Weakly-Supervised RL for Controllable Behavior

Lisa Lee[1,2], Ben Eysenbach[1,2], Ruslan Salakhutdinov[1], Shane Gu[2], Chelsea Finn[2,3]

[1] Carnegie Mellon University [2] Google AI [3] Stanford University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

How can we efficiently learn a diverse array of tasks?

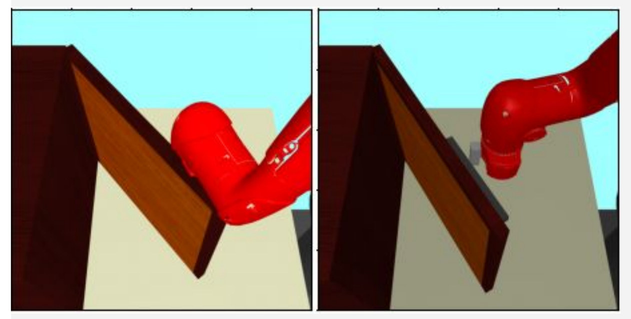
Common approach: (1) Manually design many reward functions. (2) Train RL.

- Meta-RL [e.g., Duan 16, Andrychowicz 16, Mishra 17, Finn 17]
- Multi-task RL [e.g., Rusu 16, Jaderberg 16, Hessel 18]

Problem:

- Designing reward functions is hard
- Prone to **underfitting**

Weak Supervision scales and accelerates RL.



In which image…

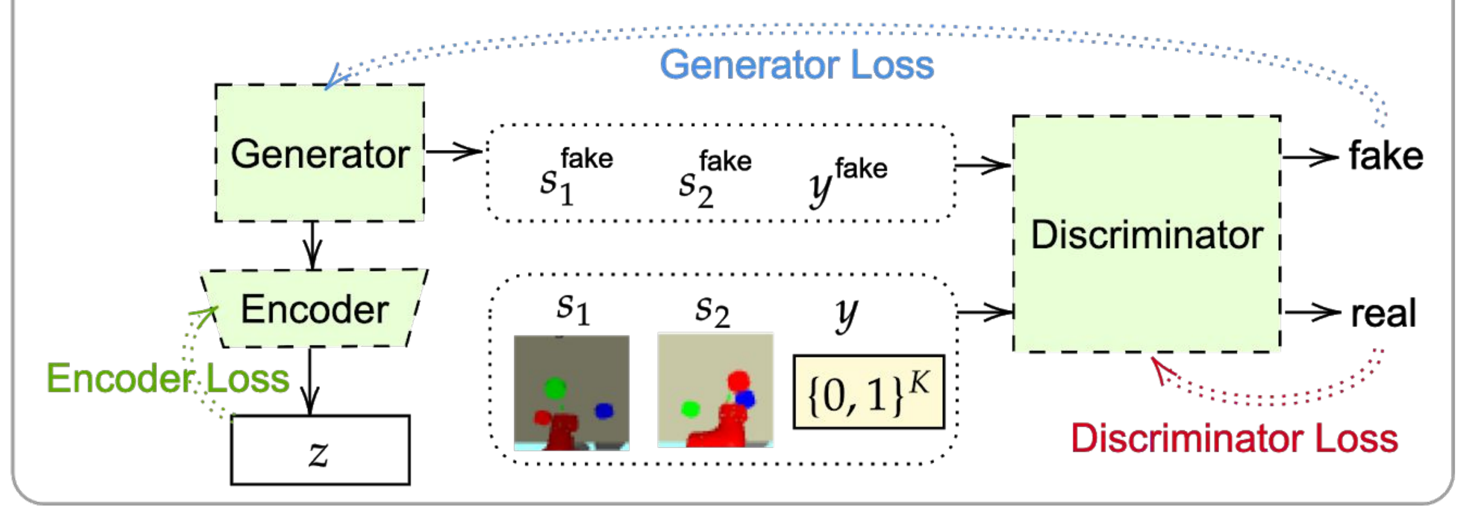
1. ...is the door opened wider?
2. ...is the lighting brighter?
3. ...is the robot closer to the door?

