Variational Curriculum Reinforcement Learning for Unsupervised Discovery of Skills

Seongun Kim [*1] **Kyowoon Lee** [*2] **Jaesik Choi** [1]

Abstract

Mutual information-based reinforcement learning (RL) has been proposed as a promising framework for retrieving complex skills autonomously without a task-oriented reward function through mutual information (MI) maximization or variational empowerment. However, learning complex skills is still challenging, due to the fact that the order of training skills can largely affect sample efficiency. Inspired by this, we recast variational empowerment as curriculum learning in goal-conditioned RL with an intrinsic reward function, which we name Variational Curriculum RL (VCRL). From this perspective, we propose a novel approach to unsupervised skill discovery based on information theory, called Value Uncertainty Variational Curriculum (VUVC). We prove that, under regularity conditions, VUVC accelerates the increase of entropy in the visited states compared to the uniform curriculum. We validate the effectiveness of our approach on complex navigation and robotic manipulation tasks in terms of sample efficiency and state coverage speed. We also demonstrate that the skills discovered by our method successfully complete a real-world robot navigation task in a zero-shot setup and that incorporating these skills with a global planner further increases the performance.

1. Introduction

Intelligent creatures are able to efficiently explore the environments and learn useful skills in the absence of external supervision. By utilizing these skills, they can quickly accomplish tasks when they are later faced with specific tasks. To scale a learning agent to the real-world, it is crucial to achieve such ability of learning skills without supervision. Recent studies on unsupervised RL suggest ways to alleviate the need for human effort. Most of these approaches focus on reducing the burden of designing objective functions by incorporating intrinsic motivation objectives or leveraging concepts from information theory. In this work, we further reconcile with the need not only to manually engineer objective functions but to craft the order of training skills.

Empowerment or MI-based RL (Klyubin et al., 2005; Salge et al., 2014) has gained traction in recent years as a means of unsupervised skill discovery due to its intuitive interpretation and empirical successes (Eysenbach et al., 2019; Sharma et al., 2019; Jabri et al., 2019). However, the common empowerment approach has been to either fix or parameterize the distribution of skills (Nair et al., 2018; Pong et al., 2020; Campos et al., 2020). The efficiency of learning skills with respect to the number of required training samples is rather limited when the agent learns complex skills from a fixed skill distribution without an organized order. The notion of *curriculum* studies the effectiveness of the order of training skills. By selecting the order of appropriate skills, a learning agent may achieve a variety of complex skills (Florensa et al., 2018; Fang et al., 2019). However, it is both necessary to define a set of tasks that can be used to generate curriculum (Klink et al., 2020; Zhang et al., 2020) and specify a form of reward functions (Racaniere et al., 2019; Ren et al., 2019; Narvekar & Stone, 2019).

To rectify this issue, we interpret empowerment as a unifying framework for curriculum learning in goal-conditioned RL (GCRL). Recasting variational empowerment as curriculum learning in GCRL with intrinsic reward function, interestingly our Variational Curriculum RL (VCRL) framework encapsulates most of the prior MI-based approaches (Nair et al., 2018; Pong et al., 2020; Campos et al., 2020). In this regard, we derive a new approach to information-theoretic skill discovery, Value Uncertainty Variational Curriculum (VUVC) that allows us to automatically generate curriculum goals which maximize the expected information approximated as the uncertainty in predictions of an ensemble of value functions. We analyze asymptotic behavior of the

[*]Equal contribution [1]Kim Jaechul Graduate School of AI, KAIST [2]Department of Computer Science and Engineering, UNIST. Correspondence to: Jaesik Choi <jaesik.choi@kaist.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Codes are available at `github.com/seongun-kim/vcrl`.

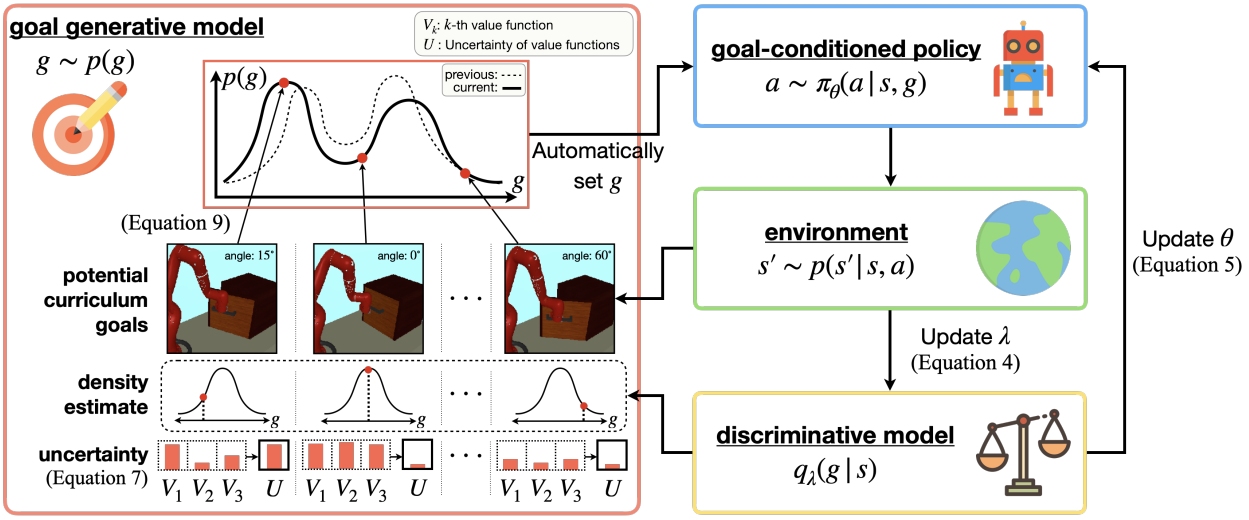

Figure 1. An overview of our proposed method, VUVC, under the unifying framework for curriculum learning in goal-conditioned RL. The value uncertainty proposes informative goals which would generate a stronger learning signal. The density estimate of potential curriculum goals indicates the novelty of the goal to the learning agent. The density estimate model is derived from the discriminative model, which is trained alongside the agent. This discriminative model provides intrinsic rewards to the agent. VUVC combines these two measures to construct a goal generative model, promoting unsupervised exploration of the entire state space by the agent.

entropy of visited states and provide the reasons why our method results in much faster coverage of the state space compared to existing methods.

The main contributions of this paper can be summarized as follows: (1) We provide the unifying framework VCRL encapsulating most of the prior MI-based approaches. (2) We propose VUVC, a value uncertainty based approach to information-theoretic skill discovery, aimed at automatically generating curricula for training skills and which is supported by theoretical justification. (3) We show the effectiveness of our approach on complex navigation, robotic manipulation in both configuration and image state space, and real-world robotic navigation tasks and illustrate that the skills discovered by our method can be further improved by incorporating them with a global planner.

2. Background

2.1. Goal-Conditioned Reinforcement Learning

Goal-conditioned RL (Kaelbling, 1993) extends the standard RL framework to enable agents to accomplish a variety of tasks. It solves the problem formulated as a goal-conditioned Markov decision process (MDP) which is defined as a tuple $\langle \mathcal{S}, \mathcal{G}, \mathcal{A}, P, R_g, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{G}$ is the set of goals, $\mathcal{A}$ is the set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, +\infty)$ is the transition probability, $R_g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the goal-conditioned reward function and $\gamma \in [0, 1]$ is the discount factor. The objective of GCRL is to find the policy $\pi_\theta(a|s, g)$ parameterized with θ where $s \in \mathcal{S}, a \in \mathcal{A}, g \in \mathcal{G}$ and $\pi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, +\infty)$ that maximizes the universal value function (Schaul et al., 2015):

$$\theta \leftarrow \arg\max_\theta V^{\pi_\theta}(s, g) \triangleq \mathbb{E}_{\substack{a_t \sim \pi_\theta(a_t|s_t,g), \\ s_{t+1} \sim P(s_{t+1}|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R_g(s_t, a_t) \middle| s_0 = s \right]. \quad (1)$$

2.2. Mutual Information and Empowerment

In the context of RL, MI maximization such as empowerment generally means maximizing the mutual information between a function of states and a function of actions to learn latent-conditioned policies $\pi(a|s, z)$ where the latent code z can be interpreted as a macro-action, skill or goal (Eysenbach et al., 2019; Sharma et al., 2019). Empowerment maximizes the following MI objective:

$$\begin{aligned} \mathcal{I}(s; z) &= \mathcal{H}(s) - \mathcal{H}(s|z) \\ &= \mathcal{H}(z) - \mathcal{H}(z|s) \\ &= \mathbb{E}_{z \sim p(z), s \sim p(s|z)} \left[\log p(z|s) - \log p(z) \right] \\ &\geq \mathbb{E}_{z \sim p(z), s \sim p(s|z)} \left[\log q_\lambda(z|s) - \log p(z) \right], \end{aligned} \quad (2)$$

where $\mathcal{H}(\cdot)$ is the Shannon entropy, $p(z)$ is the prior distribution, and $q_\lambda(z|s)$ represents the variational approximation for intractable posterior $p(z|s)$ parameterized with λ, often called a discriminator (Eysenbach et al., 2019; Sharma et al., 2019; Campos et al., 2020). This objective provides a way to train a policy that guides agents to explore diverse states by maximizing $\mathcal{H}(s)$ and makes the state s distinguishable from the latent code z by minimizing $\mathcal{H}(s|z)$.

Methods	$q_\lambda(g\|s)$	$p(g)$	Non-stationary goal distribution
GCRL (w/ sparse reward)	$\frac{1}{Z}\exp(1-2\delta_g\mathcal{U}_{[s\pm\delta_g]})$	$p^{\text{target}}(g)$	✗
GCRL (w/ dense reward)	$\mathcal{N}(s,\sigma^2 I)$	$p^{\text{target}}(g)$	✗
EDL (Campos et al., 2020)	$\mathcal{N}(\mu(s),\sigma^2 I)$	$p^{\text{explored}}(g)$	✗
RIG (Nair et al., 2018)	$\mathcal{N}(\mu(s),\sigma^2 I)$	$p_t^{\text{visited}}(g)$	✓
Skew-Fit (Pong et al., 2020)	$\mathcal{N}(\mu(s),\sigma^2 I)$	$\propto p_t^{\text{visited}}(g)^\alpha$	✓
VUVC (**ours**)	$\mathcal{N}(\mu(s),\sigma^2 I)$	$\propto U(g)p_t^{\text{visited}}(g)^\alpha$	✓

Table 1. Variants of VCRL framework which encapsulate most of the prior MI-based methods, depending on the choice of a discriminator $q_\lambda(g|s)$, a goal generative model $p(g)$, and whether $p(g)$ is stationary or not, where both $q_\lambda(g|s)$ and $p(g)$ are components of the MI objective. The discriminator determines the shape of goal-conditioned reward functions including sparse and dense shapes.

3. Variational Curriculum Reinforcement Learning

To recast the aforementioned MI-based RL as VCRL, we first present that general GCRL methods optimize empowerment objective by formulating a discriminator to represent commonly used goal-conditioned reward functions. We then expand this setting to a curriculum learning framework with a goal generative model, which we name VCRL where Table 1 summarizes variants of the VCRL framework.

Henceforth, we consider the latent code z in Equation 2 as a goal g and assume the goal space matches the state space, while VCRL framework is not limited to this assumption and trivially extended by introducing a state abstraction function (Ren et al., 2019). The objective now becomes equivalent to that of a GCRL where the resulting policy aims to reach g (Pong et al., 2020; Choi et al., 2021). Given a policy $\pi_\theta(a|s,g)$ and a discriminator $q_\lambda(g|s)$, an objective of MI-based RL is to maximize a variational lower bound:

$$\mathcal{F}(\theta,\lambda)=\mathbb{E}_{\substack{g\sim p(g),\\ s\sim\rho^\pi(s|g)}}[\log q_\lambda(g|s)-\log p(g)], \quad (3)$$

where $\rho^\pi(s|g)$ is a stationary state distribution induced by the goal-conditioned policy $\pi(a|s,g)$ (Gregor et al., 2016; Campos et al., 2020). To solve this joint optimization problem, we iteratively fix one parameter and optimize the other one at each training epoch i:

$$\lambda^{(i)}\leftarrow\arg\max_\lambda \mathbb{E}_{\substack{g\sim p(g),\\ s\sim\rho^{\pi_{\theta^{(i-1)}}}(s|g)}}[\log q_\lambda(g|s)-\log p(g)] \quad (4)$$

$$\theta^{(i)}\leftarrow\arg\max_\theta \mathbb{E}_{\substack{g\sim p(g),\\ s\sim\rho^{\pi_\theta}(s|g)}}[\log q_{\lambda^{(i)}}(g|s)]. \quad (5)$$

As described in the prior work (Warde-Farley et al., 2019; Choi et al., 2021), it has been shown that Equation 5 which is also called an intrinsic reward (Gregor et al., 2016), recovers the objective of GCRL in Equation 1 with dense rewards. By choosing a Gaussian distribution with mean s and fixed variance $\sigma^2 I$ for $q_\lambda(g|s)$ where I is the identity matrix, this objective becomes a negative l_2 distance between s and g. Similarly, one can show that the intrinsic reward represented in Equation 5 becomes a sparse reward where an agent gets 0 reward if l_2 distance between s and g is within some threshold δ_g and gets -1 otherwise. Other MI-based methods can also be considered a GCRL by modeling $q_\lambda(g|s)$ to follow $\mathcal{N}(\mu(s),\sigma^2 I)$ where $\mu(s)$ is a function approximator usually following an encoder structure.

