Pre-Training Goal-based Models for Sample-Efficient Reinforcement Learning


Haoqi Yuan


Zhancun Mu

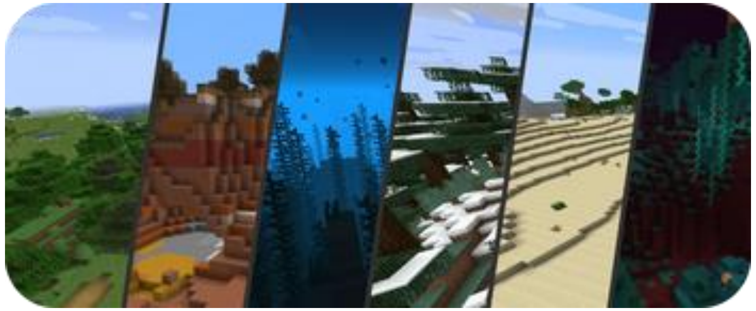

Feiyang Xie


Zongqing Lu

Pre-Training for Reinforcement Learning (RL)

- Deep RL suffers from low sample efficiency in complex, long-horizon tasks.

Long-horizon task



Complex environment

RL

Pre-Training for Reinforcement Learning (RL)

- Access to large datasets, such as human gameplay data from the Internet.

Minecraft human gameplay datasets
(Baker et al., 2022; Fan et al., 2022)

- Environment info
- Agent behavior

Pre-Training for Reinforcement Learning (RL)

- Pre-training from datasets can learn **useful priors** for RL, improving sample efficiency.

Pre-training on task-agnostic datasets

RL for downstream tasks

Pre-Training for Reinforcement Learning (RL)

- Pre-training from datasets can learn **useful priors** for RL, improving sample efficiency.

Task-agnostic dataset

$\{\tau = \{(s_i, a_i)\}_{i=0}^{T}\}$

Skills
SPiRL (Pertsch et al., 2021)
SkiMo (Shi et al., 2022)

Reward/value functions
ELE (Bruce et al., 2023)
VIP (Ma et al., 2023)

State representations
R3M (Nair et al., 2022)
MVP (Xiao et al., 2022)

Pre-training

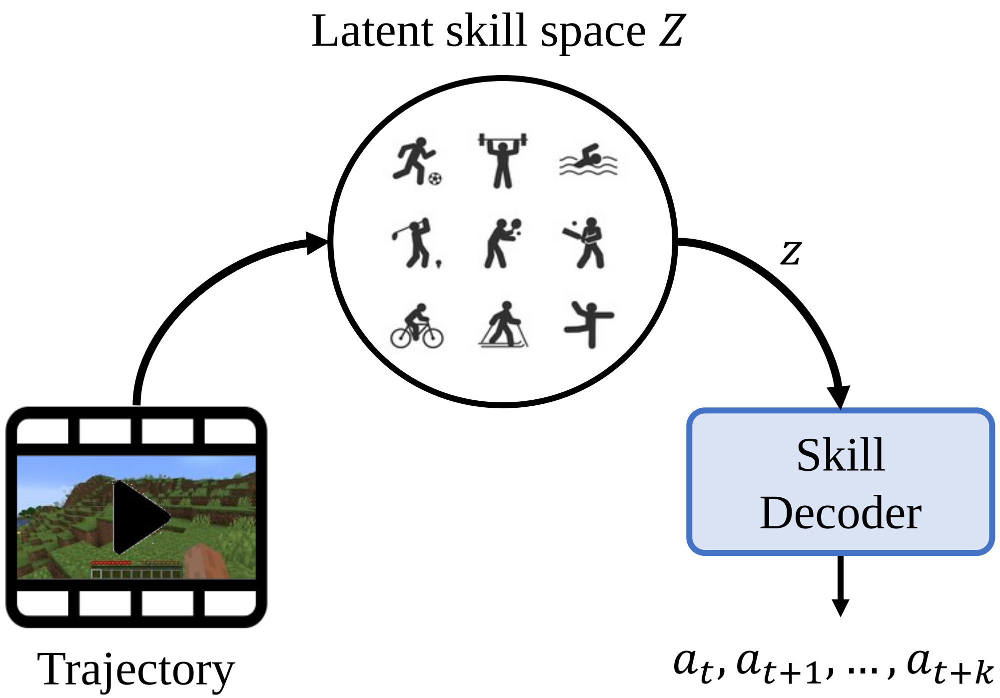
Complex, long-horizon downstream tasks.

