DMQA Open-Seminar

Hierarchical Learning for Offline Goal-Conditioned RL

Korea University
Data Mining & Quality Analytics Lab.

Jang SeongIn

0. 발표자 소개

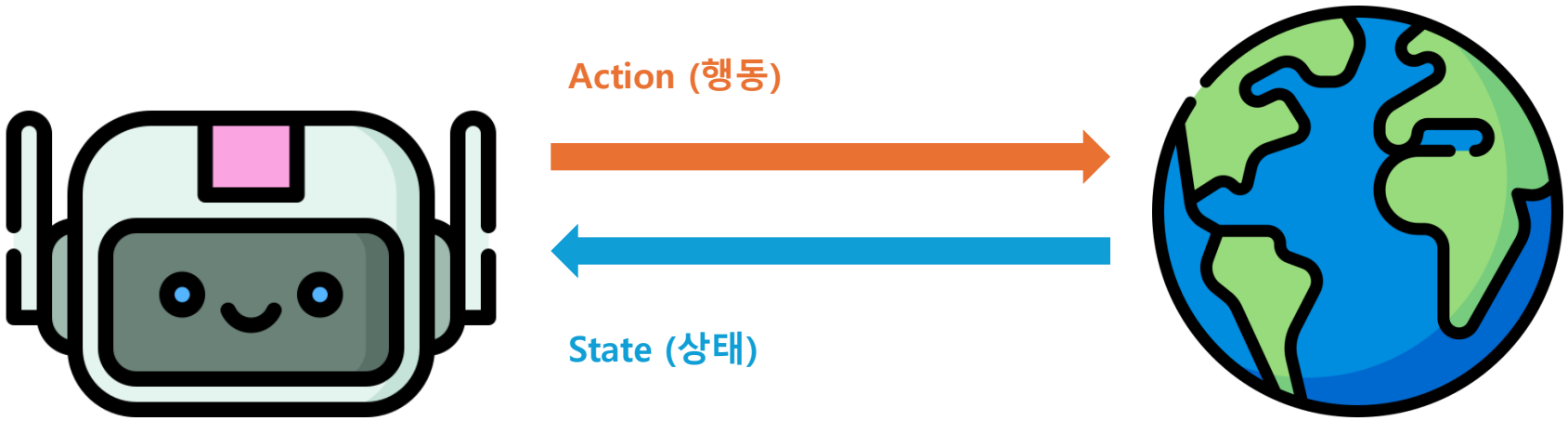
- **장성인 (Seongin Jang)**
  - 고려대학교 산업경영공학과 석사과정(26.03~Present)
  - Data Mining & Quality Analytics Lab (김성범 교수님)
- **관심 연구 분야**
  - Data Analysis
  - Reinforcement Learning
  - VLA model
- **E-mail**
  - si_jang@korea.ac.kr

1. Introduction

Reinforcement Learning

❖ **Reinforcement Learning**

- **Agent**가 **environment**와 상호작용하여 받는 **reward**가 최대로 될 수 있도록 **최적의 policy**를 학습하는 과정