Only use *semantically meaningful* tasks!

Weakly-Supervised Control Framework

Disentanglement model (fixed during phase 2) | Concat | VAE | Trained with RL loss | Next timestep

Phase 1: Weakly-supervised representation learning from agent's observations

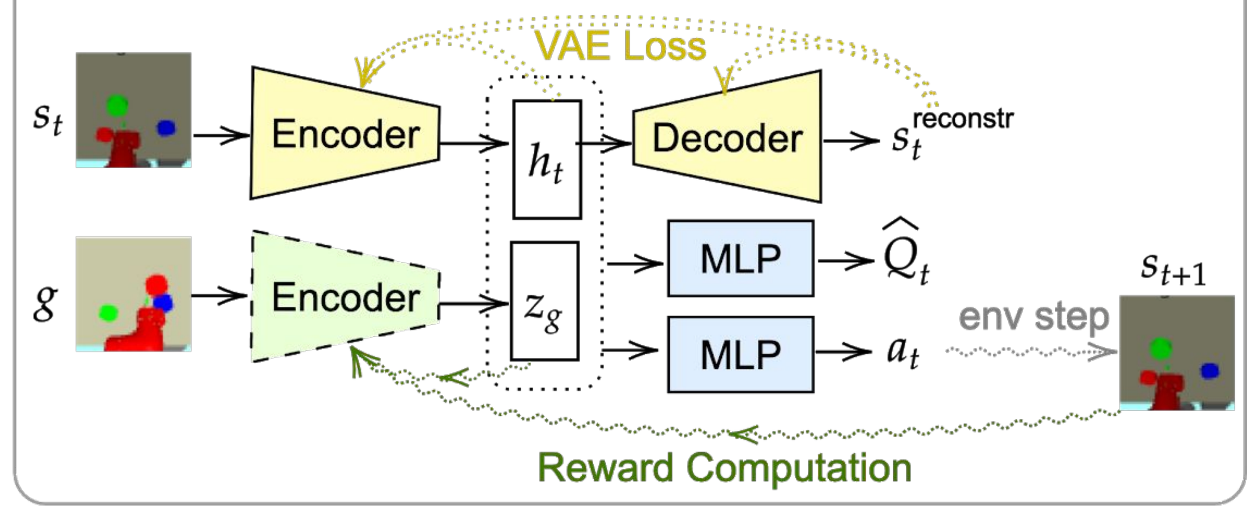


WSC is agnostic to the underlying disentangled representation learning algorithm. In our experiments, we use the method by Shu et al. [1]:

$$\min_D \ \mathbb{E}_{(s_1,s_2,y)\sim\mathcal{D}}\left[D(s_1,s_2,y)\right] + \mathbb{E}_{z_1,z_2\sim\mathcal{N}(0,I)}\left(1 - D(G(z_1),G(z_2),y^{\text{fake}})\right)$$

$$\max_G \ \mathbb{E}_{z_1,z_2\sim\mathcal{N}(0,I)}\left[D(G(z_1),G(z_2),y^{\text{fake}})\right], \qquad \max_e \ \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[e(z\mid G(z))\right]$$

Phase 2: Self-supervised control in disentangled space



In Phase 2, we use the learned disentangled representation to **guide goal generation** and **define distances (rewards)** along semantically meaningful axes:

$$r_t := R_{z_g}(s_{t+1}) := -\|e_{\mathcal{I}}(s_{t+1}) - z_g\|_2^2$$

The agent proposes its own (latent) goals to practice, attempt the proposed goals, and use the experience to update its policy.