RL for downstream tasks

$$\max \sum_t \gamma^t R(s_t, a_t)$$

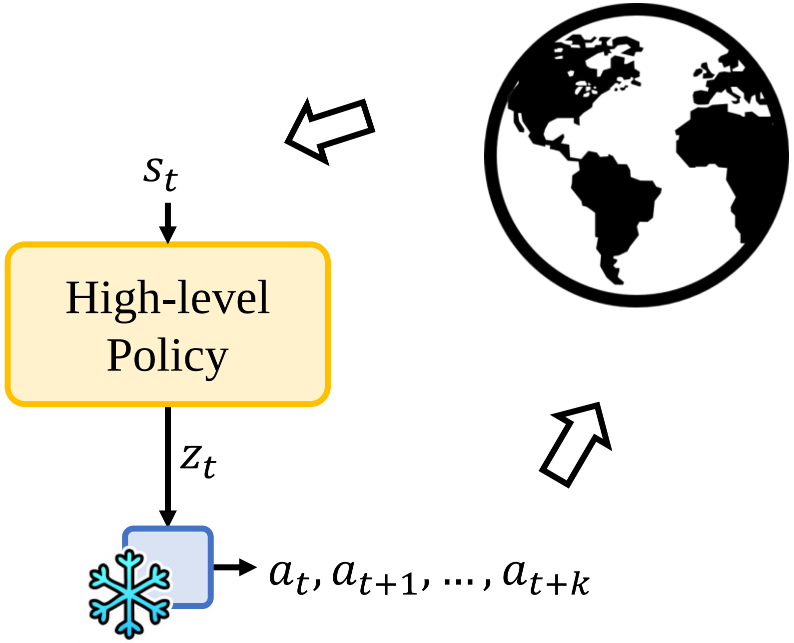
Issues in Skill Pre-Training

Skill pre-training



$$\mathcal{L} = \mathcal{L}_{\text{recon}}(\hat{a}, a) + \mathcal{L}_{\text{KL}}\big(p(z) || \mathcal{N}(0, I)\big)$$

Downstream RL



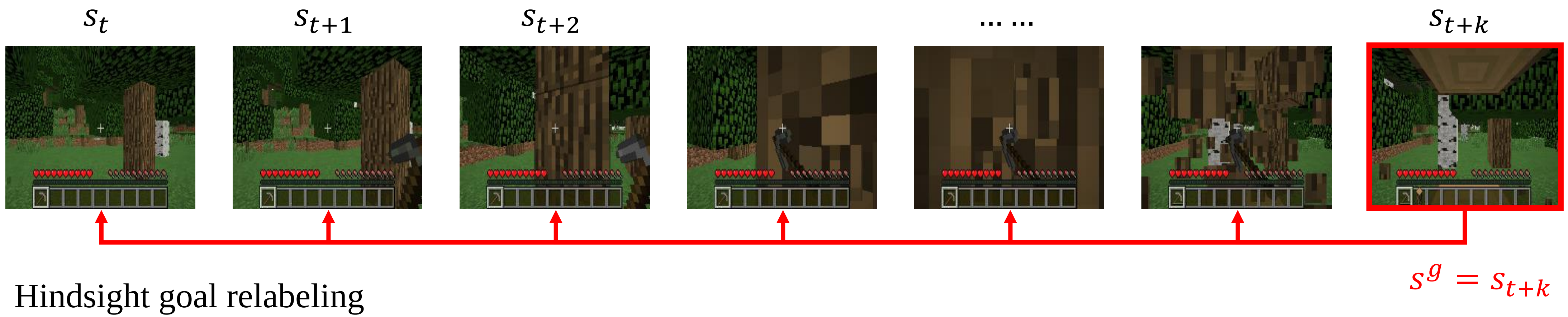
(1) VAE: trade-off between the action prediction accuracy and KL loss

(2) Continuous high-level action space Z

SPiRL(Pertsch et al., 2021); SkiMo (Shi et al., 2022); ASPiRe (Xu et al., 2022); TACO-RL (Rosete-Beas et al., 2022).

Goal-Conditioned Skill

- To address (1), we adopt a goal-conditioned behavior cloning approach (Lifshitz et al., 2023) to learn diverse skills, without trade-offs in loss functions.



- Hindsight goal relabeling

- Goal-conditioned behavior cloning

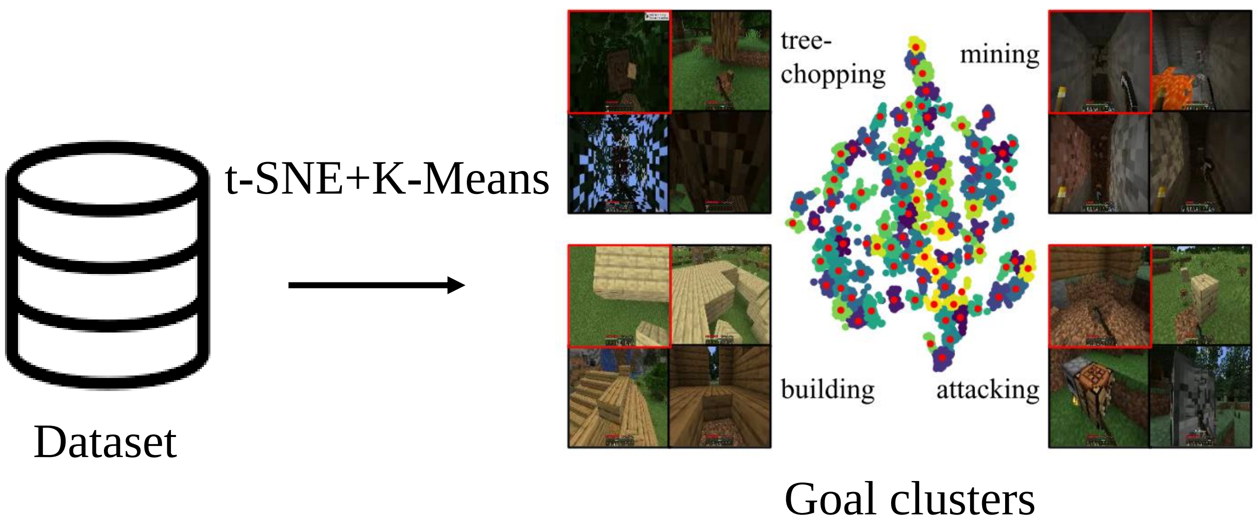
s^g → Goal-Conditioned Policy → a_i
s_i → Goal-Conditioned Policy