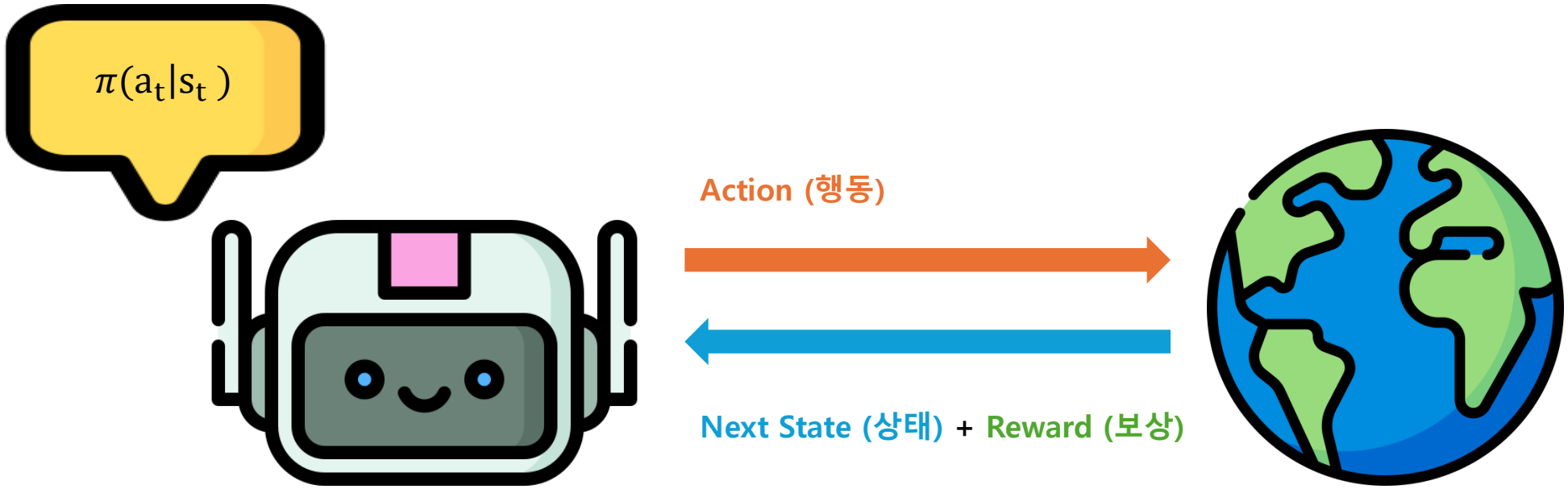
<Agent> **<Environment>**

1. Introduction

Reinforcement Learning

❖ **Reinforcement Learning**

- **Agent**가 **environment**와 상호작용하여 받는 **reward**가 최대로 될 수 있도록 **최적의 policy**를 학습하는 과정



1. Introduction

Reinforcement Learning

❖ **Reinforcement Learning**

- **Agent**가 **environment**와 상호작용하여 받는 **reward**가 최대로 될 수 있도록 **최적의 policy**를 학습하는 과정

1. Introduction

Goal-Conditioned Reinforcement Learning

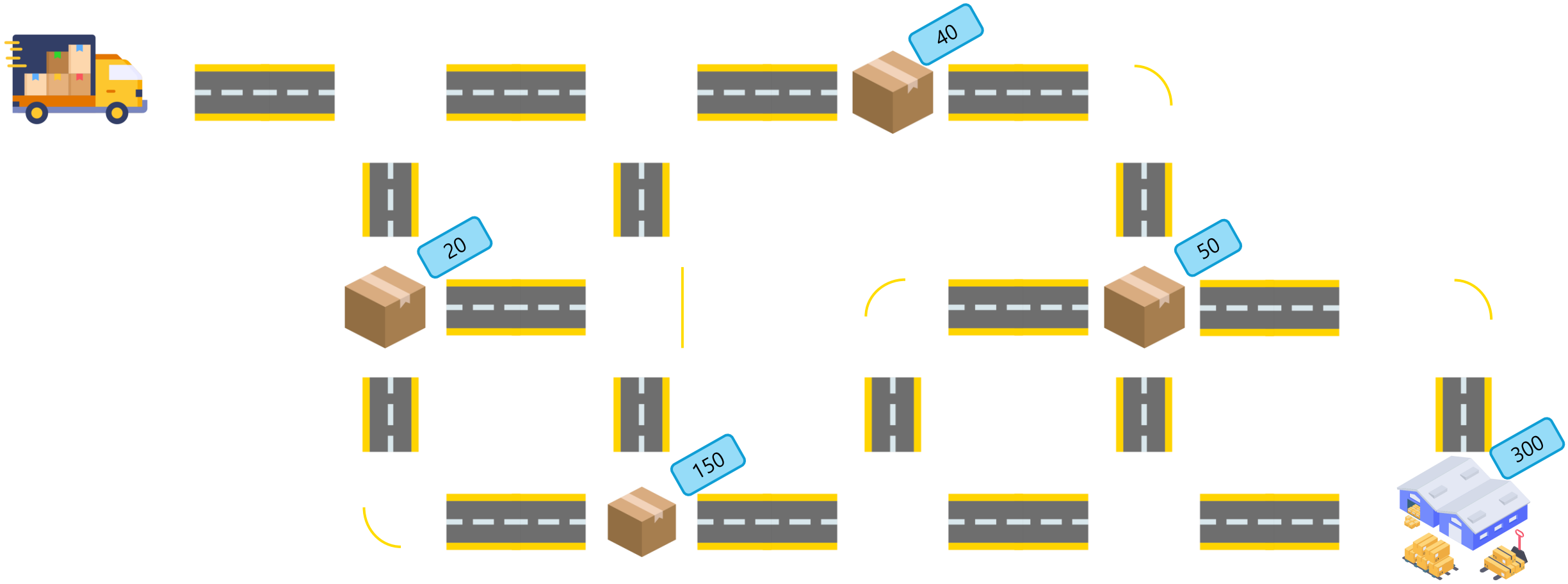
❖ **Reinforcement Learning + Goal → Goal Conditioned RL(GCRL)**