Algorithm 1 Weakly-Supervised Control

Input: Weakly-labeled dataset $\mathcal{D}$, factor subindices $\mathcal{I} \subseteq [K]$

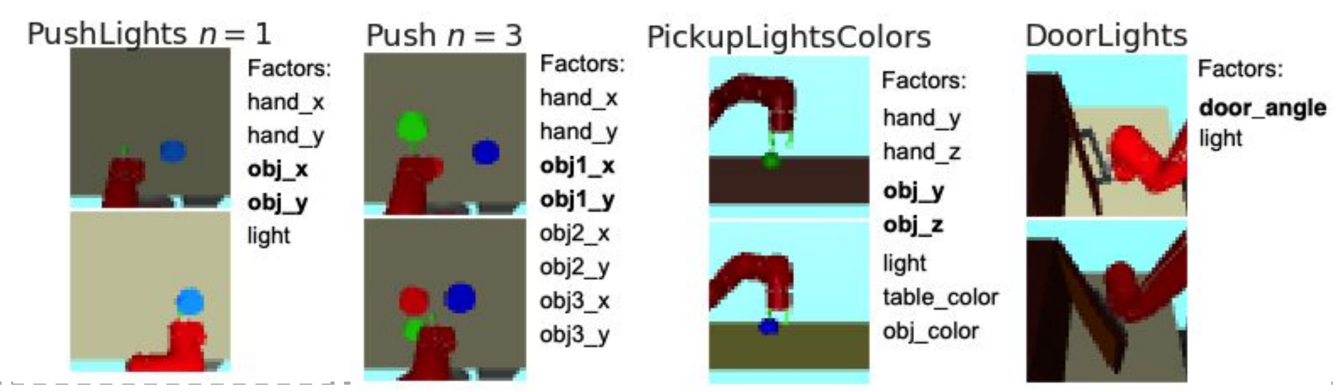
Train disentangled representation $e : \mathcal{S} \mapsto \mathcal{Z}$ using $\mathcal{D}$.
Compute $\mathcal{Z}_{\mathcal{I}}^{\min} = \min_{s\in\mathcal{D}} e_{\mathcal{I}}(s)$.
Compute $\mathcal{Z}_{\mathcal{I}}^{\max} = \max_{s\in\mathcal{D}} e_{\mathcal{I}}(s)$.
Define $p(\mathcal{Z}_{\mathcal{I}}) := \text{Uniform}(\mathcal{Z}_{\mathcal{I}}^{\min}, \mathcal{Z}_{\mathcal{I}}^{\max})$.
Initialize replay buffer $\mathcal{R} \leftarrow \emptyset$.
for iteration$= 0, 1, \ldots,$ **do**
  Sample a goal $z_g \in \mathcal{Z}$ and an initial state s_0.
  for $t = 0, 1, \ldots, H-1$ **do**
    Get action $a_t \sim \pi(s_t, z_g)$.
    Execute action and observe $s_{t+1} \sim p(\cdot \mid s_t, a_t)$.
    Store (s_t, a_t, s_{t+1}, z_g) into replay buffer $\mathcal{R}$.
  for $t = 0, 1, \ldots, H-1$ **do**
    for $j = 0, 1, \ldots, J$ **do**
      With probability p, sample $z_g' \sim p(\mathcal{Z}_{\mathcal{I}})$. Otherwise, sample a future state $s' \in \tau_{>t}$ in the current trajectory and compute $z_g' = e_{\mathcal{I}}(s')$.
      Store $(s_t, a_t, s_{t+1}, z_g')$ into $\mathcal{R}$.
  for $k = 0, 1, \ldots, N-1$ **do**
    Sample $(s, a, s', z_g) \sim \mathcal{R}$.
    Compute $r = R_{z_g}(s') = -\|e_{\mathcal{I}}(s') - z_g\|_2^2$.
    Update actor and critic using (s, a, s', z_g, r).
return $\pi(a \mid s, z)$

Experiments

Environments

12 visual manipulation tasks:

- **Push** & **Pickup** tasks: Move object to goal XY
- **Door** task: Move door to goal angle
- Randomized colors & lighting for increased difficulty.