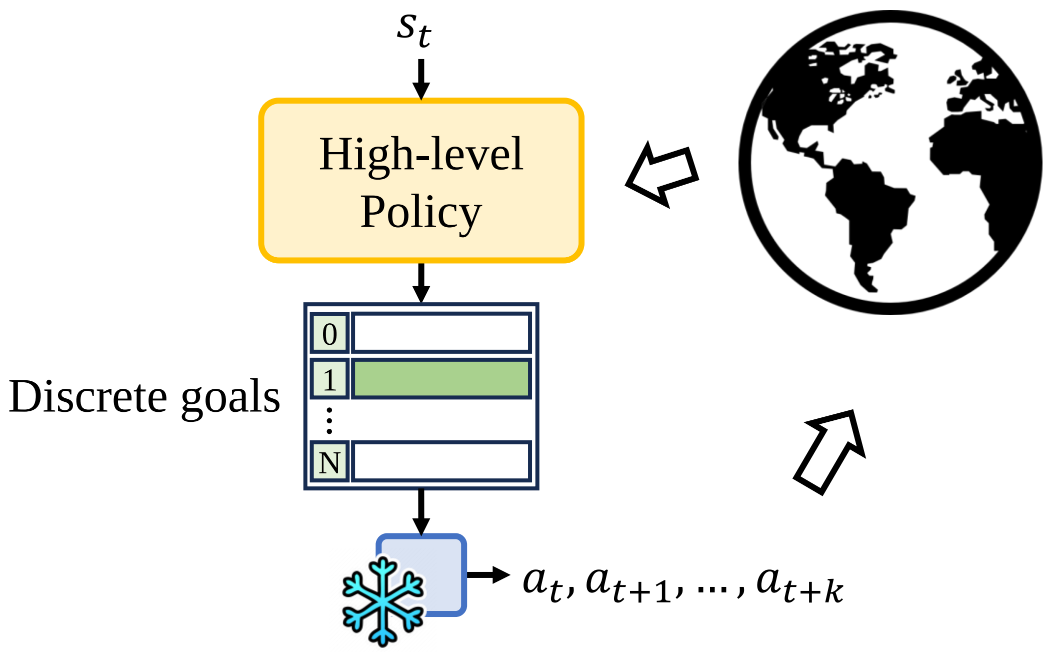
$$\mathcal{L}(\phi) = \mathbb{E}_D \left[-\log P_\phi(a_i | s_i, s^g) \right]$$

→ perform a variety of behaviors depending on the given goals s^g.