We further expand the interpretation of MI-based methods as a framework of GCRL to a framework of curriculum learning, which we term VCRL. Curriculum learning in RL studies the order of training skills or tasks. In the context of GCRL, the order of tasks, *curriculum*, is determined by characterizing a goal distribution $p(g)$ (Fournier et al., 2018; Florensa et al., 2018; Racaniere et al., 2019; Ren et al., 2019; Zhang et al., 2020; Klink et al., 2020). Without an explicit design of $p(g)$, VCRL is reduced to a simple GCRL where a target goal is given from the environment, $p^{\text{target}}(g)$. Otherwise, one can design a goal generative model to satisfy various purposes of the training. For instance, EDL (Campos et al., 2020), a variant of MI-based RL, aims to train a state space covering skill. EDL first learns $p^{\text{explored}}(g)$ along with an exploration policy (Lee et al., 2019) which tries to cover the entire state space. Then, it optimizes the MI objective (Equation 3) with the stationary goal distribution $p^{\text{explored}}(g)$. Skew-Fit (Pong et al., 2020) also seeks to learn a state space covering skill in an unsupervised manner. However, unlike EDL, it assumes a non-stationary goal distribution to ensure that the state density $p(s)$ converges to uniform distribution. This is achieved by formulating the goal distribution, $p(g)$, to be proportional to the approximate state density, $p^{\text{visited}}(s)$, raised to a skewing parameter α within the range of $[-1,0)$. Similarly, RIG samples goals directly from $p^{\text{visited}}(s)$.

4. Value Uncertainty Variational Curriculum

Despite the many empirical successes of empowerment methods, learning complex skills is still challenging since there has been little consideration of $p(g)$ in the MI objec-

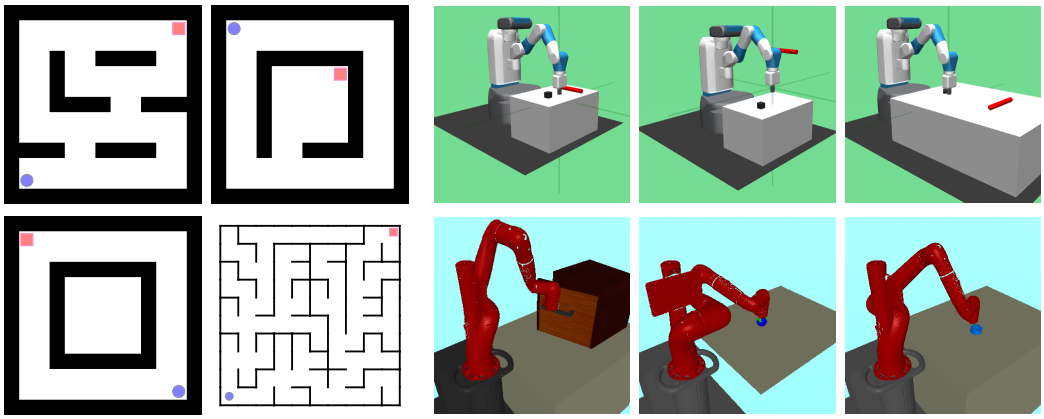

Figure 2. Illustrations of simulated environments. (Left) Point maze navigation tasks which we name *PointMazeA*, *B*, *C*, and *SquareLarge* in sequential order. The initial state and goal distribution of each task are depicted by a blue circle and red box, respectively. (Top right) Configuration-based robot manipulation tasks: *FetchPush*, *FetchPickAndPlace* and *FetchSlide*. The goal distribution which represents the target position for the puck, is illustrated by a red cylinder. (Bottom right) Vision-based robot manipulation tasks: *SawyerDoorHook*, *SawyerPickup* and *SawyerPush*.

tive (Achiam et al., 2018; Eysenbach et al., 2019; Warde-Farley et al., 2019; Campos et al., 2020). To efficiently learn complex skills, it is important to effectively optimize the variational empowerment in Equation 3. To this end, the agent should seek out goals from which it can learn the most. This can be formalized in the uncertainty of value functions which track the performance of the policy. To estimate the uncertainty, we use an ensemble of multiple value functions that has been widely adopted in the literature with empirical success (Osband et al., 2016; Lakshminarayanan et al., 2017; Osband et al., 2018; Zhang et al., 2020). Formally, we maintain an ensemble of parameters for value functions: $\psi = \{\psi_1, ..., \psi_K\}$, which is randomly initialized independently,

$$\text{Value functions } v_\psi : s, g \rightarrow V_\psi(s, g). \tag{6}$$

We quantify the uncertainty of value functions in predictions of the ensemble members from the initial state by computing the variance over the ensemble of the value functions:

$$\text{Uncertainty } U(g) : \text{Var}[\{V_\psi(s_0, g) | \psi \in \{\psi_1, ..., \psi_K\}]. \tag{7}$$

Proposition 1. *If $V_\psi(s_0, g)$ follows a log-concave distribution, then we have*

$$\mathcal{I}(V_\psi(s_0, g); \psi | s_0, g) \geq \log(2\sqrt{\text{Var}[V_\psi(s_0, g)]}). \tag{8}$$

Proof Sketch. We rewrite the mutual information as the difference between conditional entropy and marginal entropy. We then use the result in (Marsiglietti & Kostina, 2018) on a lower bound on the entropy of a log-concave random variable, expressed in terms of the p-th absolute moment to obtain the conclusion. The complete proof appears in Appendix C. □

It follows from Proposition 1 that finding a goal which maximizes the mutual information can be relaxed into the surrogate problem, which is to select a goal that maximizes the uncertainty in predictions of an ensemble of value functions when we take $K \rightarrow \infty$. With this intuition, one natural option to sample goals is to compute a goal probability proportional to the uncertainty $p(g) \propto U(g)$, where $g \in \text{support}(p_t^{\text{visited}})$. To prevent goals with lower density from being frequently proposed, we adopt the Skew strategy (Pong et al., 2020) which assigns more weight to rare samples by skewing the goal sampling probability. We therefore sample goals from the following distribution:

$$p_t^{\text{VUVC}}(g) = \frac{1}{Z_{t,\alpha}} U(g) p_t^{\text{visited}}(g)^\alpha, \quad \alpha \in [-1, 0), \tag{9}$$

where $Z_{t,\alpha}$ is the normalizing coefficient. We approximate p_t^{visited} by training a generative model on samples in the replay buffer, where we use a β-VAE (Higgins et al., 2017) in our experiments. We term a VCRL method with a goal generative model following Equation 9 as VUVC.

Definition 1. (Expected Entropy Increment over Uniform Curriculum). Given the empirical distribution of the visited state

$$p_t^{\text{visited}}(s) = \sum_{i=1}^{t} \frac{\mathbb{I}(s_i = s)}{t}, \tag{10}$$

where $\mathbb{I}(\cdot)$ is an indicator function, uniform curriculum goal distribution $p_t^{\mathcal{U}}$ and value uncertainty-based curriculum goal distribution p_t^{VU} are defined as follows:

$$p_t^{\mathcal{U}}(g) = \mathcal{U}(\text{support}(p_t^{\text{visited}}))(g), \tag{11}$$

$$p_t^{\text{VU}}(g) = \frac{1}{Z_t} U(g) p_t^{\mathcal{U}}(g), \tag{12}$$

where Z_t is the normalizing coefficient, $p_t^{\mathcal{U}}$ is uniform over the support of the p_t^{visited} and $U(g)$ is the value uncertainty.

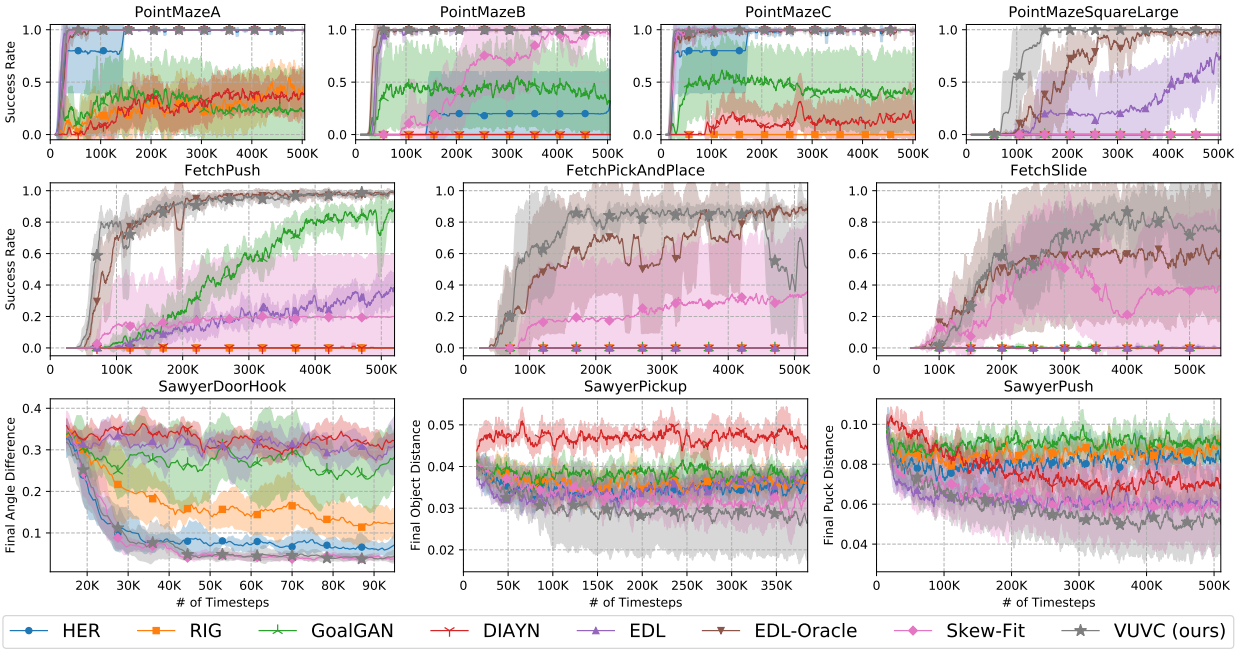

Figure 3. Learning curves for configuration-based point maze navigation tasks (top), continuous robot control tasks (middle), and vision-based continuous robot manipulation tasks (bottom). *Mean (SD)* of each performance measure over 5 random seeds are reported where results are smoothed across 10 training epochs for each seed. VUVC consistently outperforms other VCRL variants for all tasks.

Then the expected entropy increment over uniform curriculum I_t is defined as

$$I_t = \mathbb{E}_{g\sim p_t^{\mathrm{VU}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})] - \mathbb{E}_{g\sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})]. \quad (13)$$

To study the asymptotic behavior of the expected next step entropy induced by VUVC, we define the expected entropy increment over uniform curriculum in Equation 13 for the case of discrete state space. However, computing the empirical distribution of the next visited state $p_{t+1}^{\mathrm{visited}}$ requires marginalizing out the MDP dynamics which is intractable to compute. Therefore, we consider two special cases when (1) an agent always reaches the goal in Proposition 2 and (2) an agent sometimes fails to reach goals but potentially increases the amount of entropy in Proposition 3.

Proposition 2. *Given* $\epsilon = \frac{1}{t}$ *and* $\rho^{\pi_\theta}(s|g) = \mathbb{I}(s = g)$*, if*

$$\mathrm{Cov}[U(g), \log p_t^{\mathrm{visited}}(g)] \leq 0, \quad (14)$$

and take $\epsilon \to 0$*, then we have,*

$$\lim_{\epsilon\to 0} \frac{\partial}{\partial \epsilon} I_t = \lim_{\epsilon\to 0} \frac{\partial}{\partial \epsilon} \left(\mathbb{E}_{g\sim p_t^{\mathrm{VU}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})] - \mathbb{E}_{g\sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})] \right) > 0. \quad (15)$$

Proof Sketch. We begin by deriving a next step empirical distribution of the visited state given a curriculum goal g and a stationary state distribution induced by the policy $\rho^{\pi_\theta}(s|g)$, which can be written as $p_{t+1}^{\mathrm{visited}}(s) = \frac{p_t^{\mathrm{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1+\epsilon}$. Plugging this back into Definition 1, we analyze asymptotic behavior of the expected entropy increment and obtain the conclusion with the assumption $\rho^{\pi_\theta}(s|g) = \mathbb{I}(s = g)$. The complete proof is provided in Appendix C. □

With an accurate goal-conditioned policy and the model of dynamics, Proposition 2 gives us intuition that our VUVC is at least better than the uniform curriculum which Skew-Fit aims to converge to, if the uncertainty of the learned value functions $U(g)$ and the log density of p_t^{visited} are negatively correlated. We expect this negative correlation to happen frequently, since the uncertainty is positive for novel states, but it eventually reduces to zero with a sufficiently large number of samples.

Proposition 3. *Define the set* $\mathcal{G} = \mathcal{G}_{\mathrm{exploit}} \cup \mathcal{G}_{\mathrm{uninfo}} \cup \mathcal{G}_{\mathrm{info}}$ *and positive constant* Δ_1, Δ_2 *where*

$$\rho^{\pi_\theta}(s|g) = \begin{cases} \mathbb{I}(s = g) & \text{for } g \in \mathcal{G}_{\mathrm{exploit}} \\ \rho^{\pi_\theta}_{\mathrm{uninfo}}(s|g) & \text{for } g \in \mathcal{G}_{\mathrm{uninfo}} \\ \rho^{\pi_\theta}_{\mathrm{info}}(s|g) & \text{for } g \in \mathcal{G}_{\mathrm{info}}, \end{cases} \quad (16)$$

for all $g \in \mathcal{G}_{\mathrm{uninfo}}$*,*

$$\mathbb{E}_{s\sim\rho^{\pi_\theta}_{\mathrm{uninfo}}(s|g)}[\log p_t^{\mathrm{visited}}(s)] = \log p_t^{\mathrm{visited}}(g) + \Delta_1,$$

and for all $g \in \mathcal{G}_{\mathrm{info}}$*,*

$$\mathbb{E}_{s\sim\rho^{\pi_\theta}_{\mathrm{info}}(s|g)}[\log p_t^{\mathrm{visited}}(s)] = \log p_t^{\mathrm{visited}}(g) - \Delta_2.$$

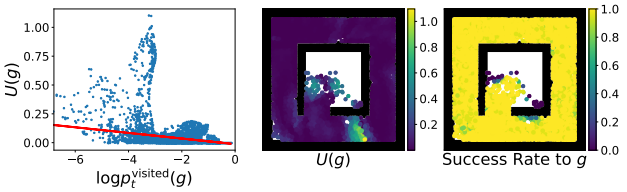

Figure 4. An illustration of the relation between value uncertainty and log density of visited states (left) and the landscape of value uncertainty (middle) and success rate (right).

