Variational Curriculum Reinforcement Learning for Unsupervised Discovery of Skills

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

*Equal Contribution [1]KAIST [2]UNIST

Motivation

How can we train robot policies **efficiently**

- **without crafting the order of training skills** and ← curriculum reinforcement learning
- **without manually engineering objective functions** ← unsupervised skill discovery
- **to achieve complex skills** for our learning agent?

Contributions

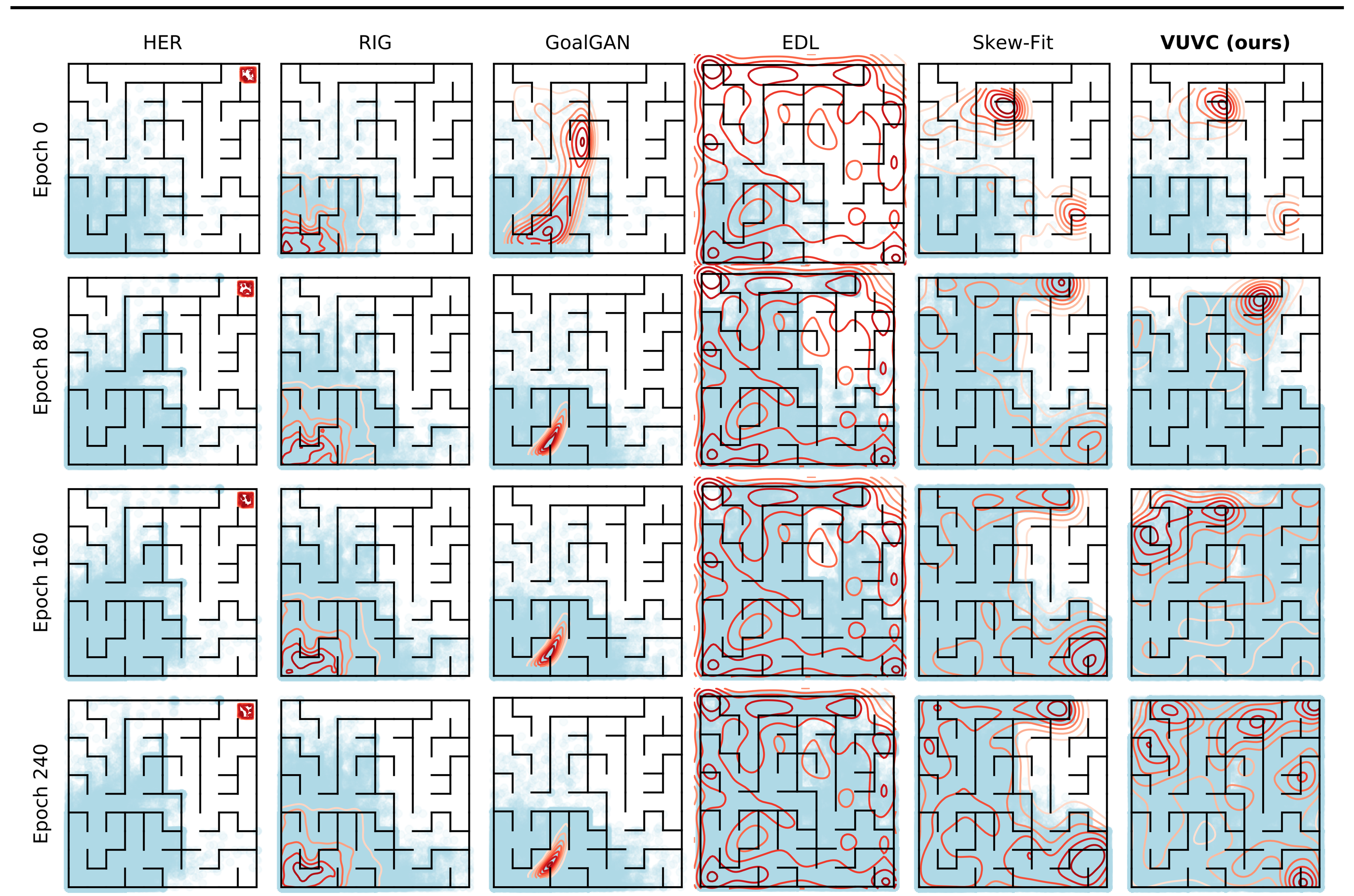
- Provide the unifying framework **Variational Curriculum Reinforcement Learning (VCRL)** encapsulating most of the prior mutual information based approaches.
- Propose **Value Uncertainty Variational Curriculum (VUVC)**, a value uncertainty based approach to information-theoretic skill discovery.