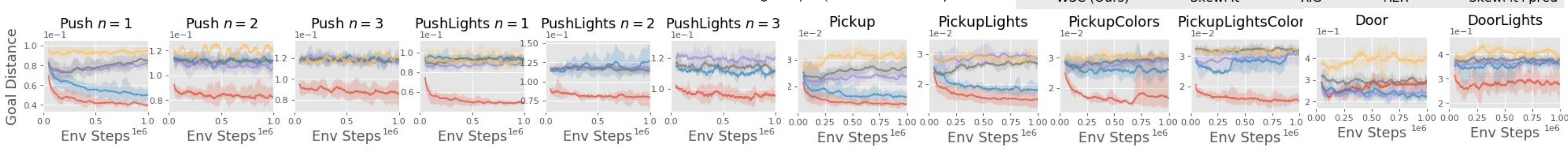


Comparisons

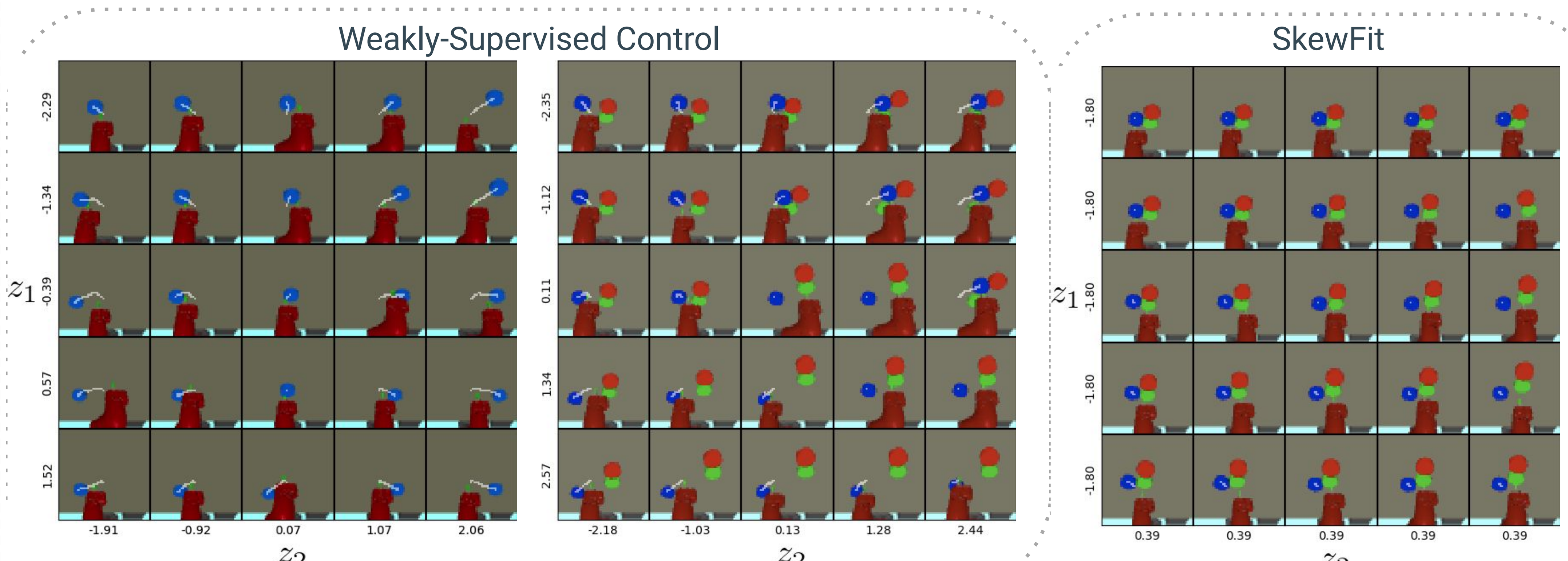
Method	$p(\mathcal{Z})$	$R_{z_g}(s')$
RIG [2]	$\mathcal{N}(0, I)$	$-\|e^{\text{VAE}}(s') - z_g\|_2^2$
SkewFit [3]	$p^{\text{skew}}(\mathcal{R})$	$-\|e^{\text{VAE}}(s') - z_g\|_2^2$
WSC	$\text{Uniform}(\mathcal{Z}_{\mathcal{I}}^{\min}, \mathcal{Z}_{\mathcal{I}}^{\max})$	$-\|e_{\mathcal{I}}(s') - z_g\|_2^2$

Does weakly-supervised control help guide exploration & learning?