Given $\epsilon = \frac{1}{t}$, *if*

$$\begin{aligned}\mathrm{Cov}[U(g), \log p_t^{\text{visited}}(g)] &\leq 0,\\ \mathbb{E}_{g\in\mathcal{G}_{\text{uninfo}}}[p_t^{\text{VU}}(g)] &\leq \mathbb{E}_{g\in\mathcal{G}_{\text{uninfo}}}[p_t^{\mathcal{U}}(g)],\\ \mathbb{E}_{g\in\mathcal{G}_{\text{info}}}[p_t^{\text{VU}}(g)] &\geq \mathbb{E}_{g\in\mathcal{G}_{\text{info}}}[p_t^{\mathcal{U}}(g)],\end{aligned}$$

and take $\epsilon \to 0$, *then we have,*

$$\lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon} I_t = \lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon}\left(\mathbb{E}_{g\sim p_t^{\text{VU}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] - \mathbb{E}_{g\sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\text{visited}})]\right) > 0.$$

Proof Sketch. The proof proceeds in a similar manner as Proposition 2 except for an assumption $\mathcal{G} = \mathcal{G}_{\text{exploit}} \cup \mathcal{G}_{\text{uninfo}} \cup \mathcal{G}_{\text{info}}$. The complete proof is in Appendix C. □

Proposition 3 extends Proposition 2 to the case where the goal-conditioned policy is sub-optimal and fails to achieve some of the goals. It implies that we need a curriculum method which can filter out uninformative states when the policy can not consistently achieve certain states, in order to achieve a rapid increment of entropy. Empirical observations indicate that VUVC achieves this effect (further details provided in Section 5).

5. Experiments

5.1. Experimental Setup and Baselines

We validate the effectiveness of VUVC on 10 different environments. They consist of point maze navigation tasks (Zhang et al., 2020; Trott et al., 2019), configuration-based robot control tasks (Plappert et al., 2018), and vision-based robot manipulation tasks (Nair et al., 2018) which are shown in Figure 2. Especially, for configuration-based robot tasks, we modify the initial state and goal distribution following from the prior work (Ren et al., 2019) to consider more complicated tasks which require extensive exploration. Further details of experimental setups are presented in Appendix D.

By comparing VUVC to HER (Andrychowicz et al., 2017), we study how effectively explicit curriculum improves sample efficiency over implicit curriculum. We examine how well value uncertainty curriculum goals encourages exploration over goals from GoalGAN (Florensa et al., 2018) which generates goals by measuring task difficulty through success rate, over goals from DIAYN (Eysenbach et al., 2019) which divides the visited state space into separate sections for each skill, or over goals from RIG (Nair et al., 2018) and Skew-Fit (Pong et al., 2020) which sample goals from the density estimate. We also investigate the importance of gradually increasing state coverage for the goal distribution by comparing it to EDL (Campos et al., 2020), and investigate how efficiently VUVC increases the visited state entropy.

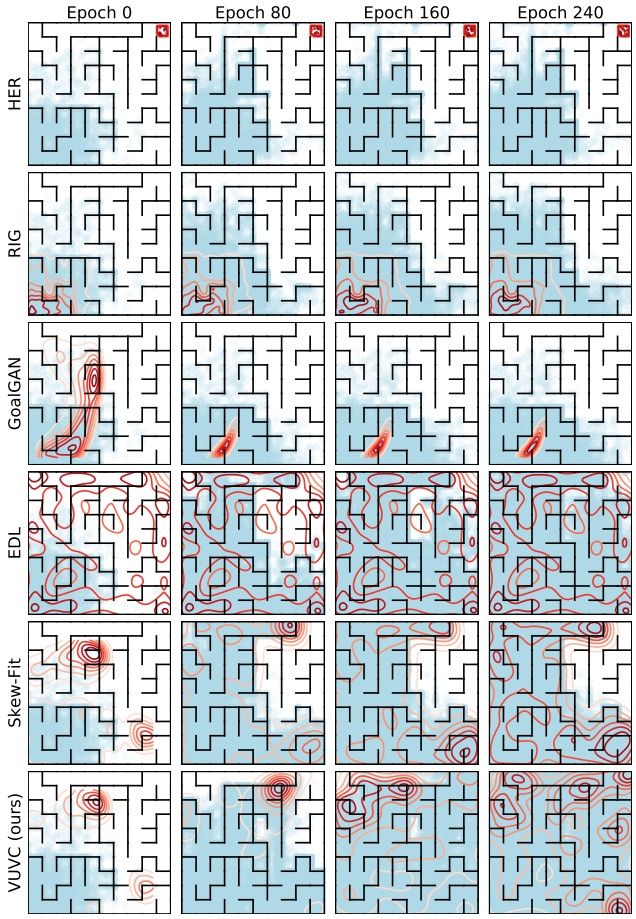

Figure 5. Curriculum goal distribution and accumulated visited states. The red contour line illustrates the curriculum goal distributions and cyan dots represent visited states by the agent. VUVC covers the state space significantly faster than the baselines.

5.2. Comparison of Sample Efficiency

We compare the number of required samples for task completion in various environments which are based on either configuration observation or image observation. Our experimental results illustrated in Figure 3 show that VUVC outperforms a variety of VCRL variants. Note that although EDL and EDL-Oracle take advantage of an additional training phase, VUVC outperforms them.

Point Maze Navigation Tasks VUVC successfully accomplishes all tasks, while some baseline methods fail. Especially in the complicated *PointMazeSquareLarge* environment, VUVC requires much less interaction for task completion. This result suggests the importance of an elaborate curriculum goal distribution in comparison to GoalGAN or Skew-Fit and emphasizes the importance of a gradually increasing state covering goal distribution when compared to EDL and EDL-Oracle.

Configuration-based Robotic Manipulation Tasks In all three tasks, VUVC significantly outperforms all baselines. It is also noteworthy that VUVC performs better than EDL-Oracle, even though our method does not make an excessive assumption (i.e., the need for an oracle uniform goal sampler). In comparison to Skew-Fit which also generates goals from a non-stationary distribution, the success rate of VUVC increases much faster. This result indicates to us that our method increases the entropy of the visited state distribution more efficiently than Skew-Fit.

Vision-based Robotic Manipulation Tasks VUVC presents the best performance compared to other VCRL variants in image observation environments. We train a policy in a latent space instead of directly training in an image space, as it has been shown that this solves RL problems in an image space efficiently (Nair et al., 2018), where an encoder of state density estimate model for a goal generator is used for a mapping function from an image observation to a latent observation. Even in a poorly-structured observation space, Figure 3 shows that VUVC consistently outperforms a variety of baseline methods. Note that DIAYN struggles in the *SawyerDoorHook* and *SawyerPickup* tasks as its policy remain close to the initial state during the training phase.

5.3. Impact of the Value Uncertainty

To see the effects of the value uncertainty in the curriculum, in Figure 4, we investigate (1) how the value uncertainty $U(g)$ and log density of visited states p_t^{visited} are correlated, and (2) how well the value uncertainty filters out uninformative states. In general, we observe that $U(g)$ and p_t^{visited} show negative correlation, indicating that VUVC covers the state space faster than the uniform curriculum for the optimal goal-conditioned policy as the regularity condition of Proposition 2 holds empirically. In addition to this, we observe a case which satisfies the regularity condition of Proposition 3 from the landscape visualization, implying that our method is more effective than the uniform curriculum. Uncertainty is low for easily reachable goals (yellow in the success rate landscape) as well as barely reachable goals (purple). On the other hand, uncertainty of goals that are moderately reachable (green) is high, which indicates that the value uncertainty focuses more on informative goals and results in better performance as we see in Figure 3.

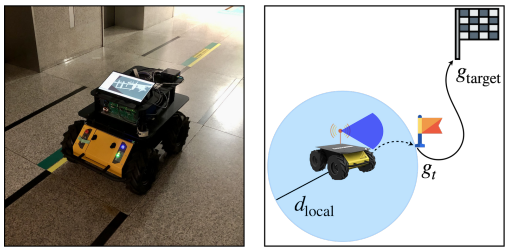

Figure 6. An illustrative example of how we utilize a global planner to generate a subgoal for our real robot platform.

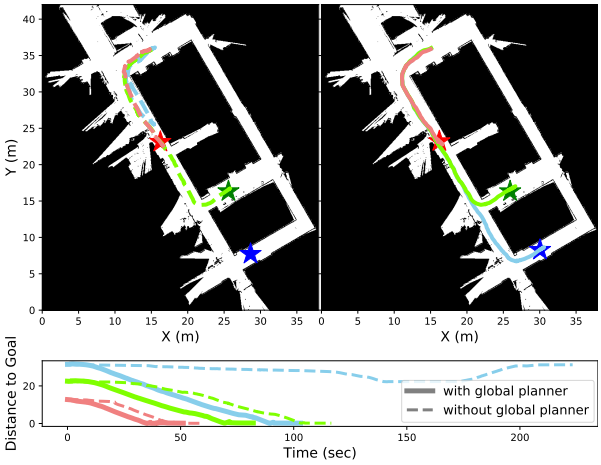

Figure 7. Building-scale navigation task with a real-world robot without (top left) and with global planner (top right). (Bottom) Evaluation on reaching the target goal.

5.4. Extensive Exploration for State Coverage

We next evaluate the effectiveness of our method by qualitatively comparing the speed of state coverage of each method in the *PointMazeSquareLarge* environment. Figure 5 demonstrates that VUVC efficiently increases the visited state entropy by considering the value uncertainty. Furthermore, after a sufficient number of exploration steps, the curriculum goal distribution induced by VUVC approaches a uniform distribution as the value uncertainty for every state converges to a consistent value. The results for other tasks are presented in Appendix F.3.

5.5. Deploying Skills on the Real-world Robot

We evaluate our method in a building-scale navigation task on the Husky A200 mobile robot which detects obstacles using a LiDAR sensor. We first apply our algorithm in a 2D navigation environment, and deploy learned navigation skills directly on the real robot in a zero-shot setup. Figures 6 and 7 show that the learned navigation skill can be directly used on our real-world robot without a manual design of complex reward functions and curriculum. We further demonstrate that combining learned navigation skills with

the help of a global planner improves navigation performance. The learned skill aims to reach the local goal g_t that is d_{local} away from the robot on the trajectory generated by the global planner. Figure 7 demonstrates that the learned skill combined with the global planner reaches the goal faster (solid line) than the learned skill itself (dashed line). Detailed description of the real-world experiment setup can be found in Appendix F.2.

6. Related Work

6.1. Curriculum RL

In GCRL, a goal relabeling scheme which samples goals from failed trajectories is proposed as an implicit curriculum method (Andrychowicz et al., 2017; Fang et al., 2018; Liu et al., 2018; Ding et al., 2019; Fang et al., 2019; Nair et al., 2018). Another line of work investigates curriculum generation methods that consider task difficulty. These methods explicitly model a curriculum generative model, generating goals based on task difficulty (Florensa et al., 2018; Racaniere et al., 2019), competence progress (Fournier et al., 2018), utilization of an additional agent (Narvekar & Stone, 2019), maximization of achieved goal distribution entropy with heuristic (Pitis et al., 2020), or progressive updating towards a predefined target distribution (Klink et al., 2020). However, prior works do not provide theoretical justification (Florensa et al., 2018; Racaniere et al., 2019), are limited to a given target distribution (Fournier et al., 2018; Narvekar & Stone, 2019; Klink et al., 2020), or depend on manually engineered heuristics (Pitis et al., 2020). The notion of uncertainty has been also considered in VDS (Zhang et al., 2020) which measures the uncertainty of the Q-functions to sample curriculum goals. However, this work lacks theoretical justification and assumes an oracle goal sampler accessing a uniform distribution over all valid states in a state space, which artificially ignores exploration problems by resetting the agent to any state in the environment, whereas our work does not require such an assumption.

6.2. Empowerment and Unsupervised Skill Learning

Recent studies on empowerment have studied the forms of mutual information-based objectives to learn state-covering skills (Campos et al., 2020; Pong et al., 2020), promote skill diversity (Achiam et al., 2018; Eysenbach et al., 2019; Liu et al., 2022), learn non-parametric reward functions (Warde-Farley et al., 2019), establish meta-training task distributions (Jabri et al., 2019), incorporate skill-transition dynamics models along with skill-conditioned policies for a model-based planning (Sharma et al., 2019), and enhance generalization through the successor feature framework (Hansen et al., 2020; Liu & Abbeel, 2021a). In addition, a number of works have studied how to extend empowerment to high-dimensional image space by using a non-parametric nearest neighbor to estimate entropy (Liu & Abbeel, 2021b; Yarats et al., 2021; Seo et al., 2021). However, most of this research assumes a fixed stationary distribution over skills (or goals) and there has been little exploration regarding the form of skill (or goal) distribution $p(z)$ (or $p(g)$). Compared to prior empowerment approaches, we investigate the effectiveness of curriculum skill distribution.

6.3. Uncertainty Quantification in RL

Measures of uncertainty have played a key role in RL. Bootstrapped DQN (Osband et al., 2016) uses a bootstrapping method to estimate the uncertainty of the Q-value, and utilizes it for efficient exploration. Plan2Explore (Sekar et al., 2020) leverages an ensemble of one-step predictive models to guide the exploration. Both bootstrapping and dropout methods are used to measure the uncertainty of the collision prediction model for safe navigation (Kahn et al., 2017). PBP-RNN (Benatan & Pyzer-Knapp, 2019) uses probabilistic backpropagation as an alternative to quantify uncertainty within a safe RL scenario. PETS (Chua et al., 2018) employs trajectory sampling with probabilistic dynamics models to bridge gap model-based RL and model-free RL.

