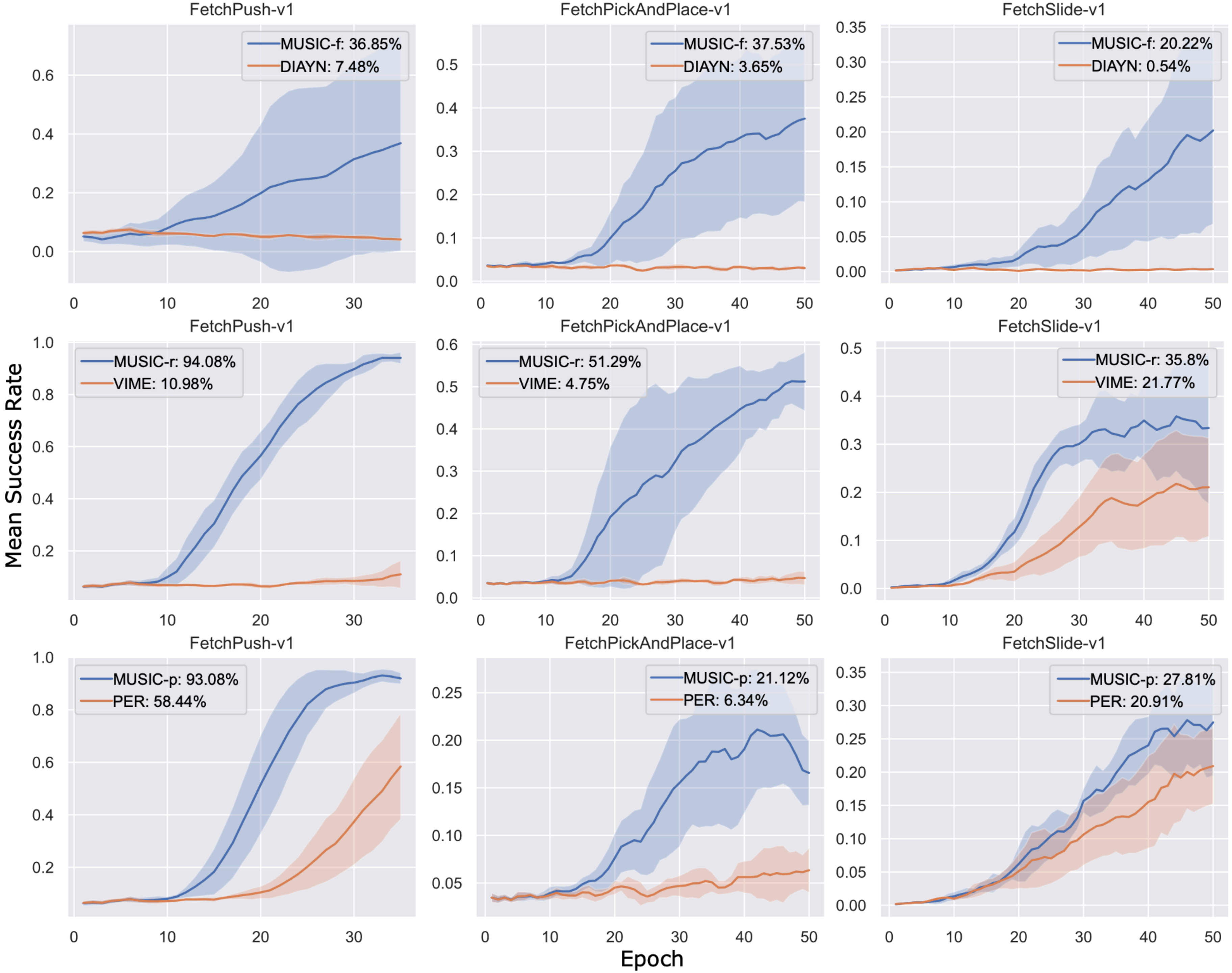




Mutual-Information State Intrinsic Control (MUSIC)

MUSIC combined with task rewards via: fine-tuning, prioritization, and intrinsic reward



