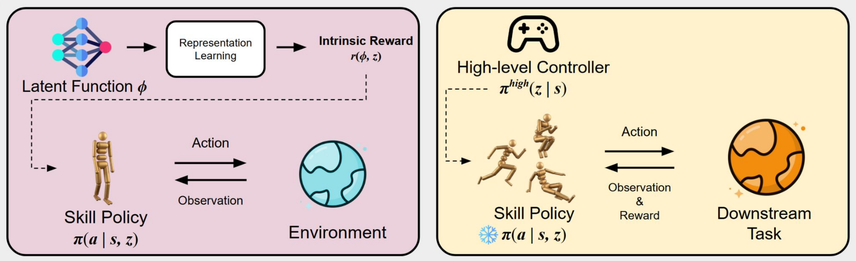
Learning Generalizable Skill Policy with Data-Efficient Unsupervised RL

Jongchan Park Seungjun Oh Seungho Baek Yusung Kim
Sungkyunkwan University

Visit our project page!

ICML International Conference On Machine Learning

Introduction & Motivation



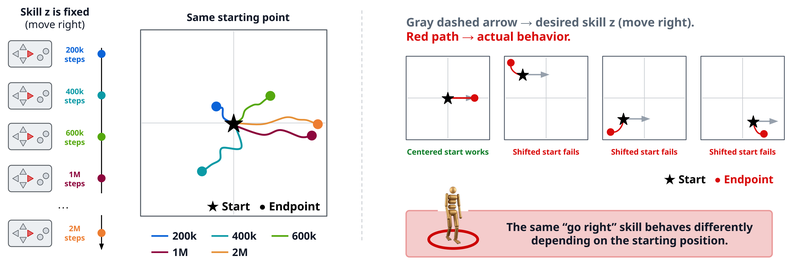
Phase 1: Unsupervised Skill Discovery with Reinforcement Learning

Phase 2: Downstream Task Application with Pre-trained Skill Policy

Fig. 1. Two Phases of URL

Unsupervised Reinforcement Learning (URL)

- ✔ URL learns a skill policy $\pi(a \mid s, z)$ and a latent mapping function $\phi(s)$ without external **supervision signals**.
- ✔ An intrinsic reward function $r(\phi, z)$ guides the reinforcement learning process.
- ✔ URL aims to learn a **scalable** skill policy that can be applied to various downstream tasks.



(a) Semantic Drift in Skill Pre-training

(b) Generalization Failure in Downstream Phase

Fig. 2. Prior SOTA Failure Cases

Problem

- ✔ **Semantic Drift** The transition (s, a, s', z) can become stale because the skill policy and latent function evolve iteratively. However, this stale z is used as a target for ϕ, leading to inefficient learning.
- ✔ **Generalization Failure** The skill policy fails to reproduce the same trajectories when specific observation factors (coordinates, background color, …) are varied. These factors are closely related to $r(\phi, z)$, so the skill policy can easily overfit to them.

Method

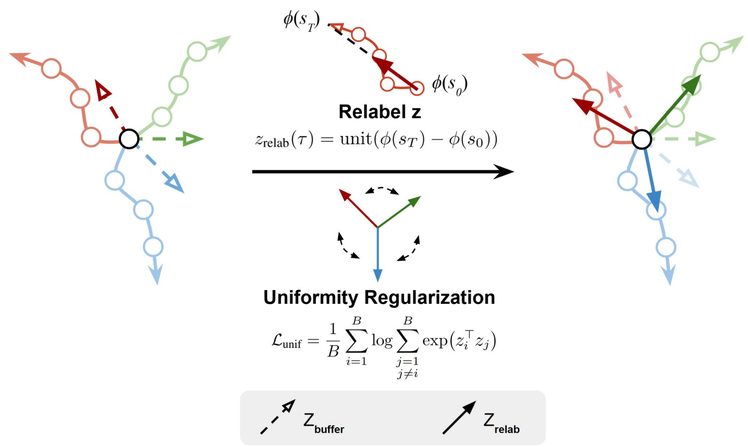


Fig. 3. Skill Relabeling with Uniformity Regularization.

Skill Relabeling

- ✔ We **relabel** stale z in the replay buffer to z_{relab} using the **current latent** function $\phi(s)$.
- ✔ To maintain high entropy in the z_{relab} distribution, we adopt a **uniformity term** that improves the lower bound on $H(z_{relab})$.