Variational Curriculum Reinforcement Learning (VCRL)

- Recast variational empowerment as curriculum learning in goal-conditioned RL.
- Objective of variational empowerment is to maximize a variational lower bound:

$$\mathcal{F}(\theta,\lambda) = \mathbb{E}_{\substack{g\sim p(g),\\ s\sim\rho^{\pi}(s|g)}} [\log q_\lambda(g|s) - \log p(g)], \quad (1)$$

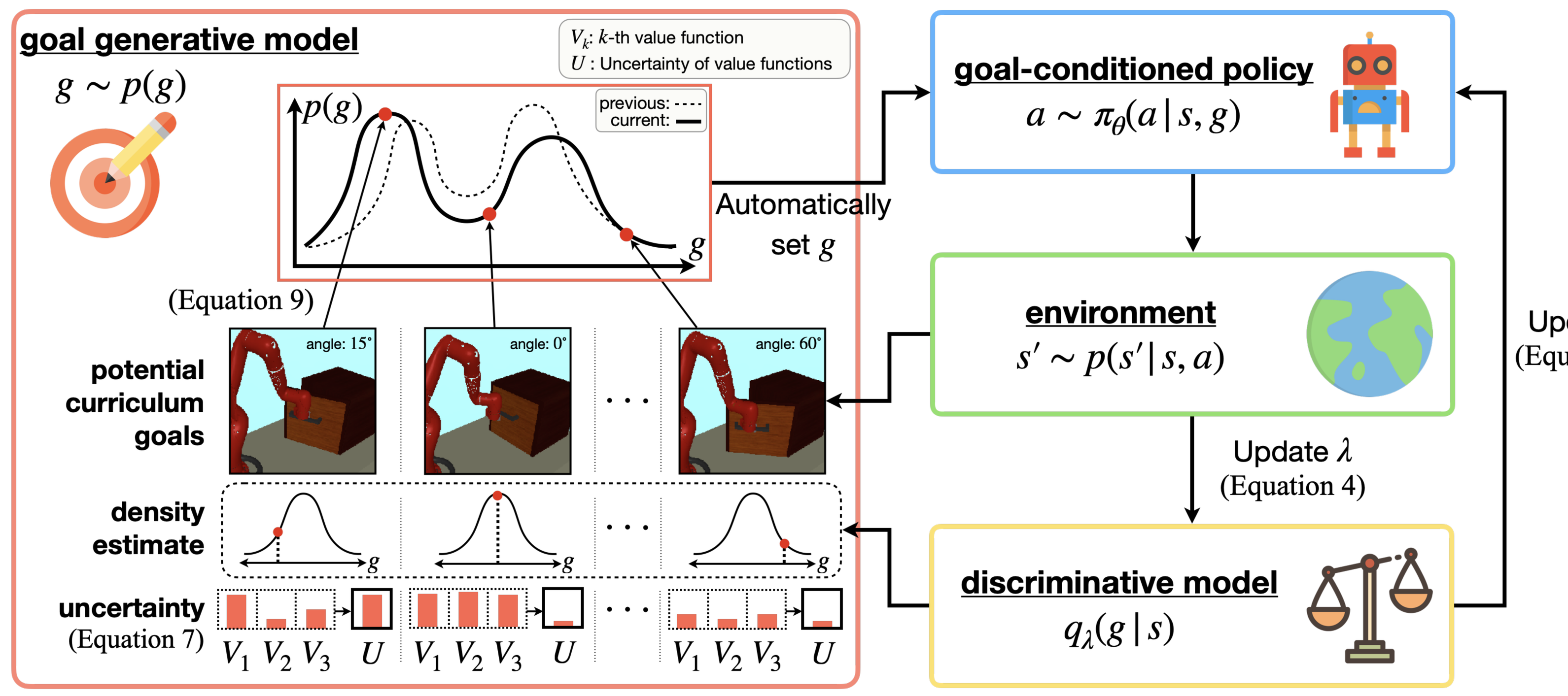
Methods	$q_\lambda(g\|s)$	$p(g)$	Non-stationary goal distribution
GCRL (w/ sparse reward)	$\frac{1}{Z}\exp(1-2\delta_g\mathcal{U}_{[s\pm\delta_g]})$	$p^{\text{target}}(g)$	✗
GCRL (w/ dense reward)	$\mathcal{N}(s,\sigma^2 I)$	$p^{\text{target}}(g)$	✗
EDL	$\mathcal{N}(\mu(s),\sigma^2 I)$	$p^{\text{explored}}(g)$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