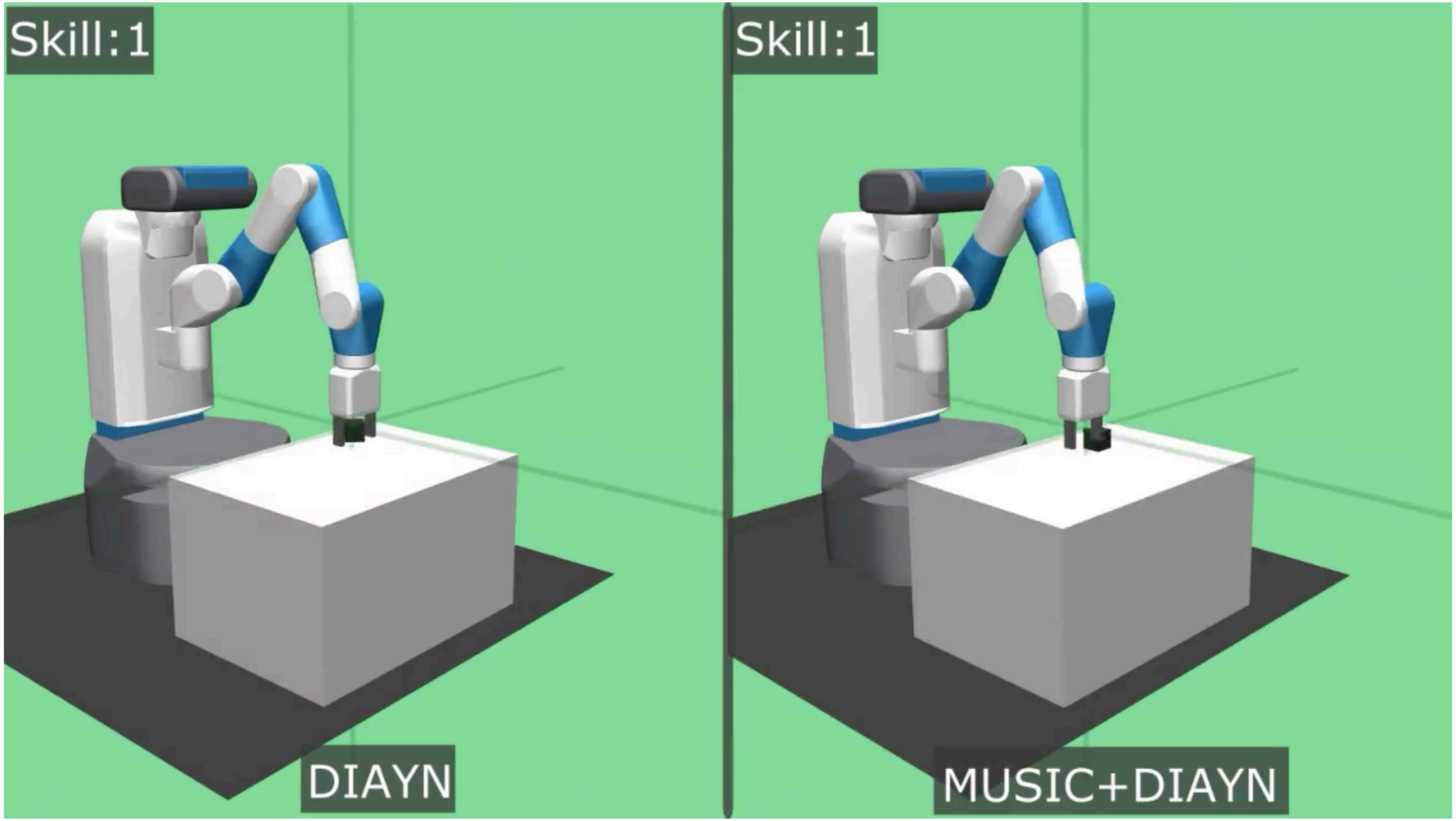


FetchPush-v1
MUSIC-f: 36.85%
DIAYN: 7.48%
FetchPickAndPlace-v1
MUSIC-f: 37.53%
DIAYN: 3.65%
FetchSlide-v1
MUSIC-f: 20.22%
DIAYN: 0.54%
FetchPush-v1
MUSIC-r: 94.08%
VIME: 10.98%
FetchPickAndPlace-v1
MUSIC-r: 51.29%
VIME: 4.75%
FetchSlide-v1
MUSIC-r: 35.8%
VIME: 21.77%
FetchPush-v1
MUSIC-p: 93.08%
PER: 58.44%
FetchPickAndPlace-v1
MUSIC-p: 21.12%
PER: 6.34%
FetchSlide-v1
MUSIC-p: 27.81%
PER: 20.91%
Mean Success Rate
Epoch