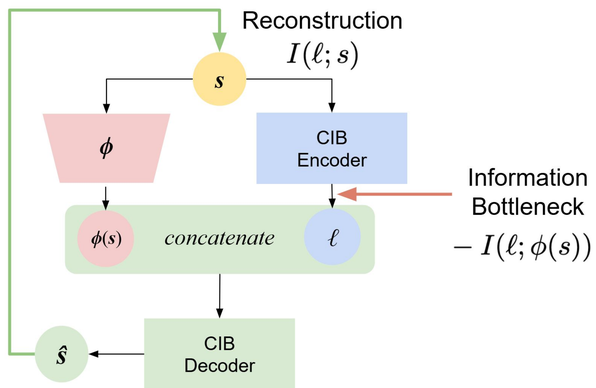
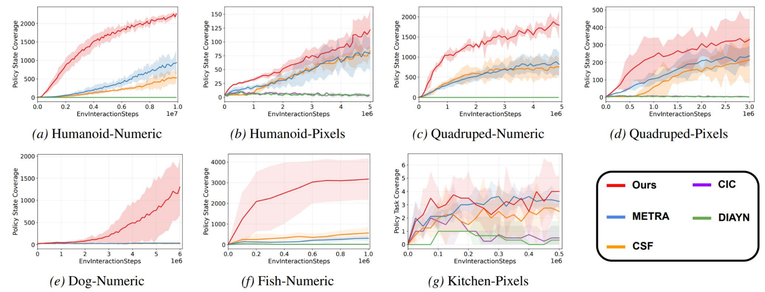


Fig. 4. Architecture of CIB

Complementary Information Bottleneck (CIB)

- ✔ We propose a **structural solution** that prevents the skill policy from overfitting to specific observations already captured by $\phi(s)$.
- ✔ The output ℓ of the encoder $q_\psi(\ell \mid s)$ **replaces** the original observation in the skill policy, yielding $\pi(a \mid \ell, z)$.

Results & Analysis



(a) Humanoid-Numeric *(b)* Humanoid-Pixels *(c)* Quadruped-Numeric *(d)* Quadruped-Pixels

(e) Dog-Numeric *(f)* Fish-Numeric *(g)* Kitchen-Pixels

Fig. 5. Skill Pre-training Results

Type	Domain	Task	METRA	**Ours**
State-based	Humanoid	(FS, RG)	0.52 ± 0.12	$\mathbf{0.83 \pm 0.08}$
		(RS, RG)	0.04 ± 0.03	$\mathbf{0.68 \pm 0.13}$
		MazeE	0.19 ± 0.27	$\mathbf{0.78 \pm 0.04}$
		MazeH	0.00 ± 0.00	$\mathbf{0.41 \pm 0.20}$
	Quadruped	(FS, RG)	$\mathbf{0.81 \pm 0.05}$	0.78 ± 0.08
		(RS, RG)	0.04 ± 0.02	$\mathbf{0.35 \pm 0.03}$
		MazeE	0.02 ± 0.03	$\mathbf{0.33 \pm 0.33}$
		MazeH	0.00 ± 0.00	$\mathbf{0.13 \pm 0.16}$
	Dog	(FS, RG)	0.01 ± 0.01	$\mathbf{0.55 \pm 0.10}$
		(RS, RG)	0.00 ± 0.00	$\mathbf{0.53 \pm 0.08}$
		MazeE	0.00 ± 0.00	$\mathbf{0.56 \pm 0.07}$
		MazeH	0.00 ± 0.00	$\mathbf{0.35 \pm 0.15}$
	Fish	(FS, RG)	0.00 ± 0.00	$\mathbf{0.79 \pm 0.06}$
Pixel-based	Humanoid	(FS, RG)	0.35 ± 0.04	$\mathbf{0.48 \pm 0.15}$
		(RS, RG)	0.16 ± 0.07	$\mathbf{0.30 \pm 0.10}$
	Quadruped	(FS, RG)	0.60 ± 0.09	$\mathbf{0.64 \pm 0.07}$
		(RS, RG)	0.30 ± 0.11	$\mathbf{0.40 \pm 0.11}$

Tab. 1. Downstream Task Results

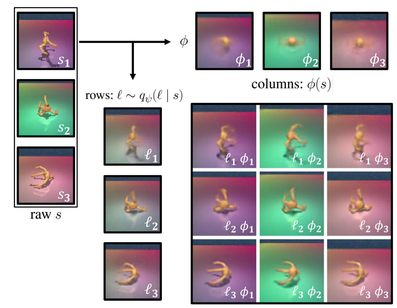


Fig. 6. Remaining Information of CIB

Key Results

- ✔ Our algorithm (GenDa) achieves **data-efficient** learning through skill relabeling during the **skill pre-training phase**.
- ✔ GenDa also scales to challenging tasks where previous approaches fail.
- ✔ CIB **mitigates overfitting** and thereby enhances **downstream-task performance**.

Conclusion

Summary

- ✔ We **identify two fundamental challenges** in off-policy unsupervised reinforcement learning that limit the scalability of skill discovery.
- ✔ GenDa significantly improves **data efficiency** and **generalizability**.