Value Uncertainty Variational Curriculum (VUVC)

- Approach for unsupervised discovery of skills which utilizes a value uncertainty for an increment in the entropy of the visited state distribution.

Value Uncertainty Variational Curriculum



$$p_t^{\text{VUVC}}(g) = \frac{1}{Z_{t,\alpha}}U(g)p_t^{\text{visited}}(g)^\alpha, \quad \alpha\in[-1,0), \quad (2)$$

where $Z_{t,\alpha}$ is the normalizing coefficient.

Definition: Expected Entropy Increment over Uniform Curriculum

Given the empirical distribution of the visited state

$$p_t^{\text{visited}}(s) = \sum_{i=1}^{t}\frac{\mathbb{I}(s_i=s)}{t}, \quad (3)$$

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:

$$p_t^{\mathcal{U}}(g) = \mathcal{U}(\text{support}(p_t^{\text{visited}}))(g), \quad (4)$$

$$p_t^{\text{VU}}(g) = \frac{1}{Z_t}U(g)p_t^{\mathcal{U}}(g), \quad (5)$$

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

$$I_t = \mathbb{E}_{g\sim p_t^{\text{VU}}}[\mathcal{H}(p_{t+1}^{\text{visited}})] - \mathbb{E}_{g\sim p_t^{\mathcal{U}}}[\mathcal{H}(p_{t+1}^{\text{visited}})]. \quad (6)$$

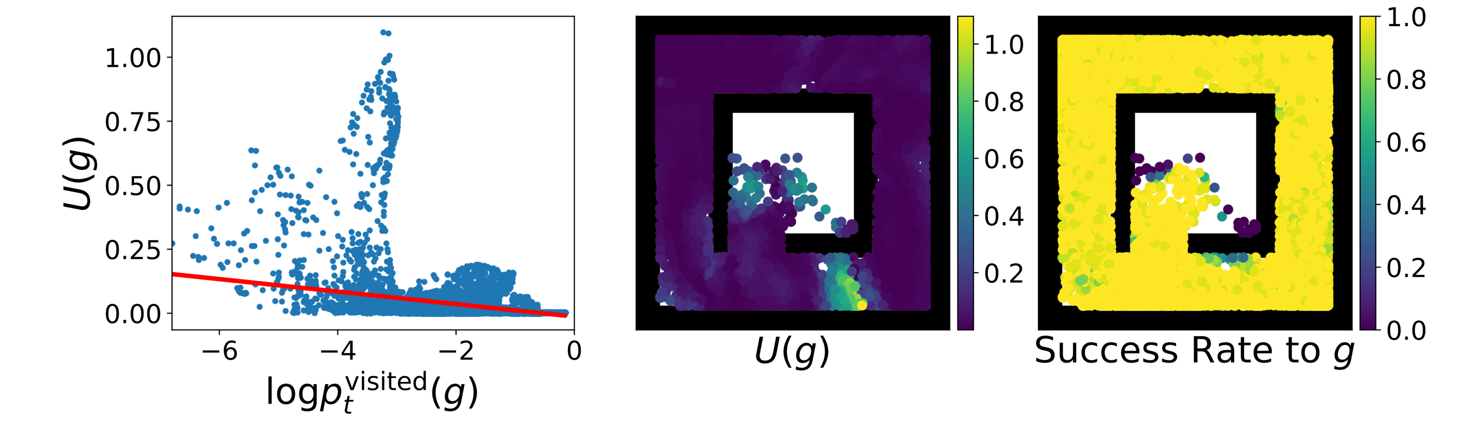
Proposition: VUVC Is At Least Better Than The Uniform Curriculum

With an accurate goal-conditioned policy and the model of dynamics, VUVC accelerates the increase of entropy in the visited states compared to the uniform curriculum, if the uncertainty of the learned value functions $U(g)$ and the log density of p_t^{visited} are negatively correlated.