6.4. Intrinsic Reward and Exploration

In a tabular setting, visit counts can be used as exploration bonus to encourage exploration (Strehl & Littman, 2008). Count-based exploration methods are further extended to non-tabular setting by introducing the pseudo-count (Bellemare et al., 2016; Ostrovski et al., 2017) or successor representation (Machado et al., 2020). Another common approach guides the agent based on prediction errors. For instance, squared prediction error in learned dynamics models is used as exploration bonus (Stadie et al., 2015). RND (Burda et al., 2019) uses errors in a randomly generated prediction problem that predicts the output of a fixed randomly initialized neural network given the observations. Our work enables agents to reach any previously visited states by learning goal-conditioned policies that cover the entire goal space. In contrast, exploration bonuses help agents visit novel states, but they cannot reuse learned policies to solve user-specified goals as those states are quickly forgotten.

7. Conclusion

We provide the unifying framework VCRL which recasts MI-based RL as curriculum learning in goal-conditioned RL. Under VCRL framework, we propose a novel approach VUVC for unsupervised discovery of skills which utilizes a value uncertainty for an increment in the entropy of the visited state distribution. Under regularity conditions, we prove that VUVC improves the expected entropy more than the uniform curriculum method. Our experimental results demonstrate that VUVC consistently outperforms a variety

of prior methods both on configuration-based and vision-based continuous robot manipulation tasks. We also demonstrate that VUVC enables a real-world robot to learn to navigate in a long-range environment without any explicit rewards, and that incorporating skills with a global planner further improves the performance.

Acknowledgements

This work was supported by the Industry Core Technology Development Project, 20005062, Development of Artificial Intelligence Robot Autonomous Navigation Technology for Agile Movement in Crowded Space, funded by the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation, No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

Achiam, J., Edwards, H., Amodei, D., and Abbeel, P. Variational option discovery algorithms. arXiv preprint arXiv:1807.10299, 2018.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. Advances in Neural Information Processing Systems (NeurIPS), 30, 2017.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. Advances in Neural Information Processing Systems (NeurIPS), 29, 2016.

Benatan, M. and Pyzer-Knapp, E. O. Fully bayesian recurrent neural networks for safe reinforcement learning. arXiv preprint arXiv:1911.03308, 2019.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In International Conference on Learning Representations (ICLR), 2019.

Campos, V., Trott, A., Xiong, C., Socher, R., Giró-i Nieto, X., and Torres, J. Explore, discover and learn: Unsupervised discovery of state-covering skills. In International Conference on Machine Learning (ICML), pp. 1317–1327. PMLR, 2020.

Choi, J., Sharma, A., Lee, H., Levine, S., and Gu, S. S. Variational empowerment as representation learning for goal-conditioned reinforcement learning. In International Conference on Machine Learning (ICML), pp. 1953–1963. PMLR, 2021.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. Advances in Neural Information Processing Systems (NeurIPS), 31, 2018.

Co-Reyes, J., Liu, Y., Gupta, A., Eysenbach, B., Abbeel, P., and Levine, S. Self-consistent trajectory autoencoder: Hierarchical reinforcement learning with trajectory embeddings. In International Conference on Machine Learning (ICML), pp. 1009–1018. PMLR, 2018.

Ding, Y., Florensa, C., Abbeel, P., and Phielipp, M. Goal-conditioned imitation learning. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. Go-explore: a new approach for hard-exploration problems. arXiv preprint arXiv:1901.10995, 2019.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. In International Conference on Learning Representations (ICLR), 2019.

Fang, M., Zhou, C., Shi, B., Gong, B., Xu, J., and Zhang, T. Dher: Hindsight experience replay for dynamic goals. In International Conference on Learning Representations (ICLR), 2018.

Fang, M., Zhou, T., Du, Y., Han, L., and Zhang, Z. Curriculum-guided hindsight experience replay. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Florensa, C., Held, D., Geng, X., and Abbeel, P. Automatic goal generation for reinforcement learning agents. In International Conference on Machine Learning (ICML), pp. 1515–1528. PMLR, 2018.

Fournier, P., Sigaud, O., Chetouani, M., and Oudeyer, P.-Y. Accuracy-based curriculum learning in deep reinforcement learning. arXiv preprint arXiv:1806.09614, 2018.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. Active inference and learning. Neuroscience & Biobehavioral Reviews, 68:862–879, 2016.

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., and Tenenbaum, J. World model learning and inference. Neural Networks, 2021.

Gregor, K., Rezende, D. J., and Wierstra, D. Variational intrinsic control. arXiv preprint arXiv:1611.07507, 2016.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International Conference on Machine Learning (ICML), pp. 1861–1870. PMLR, 2018.

Hansen, S., Dabney, W., Barreto, A., Warde-Farley, D., de Wiele, T. V., and Mnih, V. Fast task inference with variational intrinsic successor features. In International Conference on Learning Representations (ICLR), 2020.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations (ICLR), 2017.

Islam, R., Ahmed, Z., and Precup, D. Marginalized state distribution entropy regularization in policy optimization. arXiv preprint arXiv:1912.05128, 2019.

Jabri, A., Hsu, K., Gupta, A., Eysenbach, B., Levine, S., and Finn, C. Unsupervised curricula for visual meta-reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Kaelbling, L. P. Learning to achieve goals. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), volume 2, pp. 1094–8. Citeseer, 1993.

Kahn, G., Villaflor, A., Pong, V., Abbeel, P., and Levine, S. Uncertainty-aware reinforcement learning for collision avoidance. arXiv preprint arXiv:1702.01182, 2017.

Kastner, L., Cox, J., Buiyan, T., and Lambrecht, J. All-in-one: A drl-based control switch combining state-of-the-art navigation planners. In International Conference on Robotics and Automation (ICRA), pp. 2861–2867. IEEE, 2022.

Klink, P., D'Eramo, C., Peters, J. R., and Pajarinen, J. Self-paced deep reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 33, 2020.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. Empowerment: A universal agent-centric measure of control. In IEEE Congress on Evolutionary Computation, volume 1, pp. 128–135. IEEE, 2005.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems (NeurIPS), 30, 2017.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. arXiv preprint arXiv:1906.05274, 2019.

Liu, H. and Abbeel, P. Aps: Active pretraining with successor features. In International Conference on Machine Learning (ICML), pp. 6736–6747. PMLR, 2021a.

Liu, H. and Abbeel, P. Behavior from the void: Unsupervised active pre-training. Advances in Neural Information Processing Systems (NeurIPS), 34:18459–18473, 2021b.

Liu, H., Trott, A., Socher, R., and Xiong, C. Competitive experience replay. In International Conference on Learning Representations (ICLR), 2018.

Liu, J., Wang, D., Tian, Q., and Chen, Z. Learn goal-conditioned policy with intrinsic motivation for deep reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 7558–7566, 2022.

Machado, M. C., Bellemare, M. G., and Bowling, M. Count-based exploration with the successor representation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 5125–5133, 2020.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), pp. 2794–2802, 2017.

Marsiglietti, A. and Kostina, V. A lower bound on the differential entropy of log-concave random vectors with applications. Entropy, 20(3):185, 2018.

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. Discovering and achieving goals via world models. Advances in Neural Information Processing Systems (NeurIPS), 34:24379–24391, 2021.

Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. Advances in Neural Information Processing Systems (NeurIPS), 31, 2018.

Narvekar, S. and Stone, P. Learning curriculum policies for reinforcement learning. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), pp. 25–33, 2019.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. Advances in Neural Information Processing Systems (NeurIPS), 29, 2016.

Osband, I., Aslanides, J., and Cassirer, A. Randomized prior functions for deep reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS), 31, 2018.

Ostrovski, G., Bellemare, M. G., Oord, A., and Munos, R. Count-based exploration with neural density models. In International Conference on Machine Learning (ICML), pp. 2721–2730. PMLR, 2017.

Parr, T., Pezzulo, G., and Friston, K. J. Active inference: the free energy principle in mind, brain, and behavior. MIT Press, 2022.

Pitis, S., Chan, H., Zhao, S., Stadie, B., and Ba, J. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In International Conference on Machine Learning (ICML), pp. 7750–7761. PMLR, 2020.

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. arXiv preprint arXiv:1802.09464, 2018.

Pong, V., Dalal, M., Lin, S., and Nair, A. Rlkit. URL: https://github. com/vitchyr/rlkit, 2019.

Pong, V., Dalal, M., Lin, S., Nair, A., Bahl, S., and Levine, S. Skew-fit: State-covering self-supervised reinforcement learning. In International Conference on Machine Learning (ICML), pp. 7783–7792. PMLR, 2020.

Racaniere, S., Lampinen, A., Santoro, A., Reichert, D., Firoiu, V., and Lillicrap, T. Automated curriculum generation through setter-solver interactions. In International Conference on Learning Representations (ICLR), 2019.

Ren, Z., Dong, K., Zhou, Y., Liu, Q., and Peng, J. Exploration via hindsight goal generation. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Salge, C., Glackin, C., and Polani, D. Empowerment–an introduction. In Guided Self-Organization: Inception, pp. 67–114. Springer, 2014.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In International Conference on Machine Learning (ICML), pp. 1312–1320. PMLR, 2015.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In International Conference on Machine Learning (ICML), pp. 8583–8592. PMLR, 2020.

Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. State entropy maximization with random encoders for efficient exploration. In International Conference on Machine Learning (ICML), pp. 9443–9454. PMLR, 2021.

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. Dynamics-aware unsupervised discovery of skills. In International Conference on Learning Representations (ICLR), 2019.

Stadie, B. C., Levine, S., and Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. arXiv preprint arXiv:1507.00814, 2015.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for markov decision processes. Journal of Computer and System Sciences, 74(8):1309–1331, 2008.

Trott, A., Zheng, S., Xiong, C., and Socher, R. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Warde-Farley, D., de Wiele, T. V., Kulkarni, T., Ionescu, C., Hansen, S., and Mnih, V. Unsupervised control through non-parametric discriminative rewards. In International Conference on Learning Representations (ICLR), 2019.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Reinforcement learning with prototypical representations. In International Conference on Machine Learning (ICML), pp. 11920–11931. PMLR, 2021.

Zhang, Y., Abbeel, P., and Pinto, L. Automatic curriculum learning through value disagreement. Advances in Neural Information Processing Systems (NeurIPS), 33, 2020.

A. Limitations

We summarize the limitations of our work as follows:

- Although VUVC demonstrates significant improvements in both sample efficiency and the ability to cover the state space, the quantitative experimental results suggest that divergence during training is a potential problem, particularly in tasks such as *FetchPickAndPlace*, which has a higher dimensionality of state space compared to others. It is likely that divergence occurs in this task when there are states in a confined space that exhibit high value uncertainty, causing VUVC to focus on sampling goals around these states for a certain period of time.
- Demonstrating that the regularity conditions of Proposition 2 and 3 hold is limited in empirical study (see Figure 4). Therefore, it is an appealing research direction to rigorously show the regularity conditions hold.
- In our experiment, we consider a fixed initial state. Even though the core concept of our approach, which estimates the uncertainty of the learned value functions, remains applicable to variable initial states, its performance might be affected negatively due to the increased training data required to handle a wide range of initial states. We have not yet validated the scalability of our approach in environments with non-fixed initial states, and leave it as future work.

B. Relationship with Active Inference

Our work is also related to active inference (Friston et al., 2016; 2021; Parr et al., 2022). Active inference can play a crucial role in the context of world models, as it allows the agent to update its beliefs based on the actions performed by changing the gathered observations. For example, to efficiently learn a world model, Plan2Explore (Sekar et al., 2020) and LEXA (Mendonca et al., 2021) agents seeks out surprising states by leveraging ensembles of world models to guide their exploration. This can be related to our approach, VUVC, which seek out goals that the agent learns the most from. Moreover, in our approach, we focus on maximizing state-marginal mutual information $I(s; z)$, but if we maximize state-predictive mutual information $I(s'; z|s)$, as in the DADS method (Sharma et al., 2019), we would learn skill-transition dynamics models, which might be considered as world models. From an active inference perspective, this could lead the agent to select actions and collect observations in a manner that reduces the uncertainty associated with skill-transition dynamics.