Goal Distance vs. Training Steps (Lower is better)

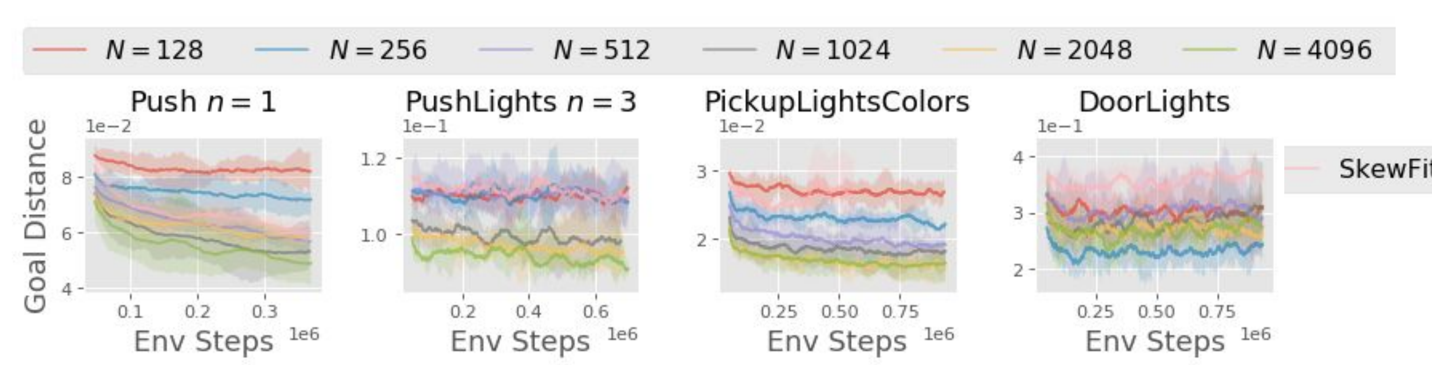


Is the policy's latent space **interpretable**?

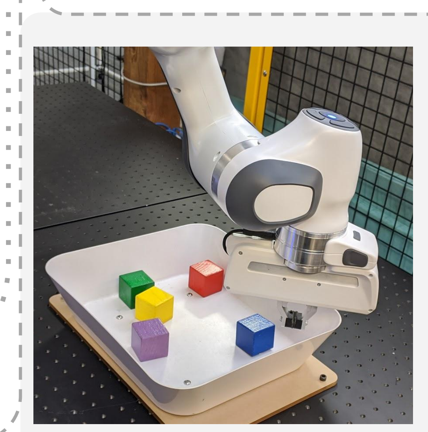


WSC: **Latent goal values** directly align with the direction in which the goal-conditioned policy moves the blue object.

How much weak supervision is needed?



(Answer: About 1000 weak labels for good performance on all domains.)