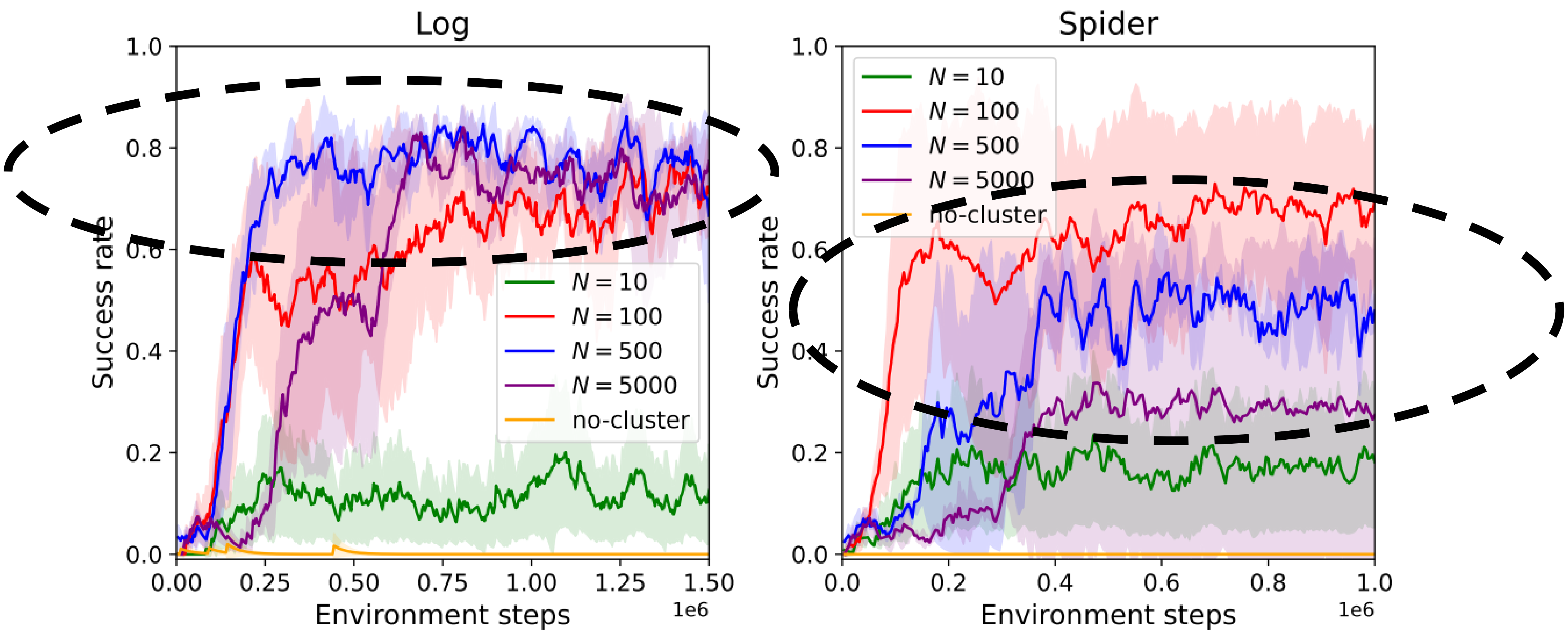
Goal Clustering

- To address (2), we propose a clustering approach.