- **Goal Conditioned RL(GCRL)** : **목표(Goal)**를 추가로 입력 받아 달성할 수 있는 **최적의 policy**를 학습하는 강화학습

1. Introduction

Goal-Conditioned Reinforcement Learning

❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

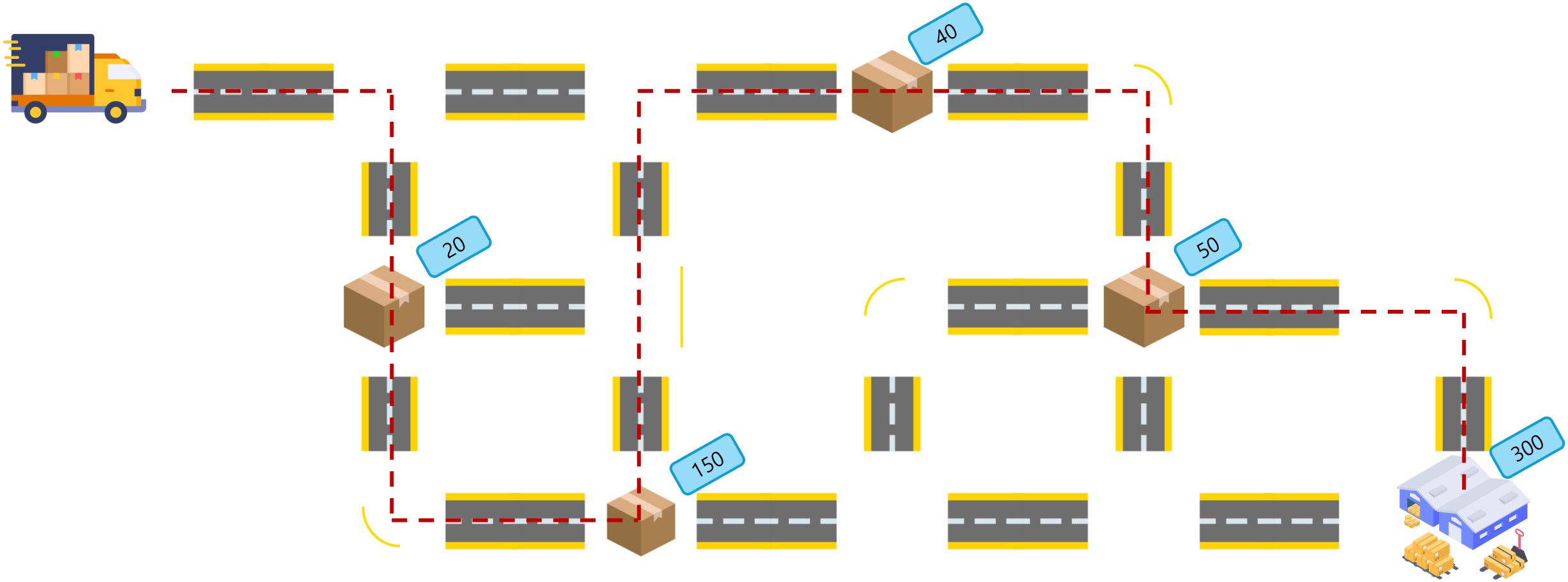


1. Introduction

Goal-Conditioned Reinforcement Learning

❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

- **RL**은 **reward를 최대화**하는 방향으로 움직임

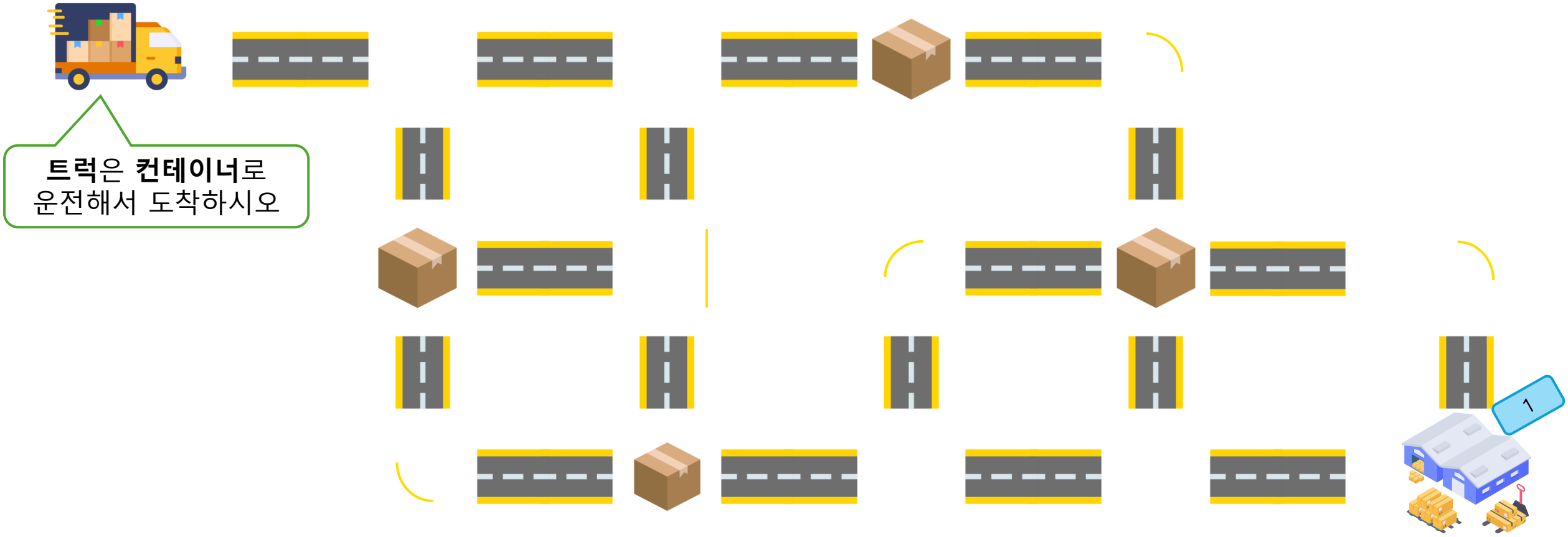