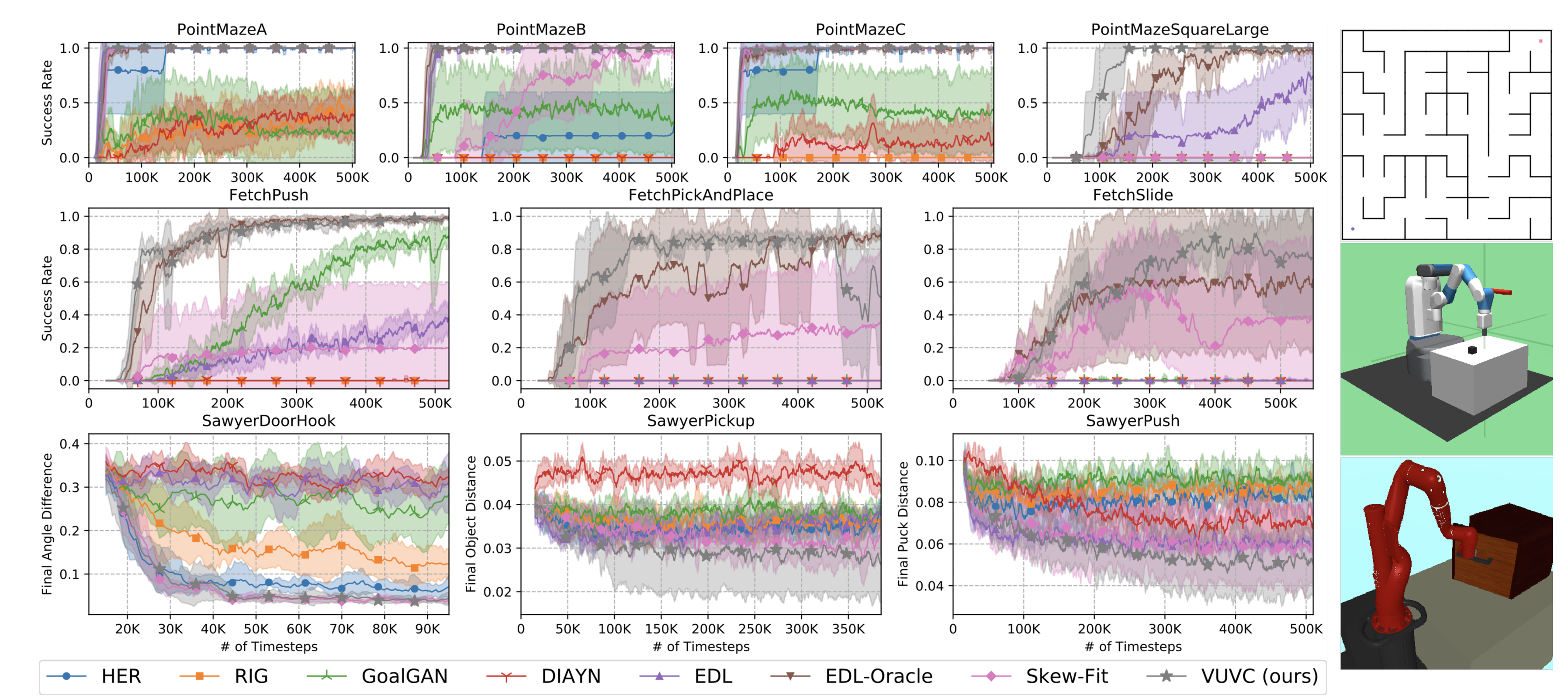
Proof Sketch. We begin by deriving a next step empirical distribution of the visited state given a curriculum goal g and a stationary state distribution induced by the policy $\rho^{\pi_\theta}(s|g)$, which can be written as $p_{t+1}^{\text{visited}}(s) = \frac{p_t^{\text{visited}}(s)+\epsilon\rho^{\pi_\theta}(s|g)}{1+\epsilon}$.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