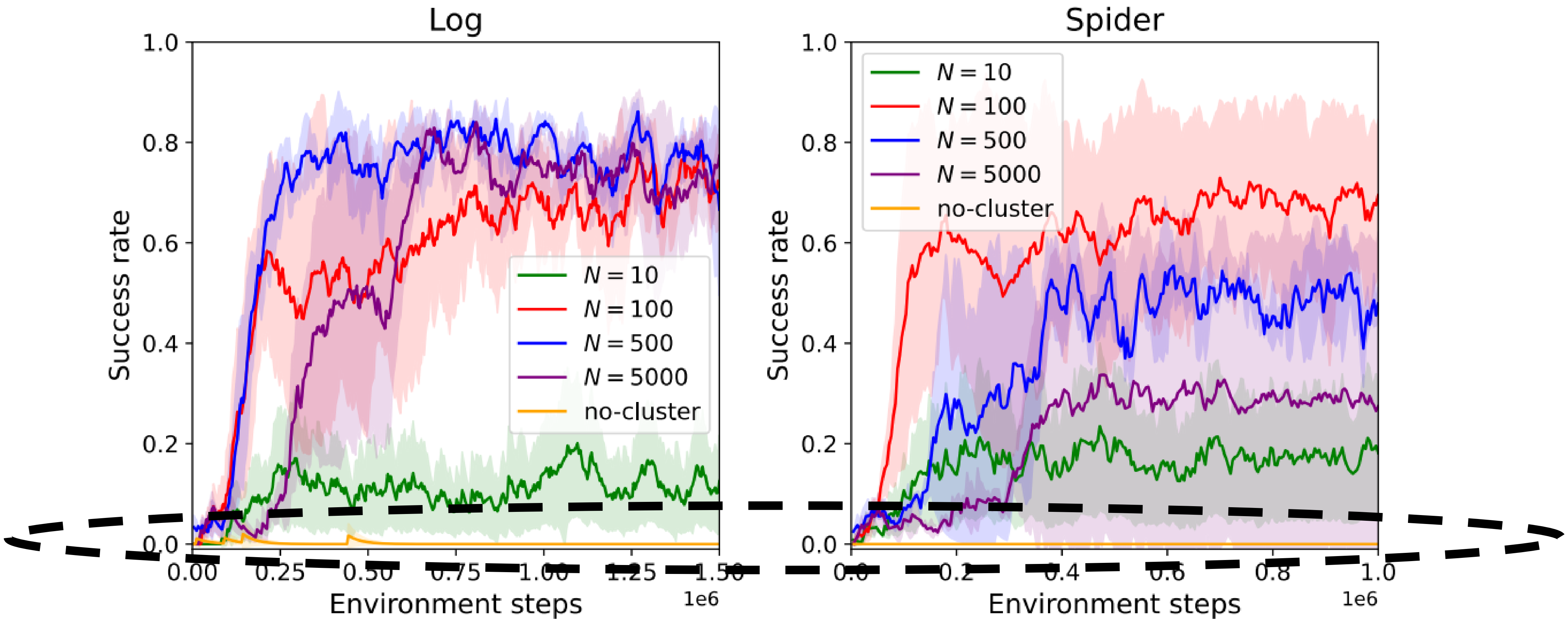




Goal Clustering

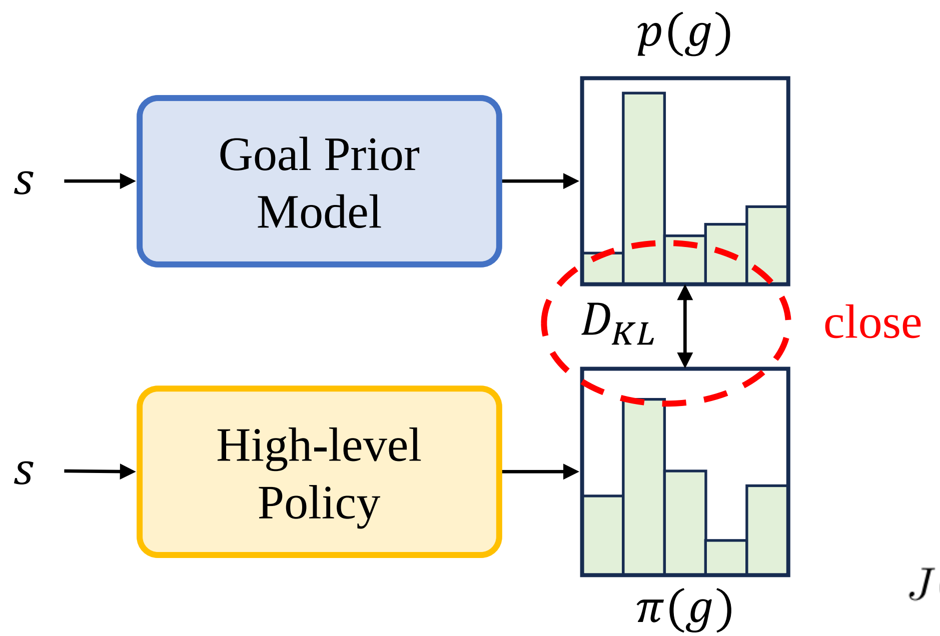


Goal Clustering

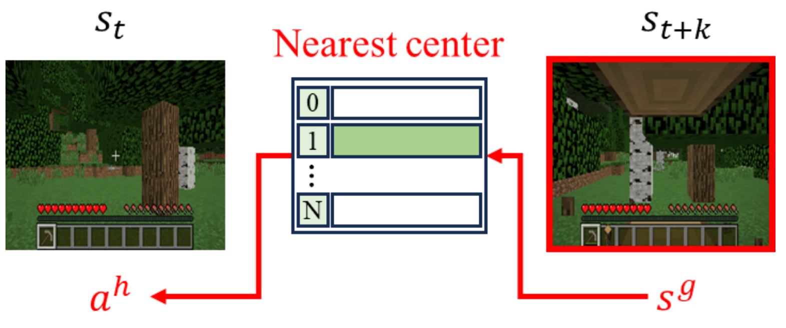


Goal Prior Model

- We have not developed a prior for the high-level RL policy: “how to select the goal”.



- Pre-training:



$$\mathcal{L}(\psi) = \mathbb{E}_D \left[-\log \pi_\psi^p (a^h | s_t) \right]$$

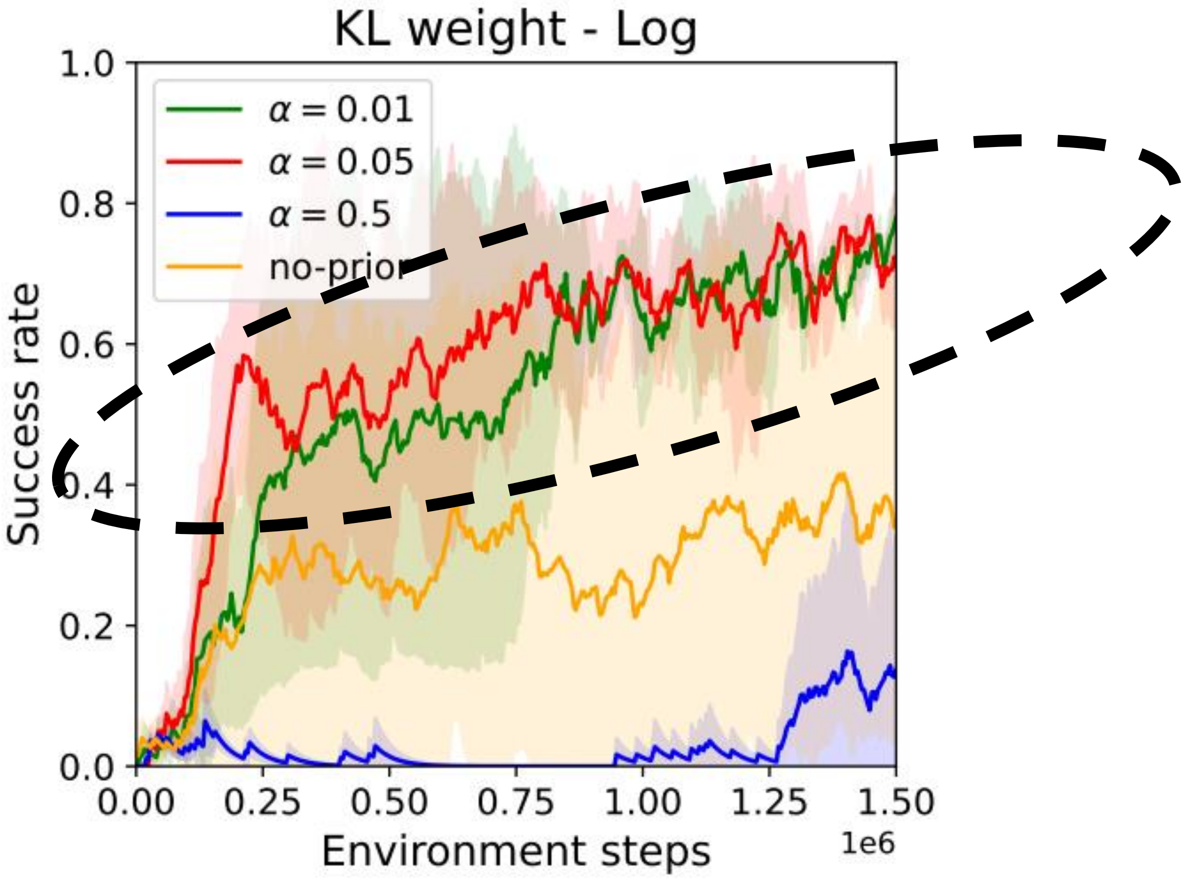
- RL: Policy regularization

$$J(\theta) = \mathbb{E}\pi_\theta \left[\sum_{t=0}^{\infty} \gamma^t \left(\sum_{i=kt}^{(k+1)t} R(s_i, a_i) \boxed{- \alpha D_{\mathrm{KL}} \left(\pi_\psi^p (a^h | s_{kt}) \| \pi_\theta (a^h | s_{kt}) \right)} \right) \right]$$