Mutual Information State Intrinsic Control

Compare MUSIC with DIAYN

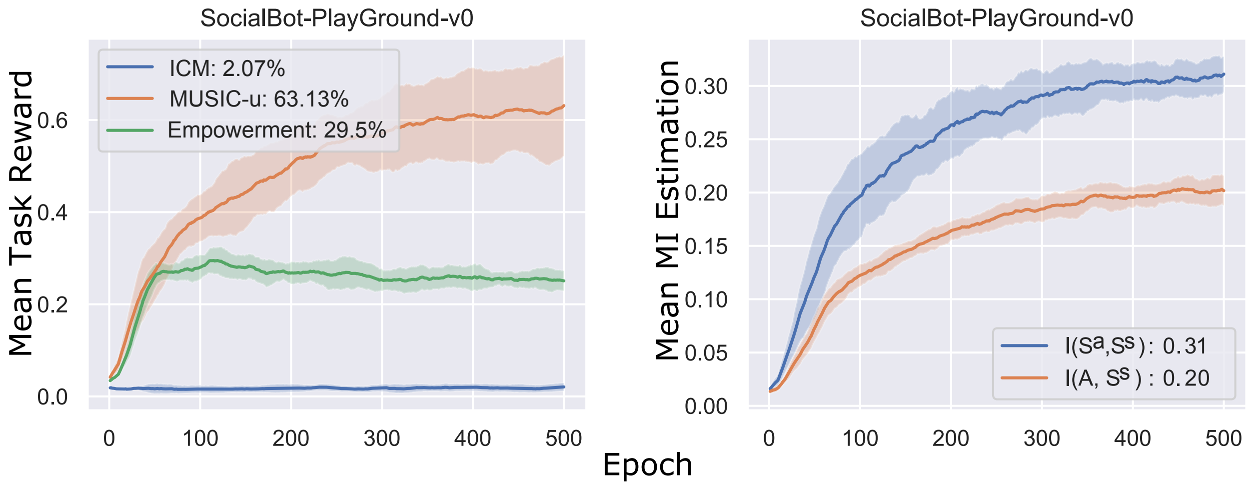
$$\mathcal{F}_{\mathrm{MUSIC+DIAYN}} = I(S^a; S^s) + I(S^s; Z) + \mathcal{H}(A \mid S, Z)$$