C. Proofs

Proposition 1. *If $V_\psi(s_0, g)$ follows a log-concave distribution, then we have*

$$\mathcal{I}(V_\psi(s_0, g); \psi|s_0, g) \geq \log(2\sqrt{\mathrm{Var}[V_\psi(s_0, g)]}). \tag{8}$$

Proof. The mutual information can be rewritten as the difference between conditional entropy and marginal entropy, which correspond to, respectively, the aleatoric uncertainty and predictive entropy:

$$\mathcal{I}(V_\psi(s_0, g); \psi|s_0, g) = \mathcal{H}(V_\psi(s_0, g)|s_0, g) - \mathcal{H}(V_\psi(s_0, g)|\psi, s_0, g). \tag{17}$$

When the value function is deterministic with zero variance, maximizing the mutual information is equal to maximizing the marginal entropy. As shown in (Marsiglietti & Kostina, 2018), a lower bound on the entropy of a log-concave random variable can be derived in terms of the p-th absolute moment:

$$\mathcal{H}(V_\psi(s_0, g)|s_0, g) \geq \log\left(\frac{2\|V_\psi(s_0, g) - \mathbb{E}[V_\psi(s_0, g)]\|_p}{\Gamma(p+1)^{\frac{1}{p}}}\right), \tag{18}$$

where Γ denotes the Gamma function. Moreover, for $p = 2$, the bound tightens as

$$\mathcal{H}(V_\psi(s_0, g)|s_0, g) \geq \log(2\sqrt{\mathrm{Var}[V_\psi(s_0, g)]}), \tag{19}$$

which implies selecting a skill/goal that maximizes the disagreement in predictions of an ensemble of value functions is equivalent to maximizing the lower bound approximation of the mutual information. □

Proposition 2. *Given $\epsilon = \frac{1}{t}$ and $\rho^{\pi_\theta}(s|g) = \mathbb{I}(s = g)$, if*

$$\mathrm{Cov}[U(g), \log p_t^{\text{visited}}(g)] \leq 0, \tag{14}$$

and take $\epsilon \to 0$, then we have,

$$\begin{aligned}&\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} I_t = \\ &\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \left(\mathbb{E}_{g \sim p_t^{\text{VU}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] - \mathbb{E}_{g \sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] \right) > 0.\end{aligned} \tag{15}$$

Proof. Given a visited state s' at time $t+1$, the next step empirical distribution of the visited state can be written as

$$p_{t+1}^{\text{visited}}(s|s') = \frac{p_t^{\text{visited}}(s) + \epsilon \mathbb{I}[s = s']}{1 + \epsilon}. \tag{20}$$

With a curriculum goal g and a stationary state distribution induced by the policy $\rho^{\pi_\theta}(s|g)$, a next step empirical distribution of the visited state can be written as

$$\begin{aligned}p_{t+1}^{\text{visited}}(s) &= \sum_{s'} p_{t+1}^{\text{visited}}(s|s') \rho^{\pi_\theta}(s'|g) \\ &= \sum_{s'} \rho^{\pi_\theta}(s'|g) \left(\frac{p_t^{\text{visited}}(s) + \epsilon \mathbb{I}[s = s']}{1 + \epsilon} \right) \\ &= \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon}.\end{aligned} \tag{21}$$

Substituting the expression of $p_{t+1}^{\text{visited}}(s)$ into entropy increment over uniform curriculum gives

$$\begin{aligned}I_t &= \mathbb{E}_{g \sim p_t^{\text{VU}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] - \mathbb{E}_{g \sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] \\ &= \sum_g (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s - \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon} \log \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon}.\end{aligned} \tag{22}$$

We take the derivative with respect to ϵ and consider the asymptotic behavior where $\epsilon \to 0$:

$$\begin{aligned}&\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} I_t \\ &= \sum_g (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} - \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon} \log \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon} \\ &= \sum_g (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \lim_{\epsilon \to 0} - \frac{1}{1 + \epsilon^2} \left((\rho^{\pi_\theta}(s|g) - p_t^{\text{visited}}(s)) (\log \frac{p_t^{\text{visited}}(s) + \epsilon \rho^{\pi_\theta}(s|g)}{1 + \epsilon} + 1) \right) \\ &= \sum_g (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s - (\rho^{\pi_\theta}(s|g) - p_t^{\text{visited}}(s)) (\log p_t^{\text{visited}}(s) + 1) \\ &= \sum_g (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\rho^{\pi_\theta}(s|g) \log p_t^{\text{visited}}(s) + p_t^{\text{visited}}(s) \log p_t^{\text{visited}}(s) \right).\end{aligned}$$

Substituting $\rho^{\pi_\theta}(s|g) = \mathbb{I}(s = g)$ simplifies

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon}I_t &= \sum_g (p_t^{\mathrm{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\mathbb{I}[s=g]\log p_t^{\mathrm{visited}}(s) + p_t^{\mathrm{visited}}(s)\log p_t^{\mathrm{visited}}(s)\right) \\
&= \sum_g (p_t^{\mathrm{VU}}(g) - p_t^{\mathcal{U}}(g))(-\log p_t^{\mathrm{visited}}(g) - \mathbb{E}[-\log p_t^{\mathrm{visited}}(g)]) \\
&= -\sum_g (p_t^{\mathrm{VU}}(g) - \mathbb{E}[p_t^{\mathrm{VU}}(g)])(\log p_t^{\mathrm{visited}}(g) - \mathbb{E}[\log p_t^{\mathrm{visited}}(g)]) \\
&\quad + \sum_g (p_t^{\mathcal{U}}(g) - \mathbb{E}[p_t^{\mathcal{U}}(g)])(\log p_t^{\mathrm{visited}}(g) - \mathbb{E}[\log p_t^{\mathrm{visited}}(g)]) \\
&= -\mathrm{Cov}[U(g), \log p_t^{\mathrm{visited}}(g)],
\end{aligned}$$

where we use the fact that $p_t^{\mathrm{VU}}(g) = \frac{1}{Z_t} p_t^{\mathcal{U}}(g)U(g)$ and $p_t^{\mathcal{U}}(g) = \mathbb{E}[p_t^{\mathcal{U}}(g)]$ for all g. Thus we can complete the proof. □

Proposition 3. *Define the set* $\mathcal{G} = \mathcal{G}_{\mathrm{exploit}} \cup \mathcal{G}_{\mathrm{uninfo}} \cup \mathcal{G}_{\mathrm{info}}$ *and positive constant* Δ_1, Δ_2 *where*

$$\rho^{\pi_\theta}(s|g) = \begin{cases} \mathbb{I}(s=g) & \textit{for } g \in \mathcal{G}_{\mathrm{exploit}} \\ \rho^{\pi_\theta}_{\mathrm{uninfo}}(s|g) & \textit{for } g \in \mathcal{G}_{\mathrm{uninfo}} \\ \rho^{\pi_\theta}_{\mathrm{info}}(s|g) & \textit{for } g \in \mathcal{G}_{\mathrm{info}}, \end{cases} \tag{16}$$

for all $g \in \mathcal{G}_{\mathrm{uninfo}}$,

$$\mathbb{E}_{s\sim\rho^{\pi_\theta}_{\mathrm{uninfo}}(s|g)}[\log p_t^{\mathrm{visited}}(s)] = \log p_t^{\mathrm{visited}}(g) + \Delta_1,$$

and for all $g \in \mathcal{G}_{\mathrm{info}}$,

$$\mathbb{E}_{s\sim\rho^{\pi_\theta}_{\mathrm{info}}(s|g)}[\log p_t^{\mathrm{visited}}(s)] = \log p_t^{\mathrm{visited}}(g) - \Delta_2.$$

Given $\epsilon = \frac{1}{t}$, *if*

$$\begin{aligned}
\mathrm{Cov}[U(g), \log p_t^{\mathrm{visited}}(g)] &\le 0, \\
\mathbb{E}_{g\in\mathcal{G}_{\mathrm{uninfo}}}[p_t^{\mathrm{VU}}(g)] &\le \mathbb{E}_{g\in\mathcal{G}_{\mathrm{uninfo}}}[p_t^{\mathcal{U}}(g)], \\
\mathbb{E}_{g\in\mathcal{G}_{\mathrm{info}}}[p_t^{\mathrm{VU}}(g)] &\ge \mathbb{E}_{g\in\mathcal{G}_{\mathrm{info}}}[p_t^{\mathcal{U}}(g)],
\end{aligned}$$

and take $\epsilon \to 0$, *then we have,*

$$\begin{aligned}
&\lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon}I_t = \\
&\lim_{\epsilon\to 0}\frac{\partial}{\partial\epsilon}\left(\mathbb{E}_{g\sim p_t^{\mathrm{VU}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})] - \mathbb{E}_{g\sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\mathrm{visited}})]\right) > 0.
\end{aligned}$$

Proof. We substitute the $\mathcal{G} = \mathcal{G}_{\mathrm{exploit}} \cup \mathcal{G}_{\mathrm{uninfo}} \cup \mathcal{G}_{\mathrm{info}}$ into the entropy increment over uniform curriculum and expand the

expression:

$$
\begin{aligned}
&\lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} I_t \\
&= \sum_{g \in \mathcal{G}_{\text{exploit}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\mathbb{I}[s=g] \log p_t^{\text{visited}}(s) + p_t^{\text{visited}}(s) \log p_t^{\text{visited}}(s) \right) \\
&\quad + \sum_{g \in \mathcal{G}_{\text{uninfo}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\rho^{\pi_\theta}_{\text{uninfo}}(s|g) \log p_t^{\text{visited}}(s) + p_t^{\text{visited}}(s) \log p_t^{\text{visited}}(s) \right) \\
&\quad + \sum_{g \in \mathcal{G}_{\text{info}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\rho^{\pi_\theta}_{\text{info}}(s|g) \log p_t^{\text{visited}}(s) + p_t^{\text{visited}}(s) \log p_t^{\text{visited}}(s) \right) \\
&= \sum_{g \in \mathcal{G}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \sum_s \left(-\mathbb{I}[s=g] \log p_t^{\text{visited}}(s) + p_t^{\text{visited}}(s) \log p_t^{\text{visited}}(s) \right) \\
&\quad + \sum_{g \in \mathcal{G}_{\text{uninfo}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \left(\log p_t^{\text{visited}}(g) - \sum_s \rho^{\pi_\theta}(s|g) \log p_t^{\text{visited}}(s) \right) \\
&\quad + \sum_{g \in \mathcal{G}_{\text{info}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) \left(\log p_t^{\text{visited}}(g) - \sum_s \rho^{\pi_\theta}(s|g) \log p_t^{\text{visited}}(s) \right) \\
&= -\text{Cov}[U(g), \log p_t^{\text{visited}}(g)] - \Delta_1 \cdot \sum_{g \in \mathcal{G}_{\text{uninfo}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g)) + \Delta_2 \cdot \sum_{g \in \mathcal{G}_{\text{info}}} (p_t^{\text{VU}}(g) - p_t^{\mathcal{U}}(g))
\end{aligned}
$$

Following the assumptions, we can conclude the proof. □

D. Experimental Setup Details

D.1. Environments

As we described, we adopt the point mazes from the prior works (Zhang et al., 2020; Trott et al., 2019). Following these works, an agent observes a position of a point and takes action given a 2-dimensional goal position. For configuration-based robotic manipulation tasks, we utilize *Fetch* environments (Plappert et al., 2018) whose initial and goal distribution are modified to consider more complicated tasks, following the prior work (Ren et al., 2019). In these environments, an observation includes gripper position and velocity, gripper state, and object position and velocity. Given a 3-dimensional desired object position as a goal, it takes action to move its end-effector in Cartesian coordinates and open/close the gripper. For vision-based manipulation tasks, we adopt *Sawyer* environments (Nair et al., 2018) which manipulate a 7-DoF Sawyer robotic arm solely from a visual input without any explicit positional information of either a robotic arm or an object. A task to solve is given as a desired goal image which the agent should match its observation image with. *HuskyNavigate* is the environment in which we train navigation skills that we deployed on the real robot. An observation consists of the raw 2D laser measurements, the relative goal position, and current robot velocity. Given a 2-dimensional goal position, an action command which consists of linear velocity and angular velocity is given to the robot. Details about the environments are summarized in Table 2.

Parameter	**all *PointMaze***	**all *Fetch***	***SawyerDoorHook***	***SawyerPickup***	***SawyerPush***	***HuskyNavigate***
State space $\mathcal{S}$	$\in \mathbb{R}^2$	$\in \mathbb{R}^{25}$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^{364}$
Action space $\mathcal{A}$	$\in \mathbb{R}^2$	$\in \mathbb{R}^4$	$\in \mathbb{R}^3$	$\in \mathbb{R}^3$	$\in \mathbb{R}^2$	$\in \mathbb{R}^2$
Goal space $\mathcal{G}$	$\in \mathbb{R}^2$	$\in \mathbb{R}^3$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^{48\times48\times3}$	$\in \mathbb{R}^2$
Episode length	50	50	100	50	50	1000

Table 2. Environment details for each experiment.

D.2. Baseline Algorithms

We evaluate sample efficiency and state coverage speed of VUVC compared to the following baseline methods:

- **Hindsight Experience Replay (HER)** (Andrychowicz et al., 2017): HER is a naïve goal-conditioned RL method. The key idea of HER is to construct implicit curriculum goals by revisiting previous states in the experience replay. By storing additional trajectories using these curriculum goals, HER generates reward signals, even in situations where the initial sparse reward fails to provide meaningful feedback.

- **Reinforcement learning with Imagined Goals (RIG)** (Nair et al., 2018): RIG trains a goal-conditioned policy in an unsupervised manner by estimating the visited state distribution and automatically setting curriculum goals sampled from this distribution.

- **GoalGAN** (Florensa et al., 2018): GoalGAN encourages an agent to explore by suggesting curriculum goals from the generative model (Mao et al., 2017). To encourage an agent to explore the environment, it generates goals of intermediate difficulty where the difficulty of task (or goal) is measured from a success rate over some number of trials to solve the task. In vision-based robotic manipulation tasks, we compute the success rate in the latent space where an encoder is inherited from RIG which adopts a VAE as a state density estimate model.

- **Diversity Is All You Need (DIAYN)** (Eysenbach et al., 2019): DIAYN learns a latent skill based on mutual information maximization between skills and visited states with policy entropy regularization. It also reduces the mutual information between actions and skills, given the state, in order to separate the skills from each other, and partitions the visited state space into separate sections for each skill, each of which has a uniform stationary prior distribution.

- **Explore, Discover and Learn (EDL)** (Campos et al., 2020): EDL overcomes the limitation of existing variational empowerment methods which provide a poor coverage of the state space. Unlike RIG which makes use of the current goal-conditioned policy to approximate the state distribution, EDL utilizes a *fixed* uniform distribution over all $\mathcal{S}$ which would require an oracle sampler from the set of valid states. If the oracle is unavailable, an exploration policy is employed to induce the uniform distribution across valid states. The skill (or goal) distribution is inferred from this uniform state distribution by using a VAE. Then, the state-covering policy is trained based on the learned skill distribution following the variational empowerment objective.

- **Skew-Fit** (Pong et al., 2020): Skew-Fit aims to achieve a general-purpose policy that can accomplish new user-specified goals, in an unsupervised manner. To achieve such goal-conditioned policy, Skew-Fit estimates the visited state distribution like RIG, and skews this distribution with the negative exponent so that the skewed distribution converges to a uniform distribution over states. Under the assumption that the goal space is equivalent to the state space, goals are sampled from the skewed distribution when training, which implies that the goal distribution is non-stationary as the visited state space gets larger.