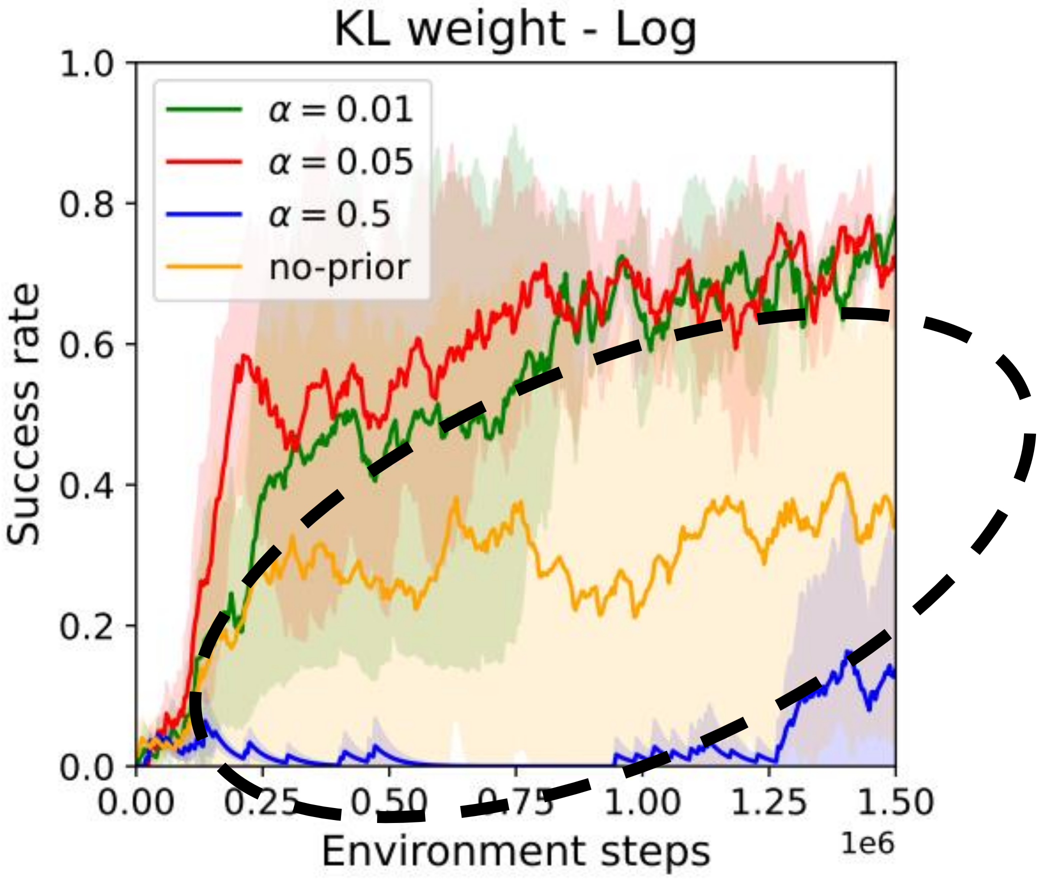
Task reward

Goal prior reward

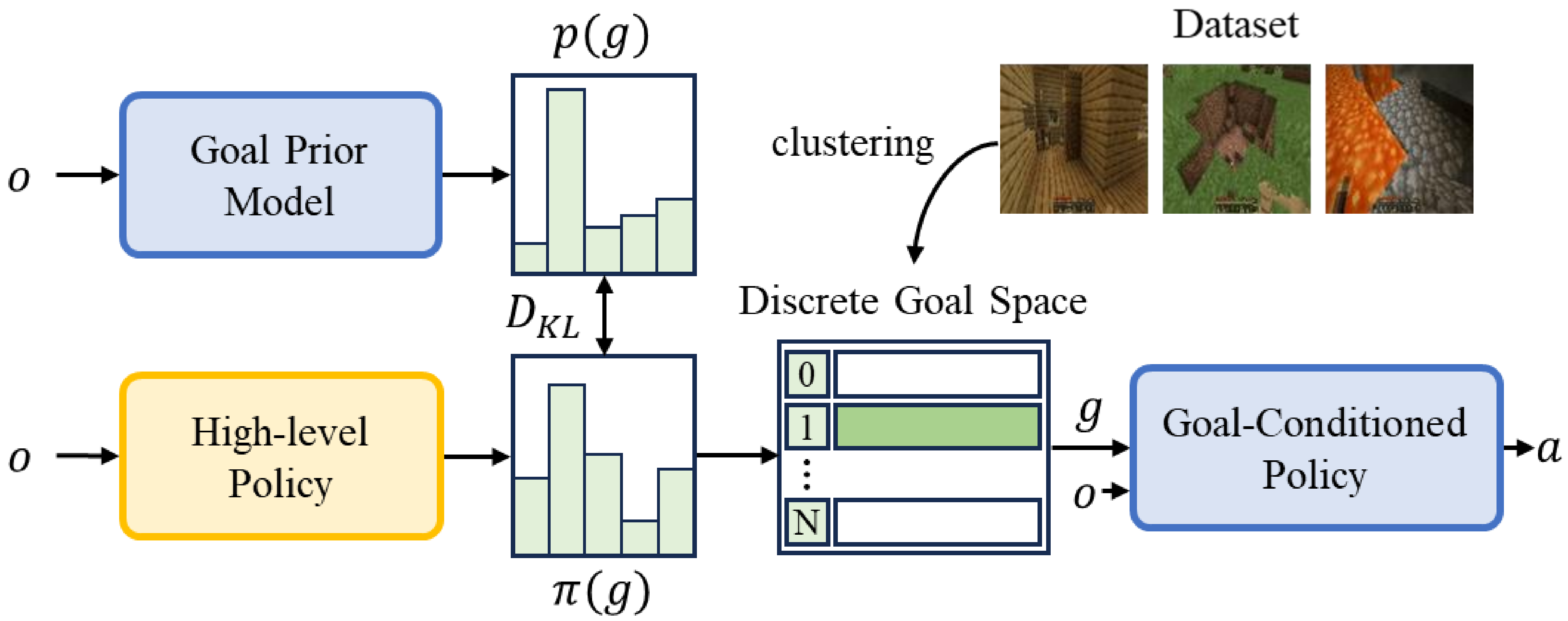
Goal Prior Model



Goal Prior Model



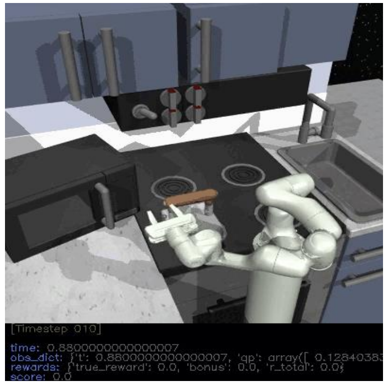
Summary



$$J(\theta) = \mathbb{E}\pi_\theta\left[\sum_{t=0}^{\infty}\gamma^t\left(\sum_{i=kt}^{(k+1)t} R(s_i, a_i) - \alpha D_{\mathrm{KL}}\left(\pi_\psi^p(a^h|s_{kt})\|\pi_\theta(a^h|s_{kt})\right)\right)\right]$$