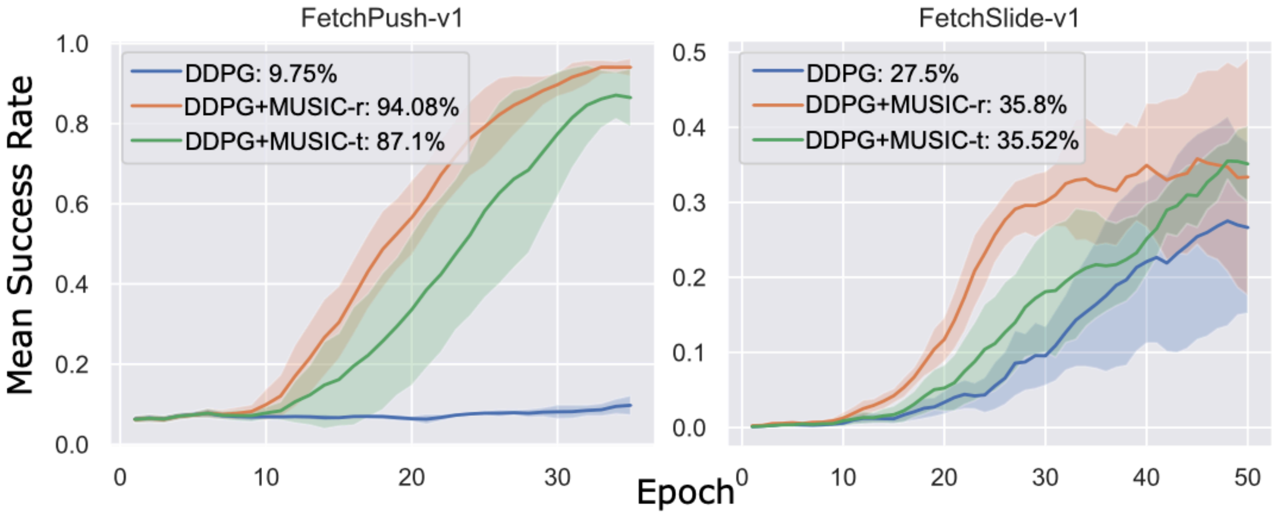


Mutual Information State Intrinsic Control

Compare MUSIC with Empowerment and DISCERN

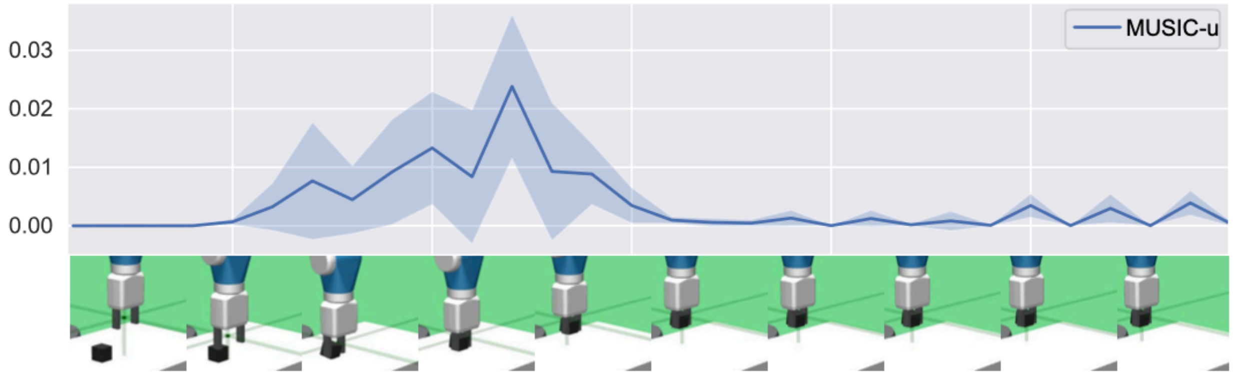
Transfer Learning and reward distribution





Comparison of DISCERN with and without MUSIC

Method	Push (%)	Pick & Place (%)
DISCERN	7.94% ± 0.71%	4.23% ± 0.47%
R (Task Reward)	11.65% ± 1.36%	4.21% ± 0.46%
R+DISCERN	21.15% ± 5.49%	4.28% ± 0.52%
R+DISCERN+MUSIC	95.15% ± 8.13%	48.91% ± 12.67%



Mutual Information State Intrinsic Control
Future Research Directions

- When the existing separation is suboptimal, new methods are needed to divide and select the states automatically.
  - with different combination of state-pairs, the agent can learn different skills

- Using learned skills for hierarchical reinforcement learning
  - Action spaces: learned skill-options
  - Playing billiards, building Lego

Mutual Information estimation prior and post to the training

Mutual Information Objective	Prior-train Value	Post-train Value
$\mathrm{MI}(\mathtt{grip\_pos}; \mathtt{object\_pos})$	0.003 ± 0.017	0.164 ± 0.055
$\mathrm{MI}(\mathtt{grip\_pos}; \mathtt{object\_rot})$	0.017 ± 0.084	0.461 ± 0.088
$\mathrm{MI}(\mathtt{grip\_pos}; \mathtt{object\_velp})$	0.005 ± 0.010	0.157 ± 0.050
$\mathrm{MI}(\mathtt{grip\_pos}; \mathtt{object\_velr})$	0.016 ± 0.083	0.438 ± 0.084

Summary and Take-home Message

- To encourage the agent to control its surroundings help the agent to explore and learn new skills.

- The learned skills or the proposed intrinsic reward help the agent to quickly learn to solve different downstream tasks.

Thank you!

Questions?