E. Implementation Details and Hyperparameters

For all experiments, the agents are trained with SAC (Haarnoja et al., 2018) with an automatically tuned entropy coefficient. During the training of the RL agent, we relabel transitions with goals by sampling from the curriculum goal distribution with probability 0.5 and the future goals with probability 0.3. We use β-VAE for both modeling a state density and computing an intrinsic reward, $\log q_\lambda(g|s)$. For DIAYN, we use a fixed set of 100 skills. To evaluate DIAYN's goal-reaching performance, we estimate the target skill from the desired goal using a discriminator. This estimated skill is then used as the goal for goal-conditioned policy. For RIG and Skew-Fit in *Sawyer* experiments, we use hyperparameters inherited from the official implementation of Skew-Fit (Pong et al., 2019). For the *PointMaze* and *Fetch* experiments, we add exploration noise into the action after a goal is reached by following the Go-Explore (Ecoffet et al., 2019) and normalize the observations using a running mean and standard deviation. For the *Sawyer* experiment, we normalize the image observations to be in the interval $[0, 1]$ by dividing by the maximum pixel intensity. Normalization is especially crucial for training β-VAE on all environments. For training an exploration policy in EDL, we use the entropy of the marginal state distribution (Co-Reyes et al., 2018; Islam et al., 2019) as a reward to encourage the agent to visit less visited states more.

The training time on a single NVIDIA Quadro 8000 GPU can range from 6 to 30 hours depending on the task and the situation.

`github.com/rail-berkeley/rlkit`

Hyperparameter	Value
Discount factor	0.98
Replay buffer size	1000000
Episode length	50
RL batch size	2048
Observation normalization	{**Yes**, No}
Polyak averaging coefficient for target networks	{0.001, **0.005**}
Policy hidden activation	ReLU
Policy learning rate	{0.0003, **0.001**, 0.003}
Q-Function hidden activation	ReLU
Q-Function learning rate	{0.0003, **0.001**, 0.003}
Ensemble size for quantifying value uncertainty	{**3**, 5, 7}
VAE batch size	256
VAE latent dimension size	2
VAE encoder activation	ReLU
VAE decoder activation	ReLU
VAE learning rate	{0.0003, **0.001**, 0.003}
β for β-VAE	{5, **10**, 20}
α for Skew	-1

Table 3. General hyperparameters used for all *PointMaze* and *Fetch* experiments. Values between brackets are tuned independently using a grid search.

Hyperparameter	all *PointMaze*	*FetchPush/PickAndPlace*	*FetchSlide*
Minimum # steps in replay buffer before training	5000	{5000, **20000**, 50000}	{5000, 20000, **50000**}

Table 4. Specific hyperparameters for all *PointMaze* and *Fetch* experiments. Values between brackets are tuned independently using a grid search.

F. Additional Experiments

F.1. Ablation Study: Ensemble Size

To study the robustness of the ensemble size for quantifying value uncertainty, we compare the ensemble size of 3, 5, and 7 in *PointMazeSquareLarge* whose results are shown in Figure 8. As illustrated in Figure 8, the performance of VUVC with different ensemble size does not differ substantially.

F.2. Real-World Robot Experiments

Setup The training of our navigation policy is performed using an OpenAI-gym-compatible simulator that we specially design to integrate it into the robot operating system (ROS). We generate an indoor map (Kastner et al., 2022) of size 25m $\times$ 25m, as shown in Figure 9, and simulate a robot with a 360° field of view 2D LiDAR sensor with a resolution set to 512. We also pre-define a collision threshold, where the agent would be notified of a collision if the sensor measurement is within the threshold. The navigation policy samples an action $a_t \in \mathbb{R}^2$, consisting of linear velocity $v_t \in [0.0, 0.5]$ and angular velocity $w_t \in [-0.64, 0.64]$, at 5Hz.

We illustrate an example of how VUVC can be used to assist a 2D mobile robot with its navigation task using a laser sensor whose training environment and results are shown in Figure 9. By using VUVC, the robot is able to extensively cover the state space, eventually reaching the full state space that would have otherwise been impossible to achieve without detouring through long-range navigation. In this environment, our method successfully discovers skills which enable a robot to complete a real-world navigation task in a zero-shot setup, and which can be incorporated with a global planner that further expands reachable distance.

Simulator is available at `github.com/leekwoon/nav-gym`

Hyperparameter	Value
Discount factor	0.99
Replay buffer size	100000
RL batch size	1024
Policy hidden activation	ReLU
Policy learning rate	0.001
Q-Function hidden activation	ReLU
Q-Function learning rate	0.001
Ensemble size for quantifying value uncertainty	3
VAE batch size	64

Table 5. General hyperparameters used for all *Sawyer* experiments.

Hyperparameter	***SawyerDoorHook***	***SawyerPickup***	***SawyerPush***
Episode length	100	50	50
VAE latent dimension size	16	16	4
β for β-VAE	20	30	20
α for Skew	-0.5	-1	-1

Table 6. Specific hyperparameters for the *Sawyer* experiments.

To evaluate the performance of the trained policy on a real robot, we deploy skills on a real Husky A200 mobile robot in the building which is depicted in the left of Figure 6. The local goal g_t is selected to be d_{local} away from the robot on the trajectory generated by the global planner, which utilizes the A* algorithm, at 2Hz. To clarify, we update a global plan from the A* algorithm where g_t is selected, and pass g_t to the goal-conditioned policy. We evaluate the performance of the trained policy in the simulator using two key metrics: 1) the robot's success in reaching the target goal and 2) the time required to traverse to it, as shown at the bottom of Figure 7. We compare these metrics for skills with and without a global planner. For navigation skills without a global planner, a fixed target goal is given, while for skills with a global planner, g_t is given. In a zero-shot setup, the robot is able to successfully navigate to two target goals (red and green stars) located 13m and 22m away from its initial position, respectively, without the use of a global planner. When assisted by a global planner, the robot is able to reach a farthest goal (a blue star) located 31m away from the initial position. Notably, the traverse time to the closest and intermediate goals are reduced from 43 seconds to 33 seconds and from 101 seconds to 67 seconds, respectively. These results demonstrate the effectiveness of our approach in the real-world.

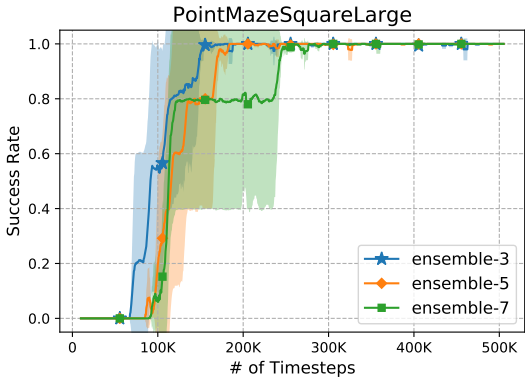

Figure 8. Learning curves for configuration-based point maze navigation task when the ensemble size is 3, 5, and 7, respectively. *Mean (SD)* of success rate over 5 random seeds are reported.

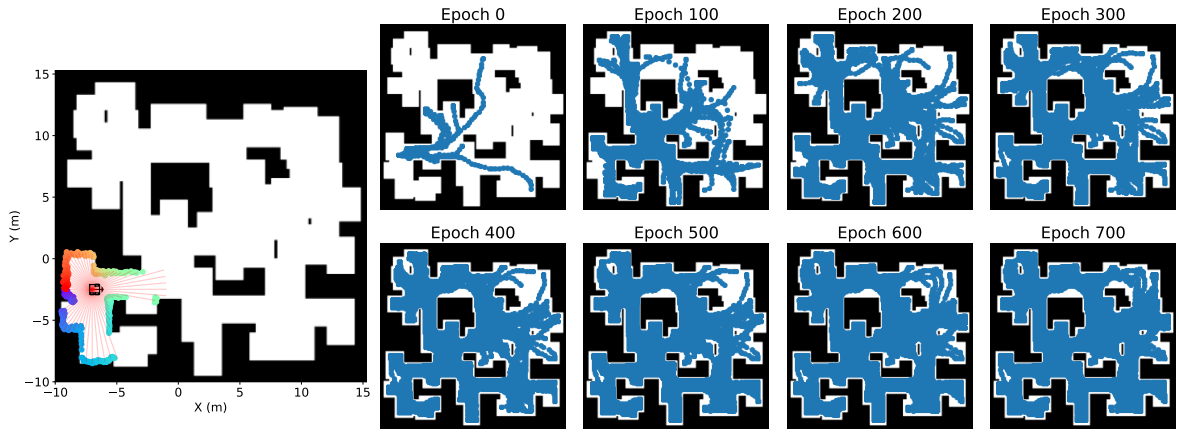

Figure 9. A simulation environment for training a mobile robot (left) and accumulated visited states for every 100 epoch (right). We illustrate an example of how VUVC can be used to assist a 2D mobile robot with its navigation task using a laser sensor. By using VUVC, the robot is able to cover more of the state space, eventually reaching the full state space that would have otherwise been impossible to achieve without detouring through long-range navigation.

F.3. Additional Results

Figure 10∼18 demonstrate the curriculum goal distribution and how it changes in the point navigation environments as well as in the robotic manipulation environments which we omit due to space constraint.