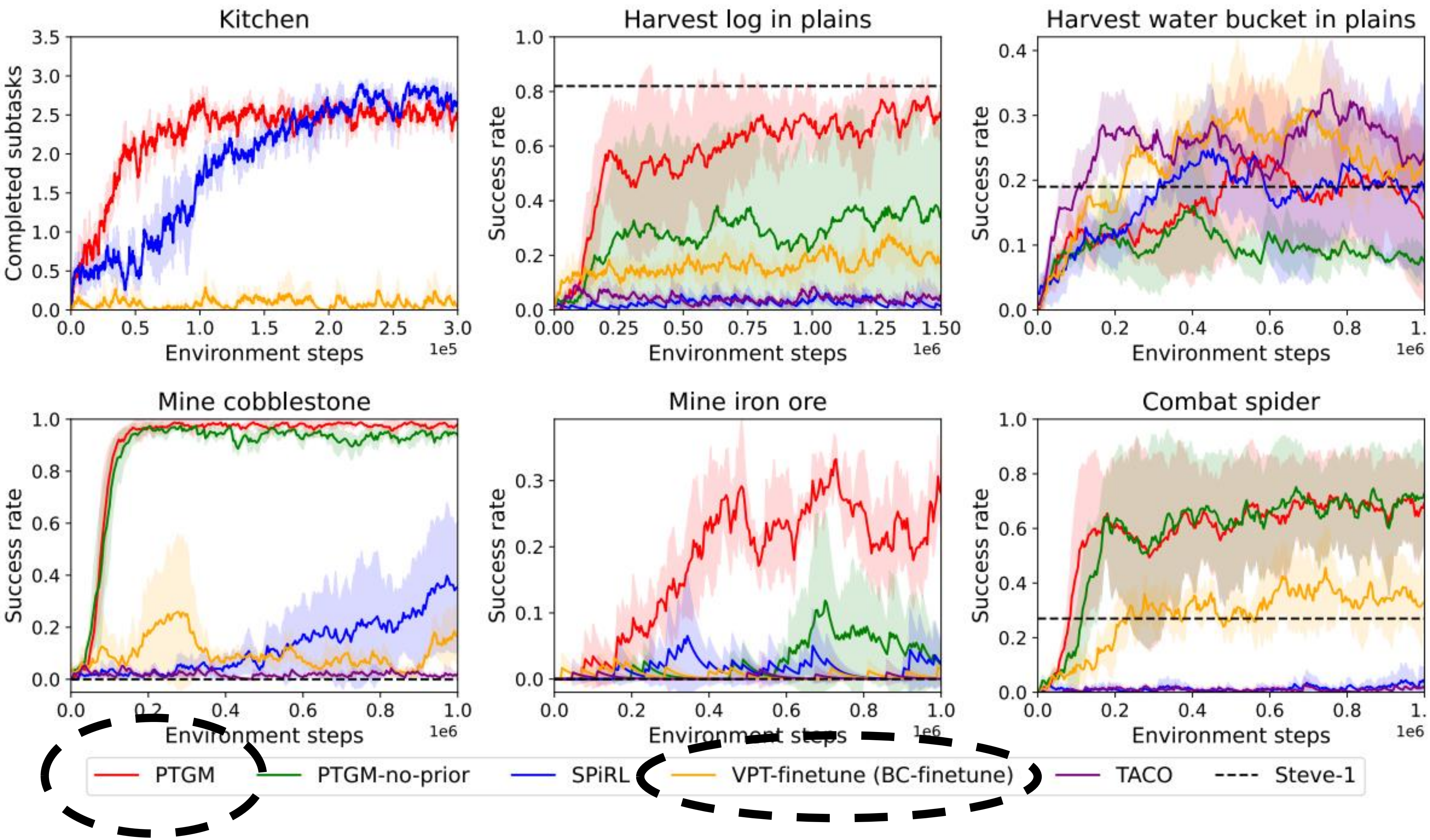
Playing Minecraft

- >10K combinatorial action space; 30 fps control; long-horizon tasks with 2K steps.
- 39M human gameplay dataset (Baker et al., 2022).






Playing Minecraft



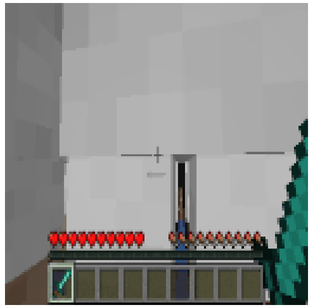
Capacity of the Discrete Goal Space

- The clustering approach may discard some useful goals.
- Why is the discrete goal space still capable of completing diverse tasks?

Capacity of the Discrete Goal Space

Goal **Scenario:** Sheep Pig Chicken

Behavior:



Test task	Sheep	Pig	Chicken
Success rate	0.82	0.36	0.94

Test task	Place	Water	Wool
Success rate	0.65	0.16	0.44

Capacity of the Discrete Goal Space

A single goal can lead to varied behaviors conditioned on different states.

Conclusion

- PTGM is a goal-based approach for skill pre-training in RL, overcoming the **two weaknesses** of previous approaches.
- Advantages in the **sample efficiency, learning stability, interpretability, and generalization** of the low-level skills.
- Promising results in different domains including the challenging **Minecraft** tasks.

Thank you!

https://sites.google.com/view/ptgm-iclr/

Poster Session 4, 16:30~18:30