Experiments

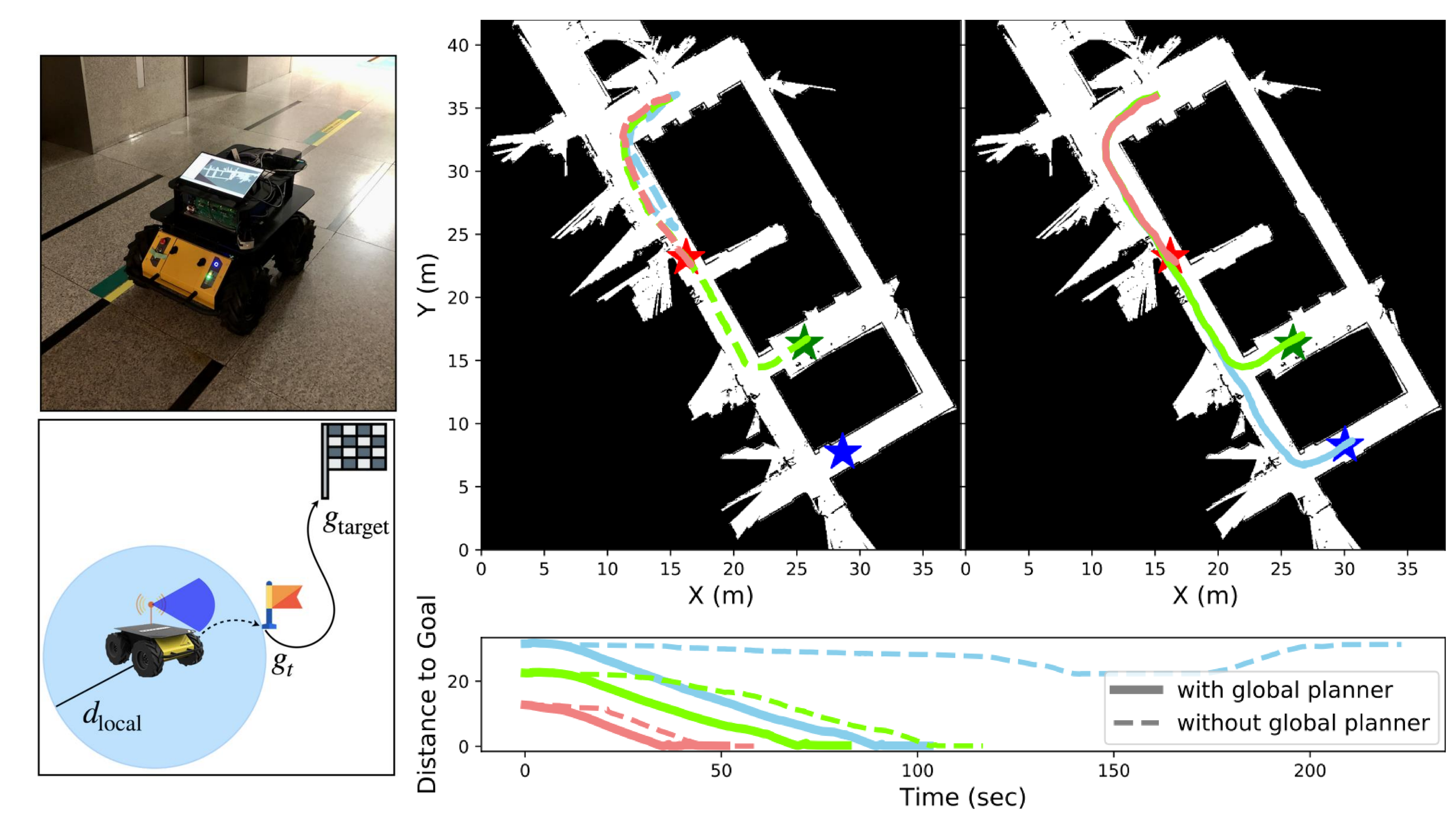
Impact of Value Uncertainty



Comparison of Sample Efficiency



Deploying Skills on the Real-world Robot



seongun@kaist.ac.kr, leekwoon@unist.ac.kr