1. Introduction

Goal-Conditioned Reinforcement Learning

❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

- **RL**은 **reward를 최대화**하는 방향으로 움직임
- **GCRL**은 중간에 상자가 있어도 **목표 g에 도달했는지에 대한 여부**만 확인

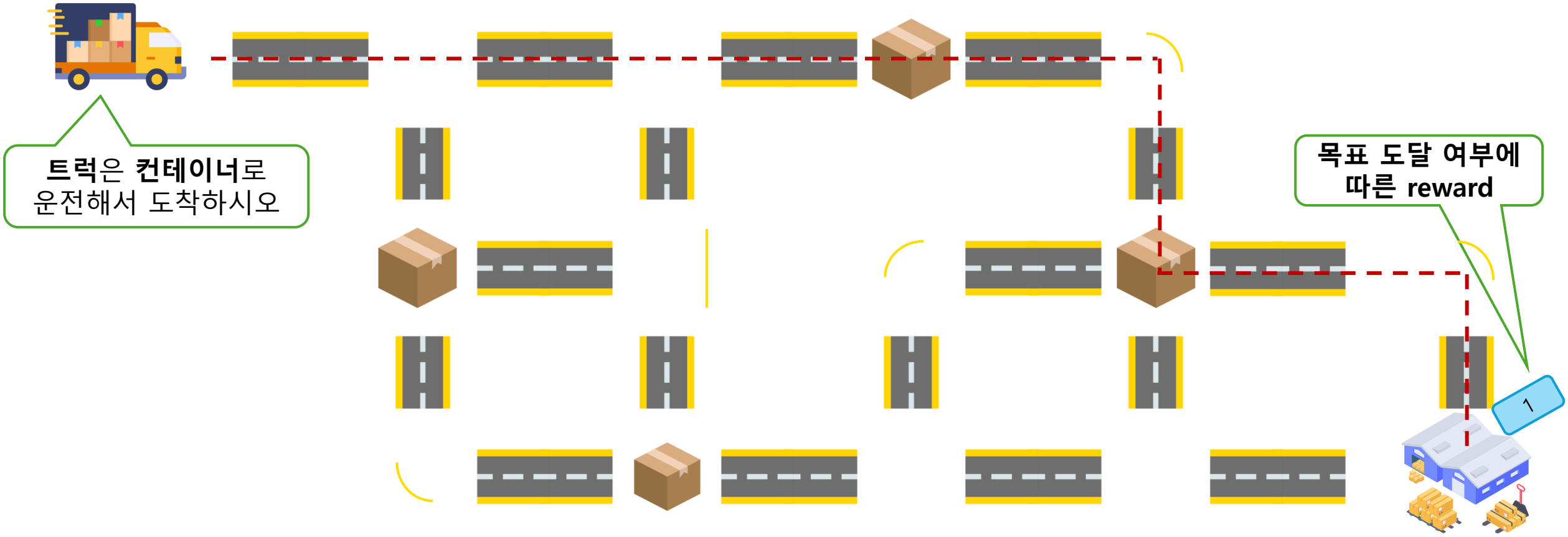


1. Introduction

Goal-Conditioned Reinforcement Learning

❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

- **RL**은 **reward를 최대화**하는 방향으로 움직임
- **GCRL**은 중간에 상자가 있어도 **목표 g에 도달했는지에 대한 여부**만 확인