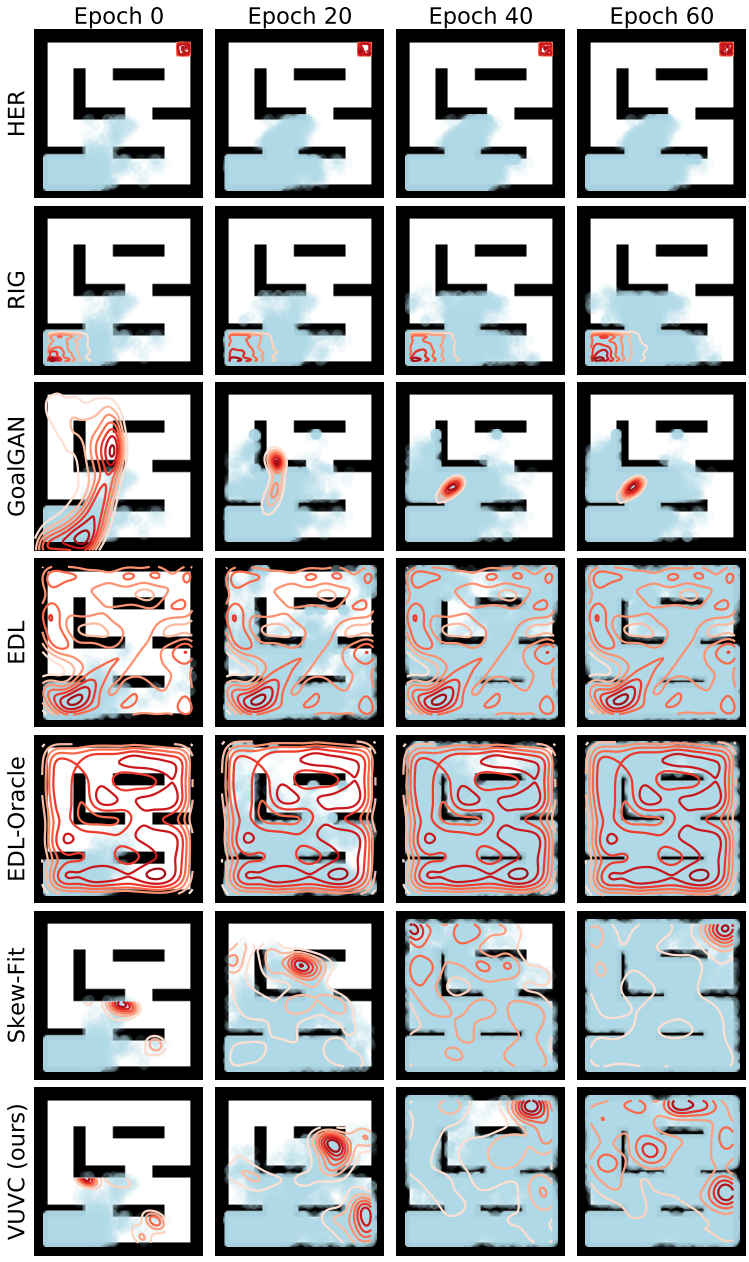

Figure 10. (*PointMazeA*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

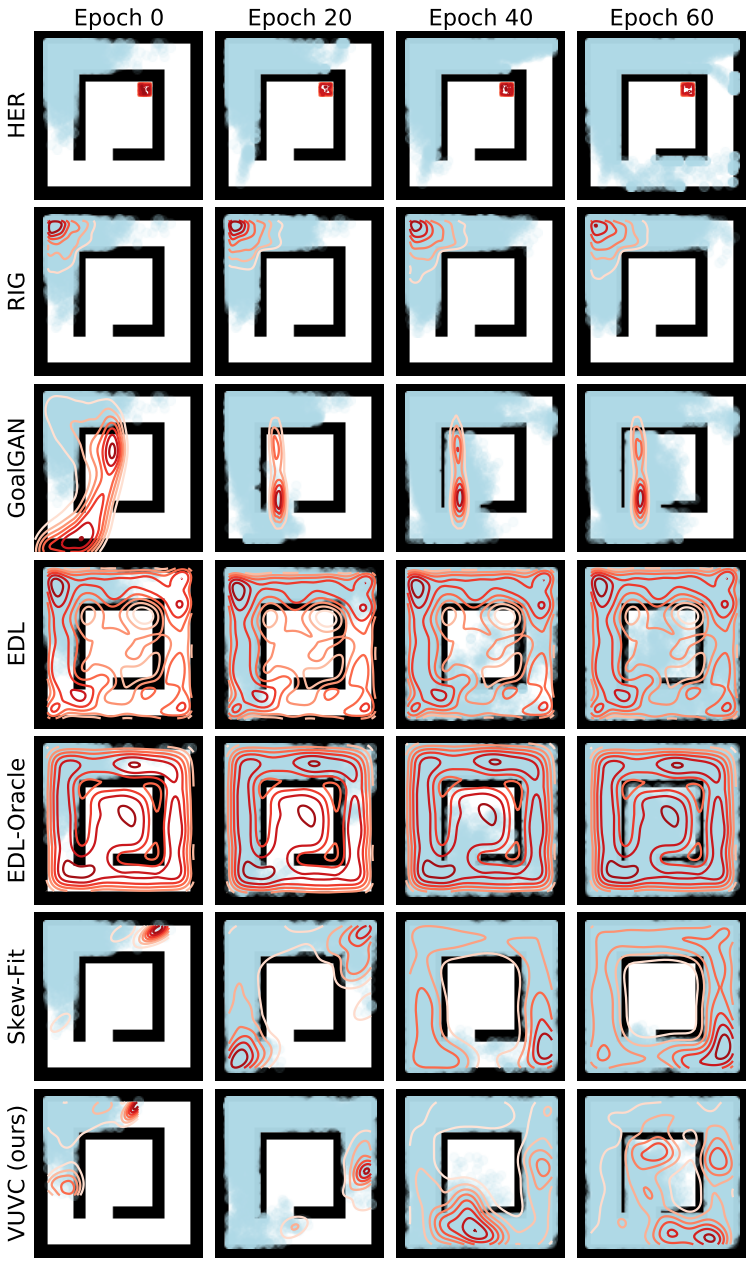

Figure 11. (*PointMazeB*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

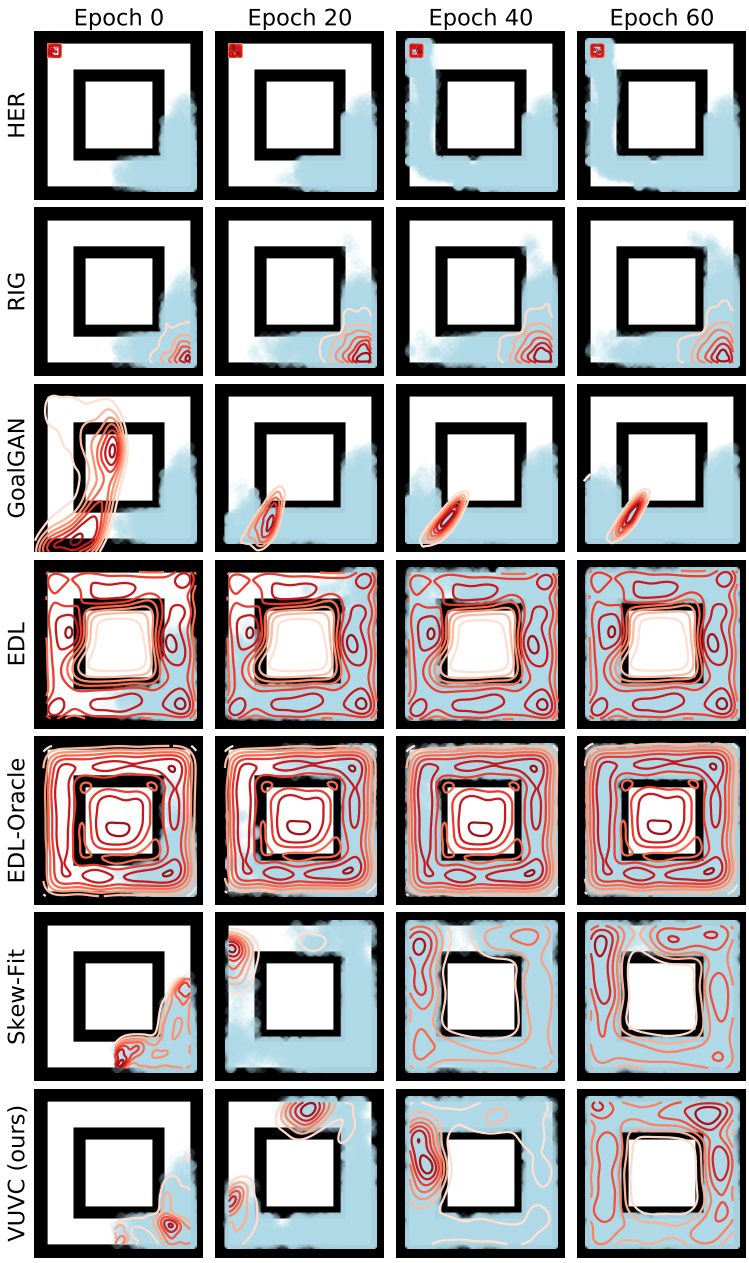

Figure 12. (*PointMazeC*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

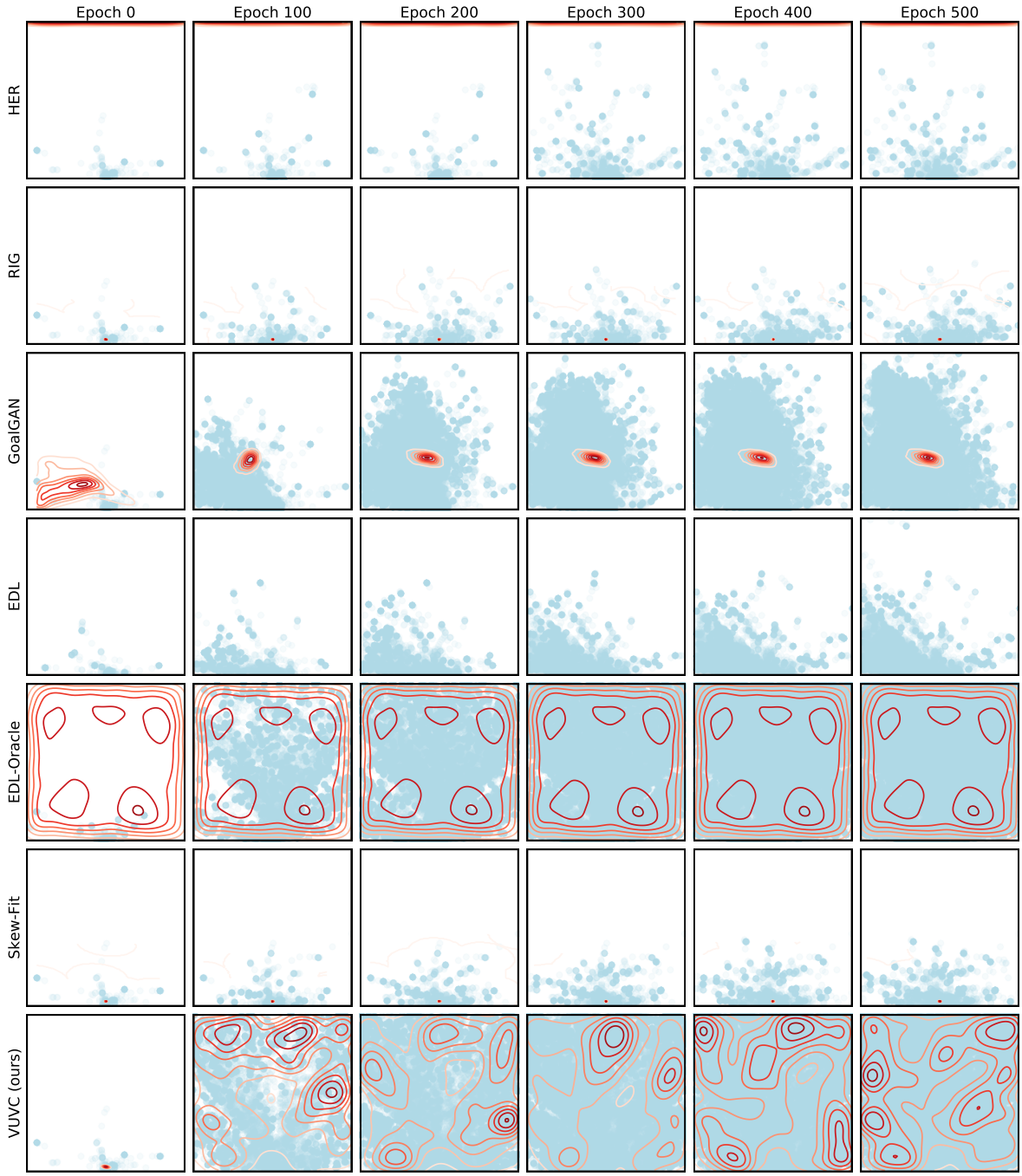

Figure 13. (*FetchPush*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

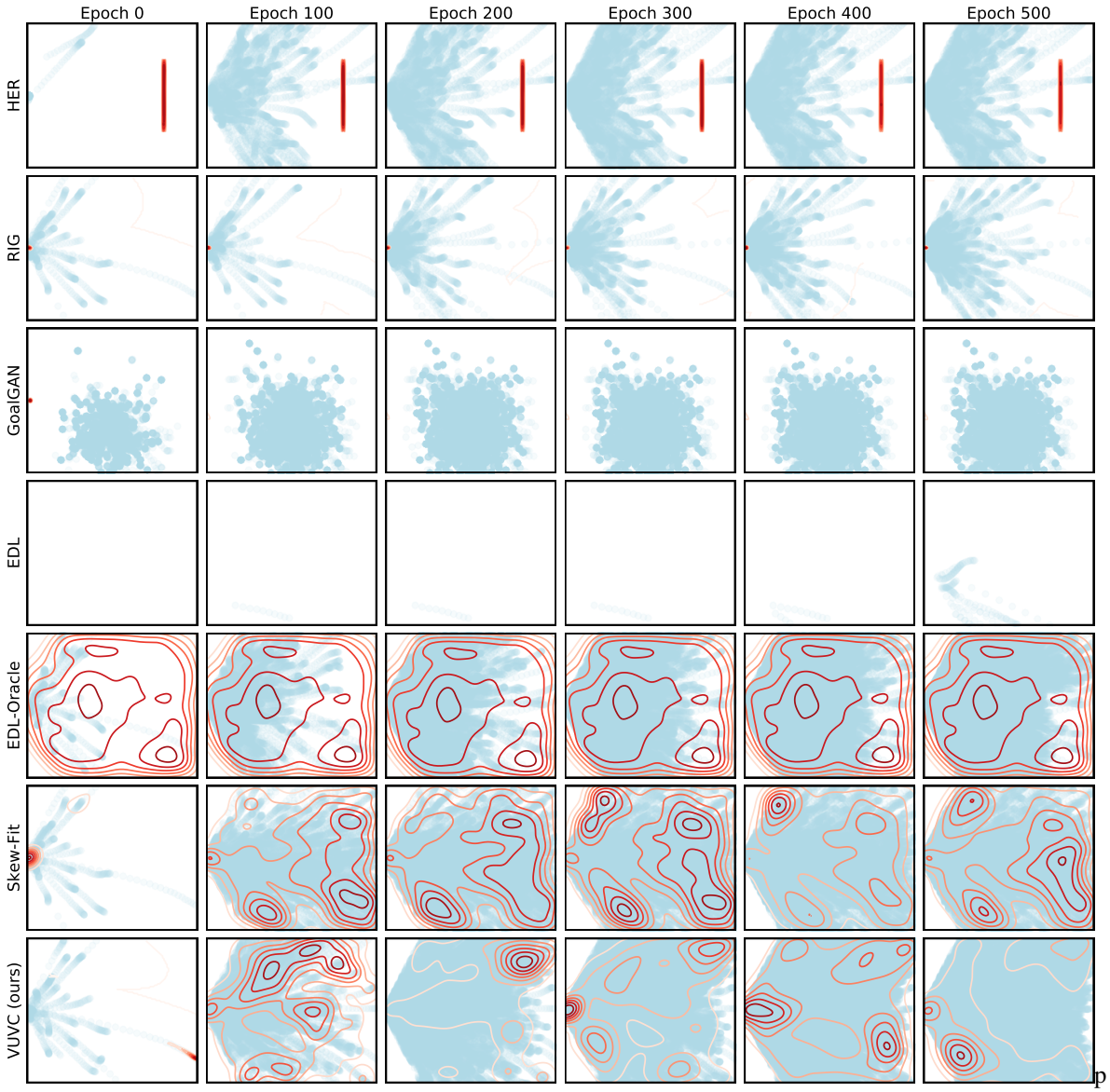

Figure 14. (*FetchSlide*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

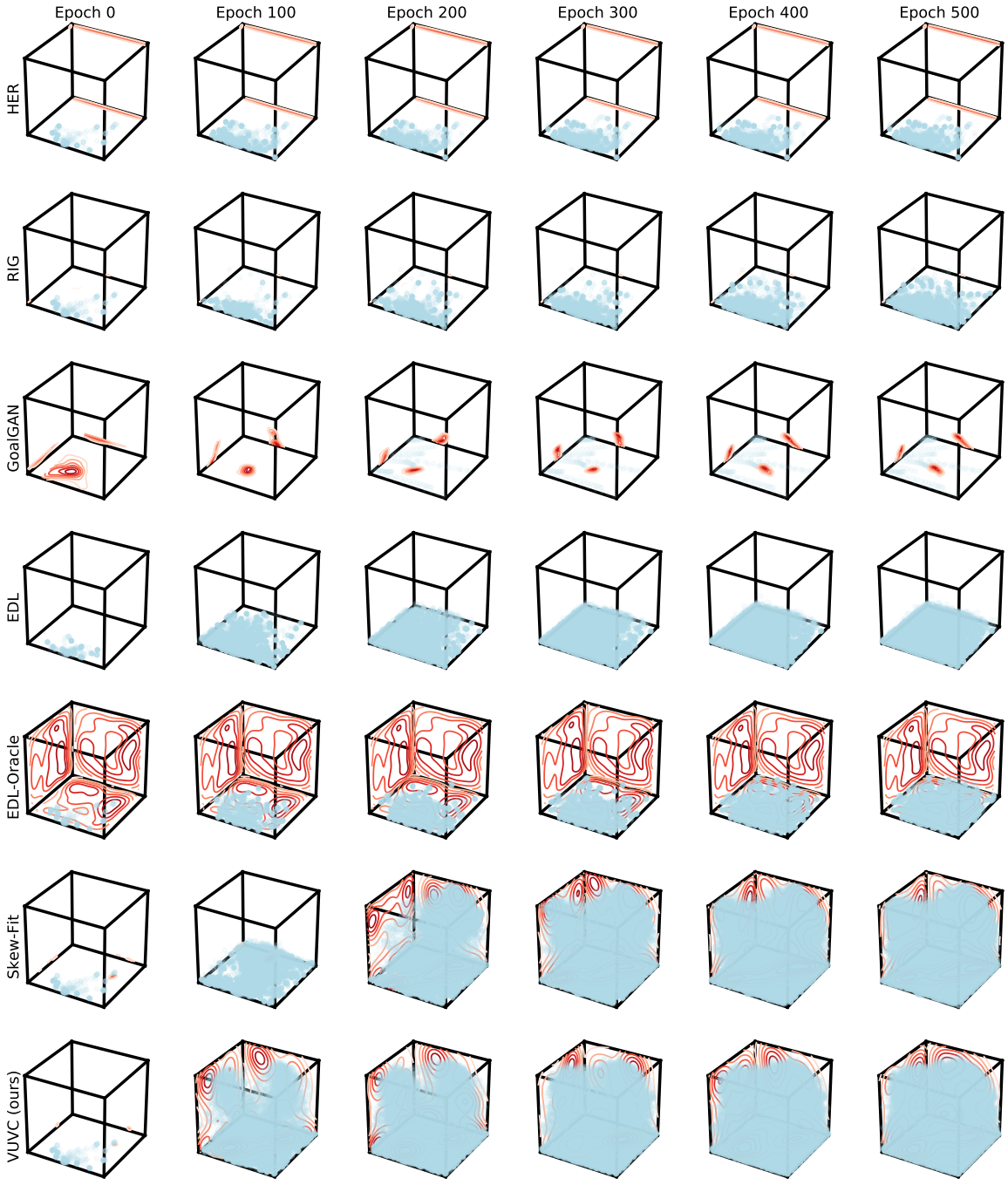

Figure 15. (*FetchPickAndPlace*) Curriculum goal distribution and accumulated visited states for a fixed seed for each method. The red contour line illustrates the curriculum goal distribution and cyan dots represent visited states by the agent.