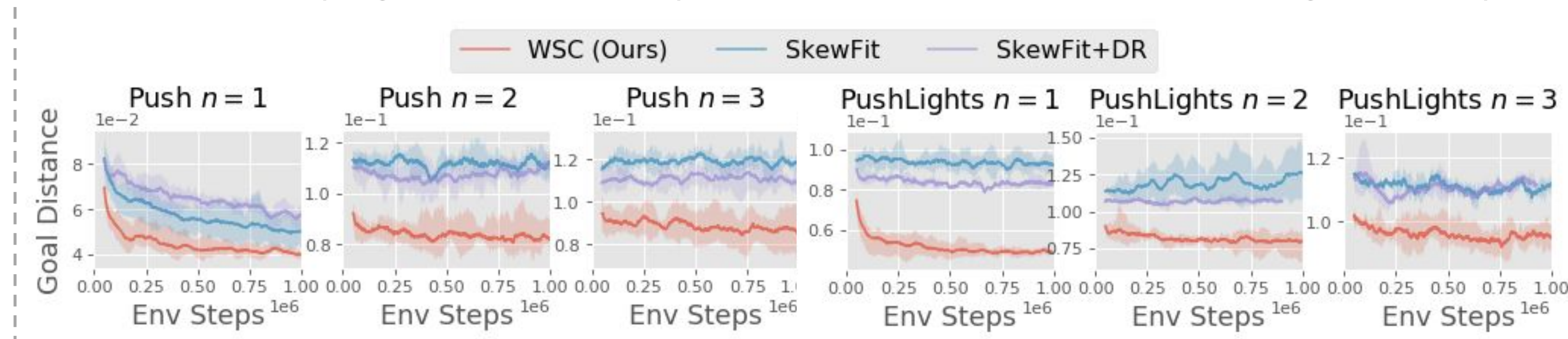
Check out our paper for additional experiments on **noisy & real-world datasets**:

Paper: arxiv.org/abs/2004.02860

Code: https://github.com/google-research/weakly_supervised_control

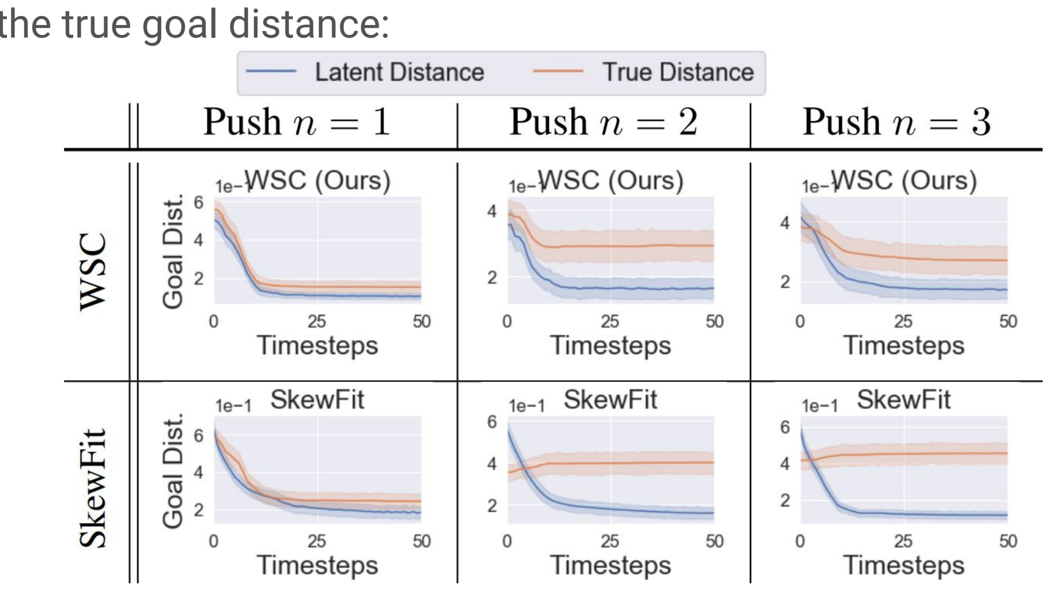
Ablation: What is the role of distances vs. goals?

SkewFit+DR = Sample goals in VAE latent space, but use reward distances in disentangled latent space.



⇒ Disentangled distance metric can help SkewFit slightly in harder environments, but the goal generation mechanism of WSC is crucial to achieving efficient exploration.

Next, we roll out trained policies conditioned on a goal image, and measure the latent distance vs. the true goal distance:



⇒ The disentangled distance optimized by WSC is **more indicative of the true goal distance** than the latent VAE distance optimized by SkewFit, especially for more complex tasks (n > 1).

References

[1] Shu et al., 2019, "Weakly-supervised disentanglement with guarantees".
[2] Nair et al., 2018, "Visual reinforcement learning with imagined goals".
[3] Pong et al., 2019, Skew-fit: State-covering self-supervised reinforcement learning.