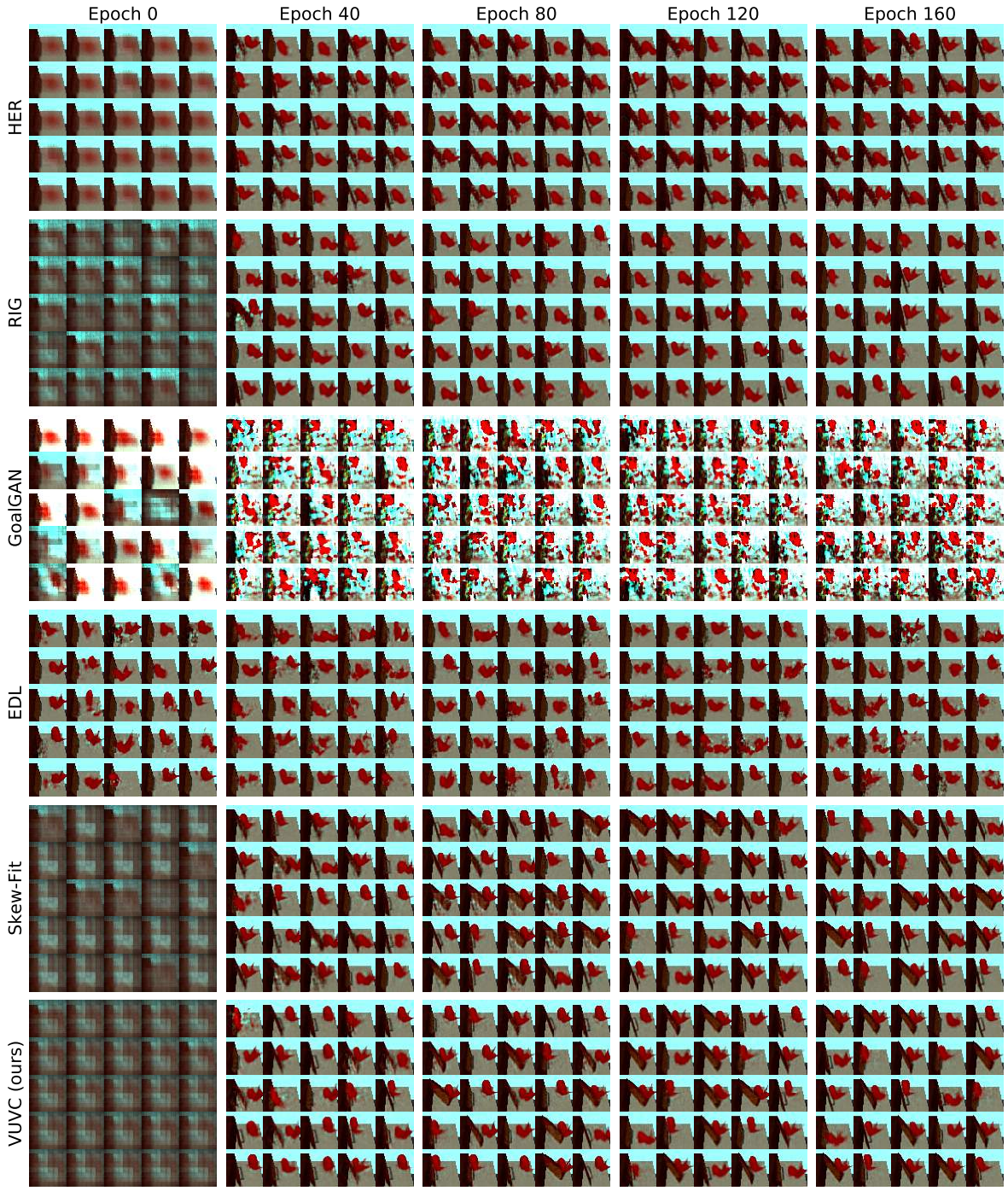

Figure 16. (*SawyerDoorHook*) Examples of curriculum goals for a fixed seed for each method. Latent codes are given as a curriculum goal and their reconstructed images are illustrated for visualization.

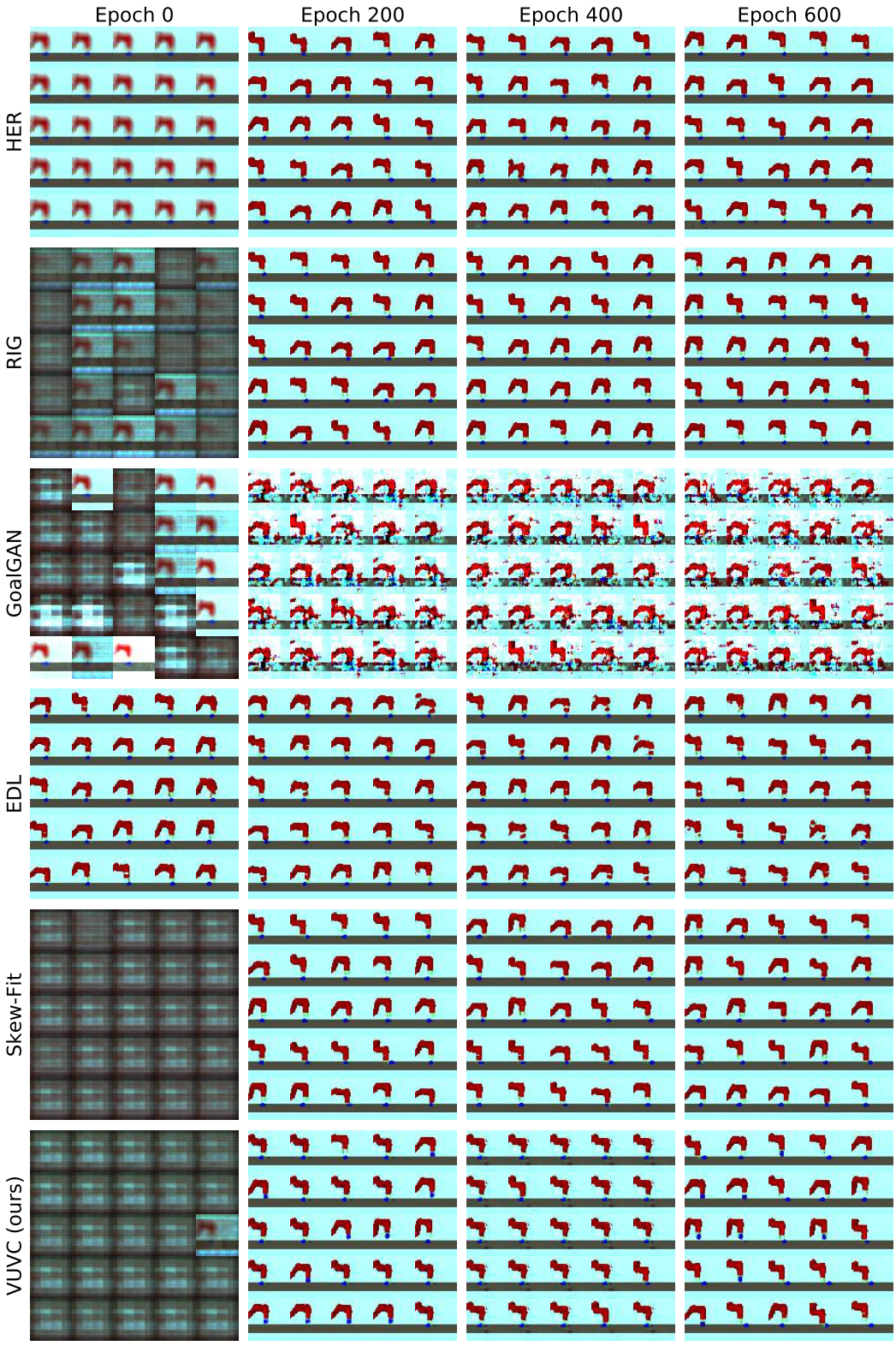

Figure 17. (*SawyerPickup*) Examples of curriculum goals for a fixed seed for each method. Latent codes are given as a curriculum goal and their reconstructed images are illustrated for visualization.

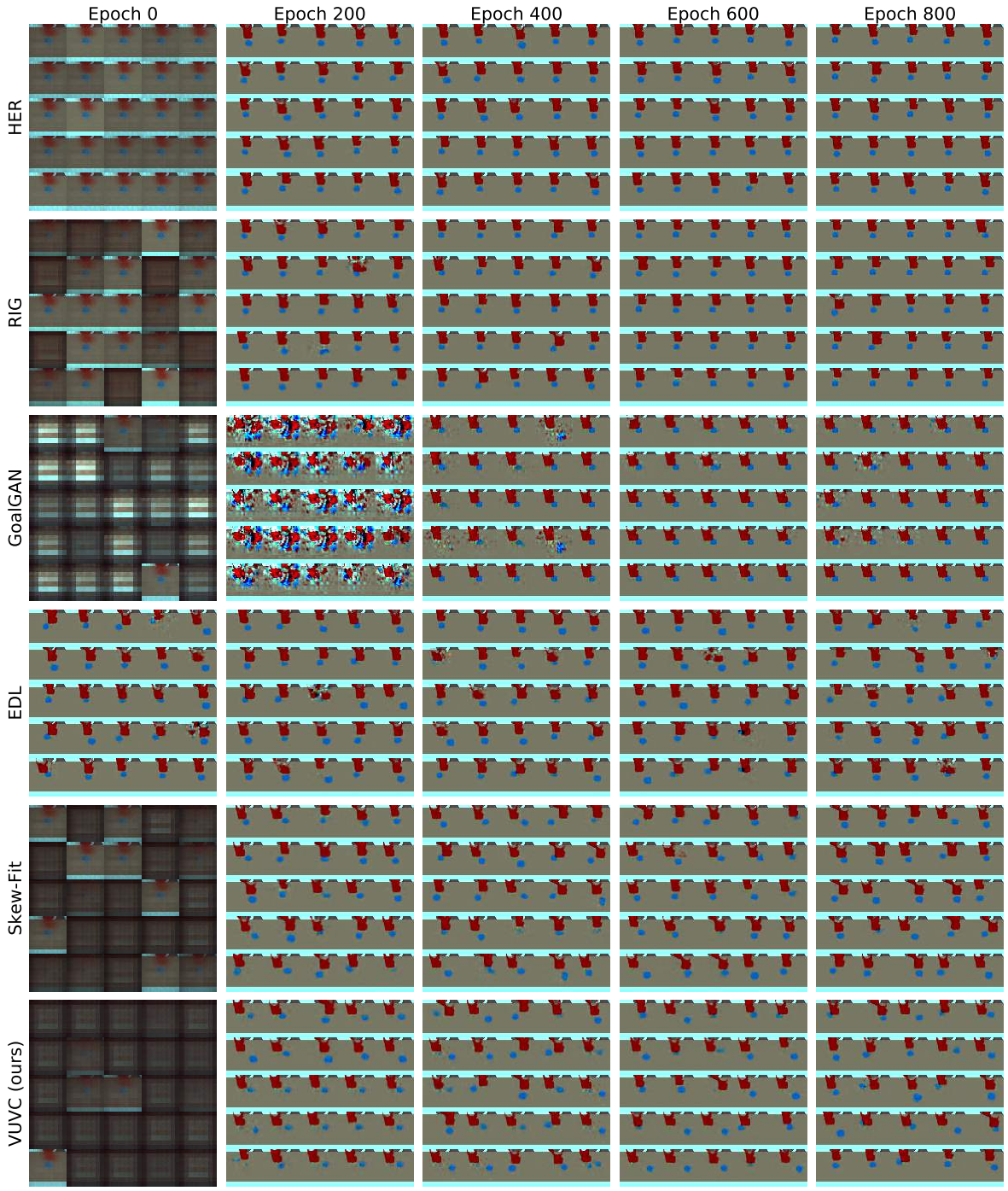

Figure 18. (*SawyerPush*) Examples of curriculum goals for a fixed seed for each method. Latent codes are given as a curriculum goal and their reconstructed images are illustrated for visualization.