1. Introduction

Goal-Conditioned Reinforcement Learning

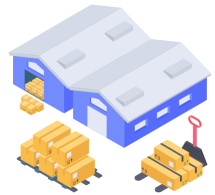
❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

- **RL**은 **reward를 최대화**하는 방향으로 움직임
- **GCRL**은 중간에 상자가 있어도 **목표 g에 도달했는지에 대한 여부**만 확인

RL : reward 최대

GCRL : goal에 도달했는지

Episode마다 다른 goal 설정 가능

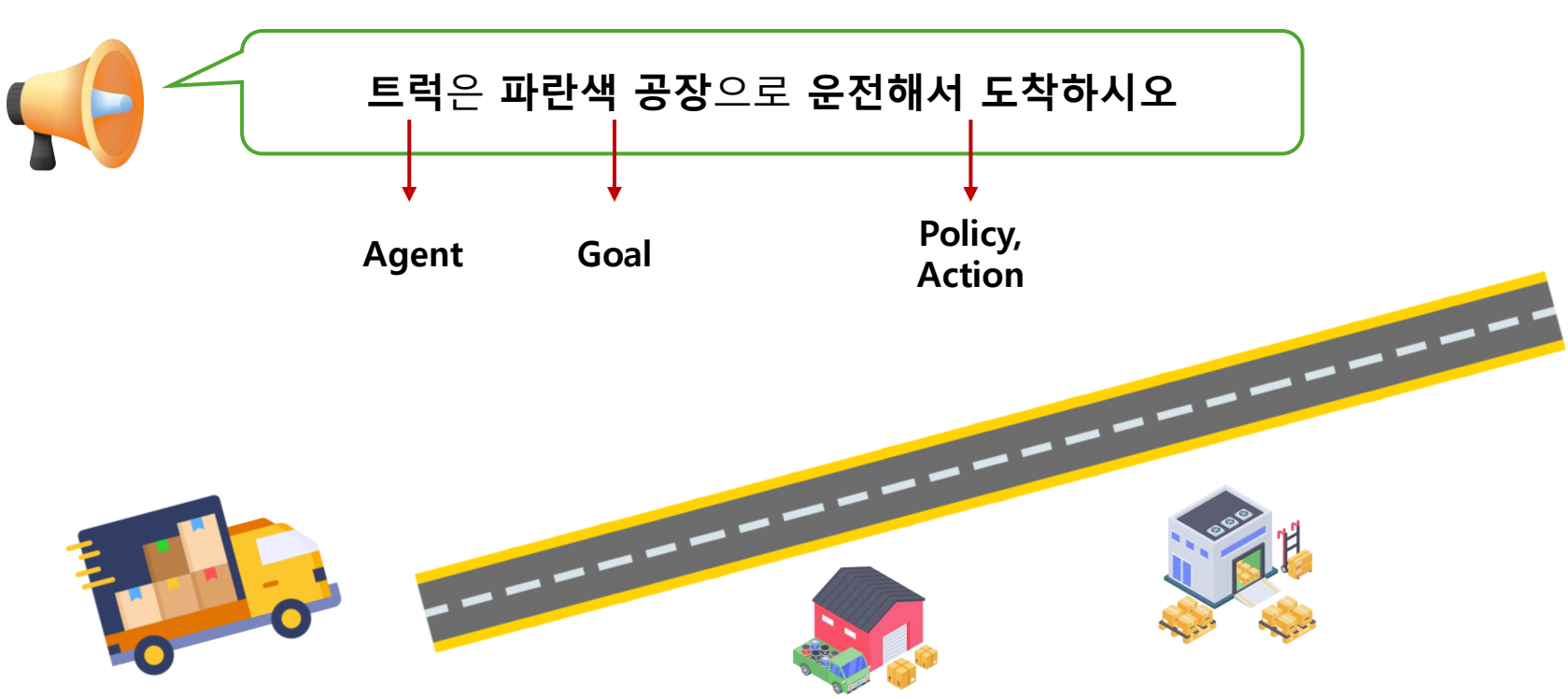


1. Introduction

Goal-Conditioned Reinforcement Learning

❖ **RL과 GCRL은 어떤 차이가 있는 것일까?**

- **RL**은 **reward를 최대화**하는 방향으로 움직임
- **GCRL**은 중간에 상자가 있어도 **목표 g에 도달했는지에 대한 여부**만 확인

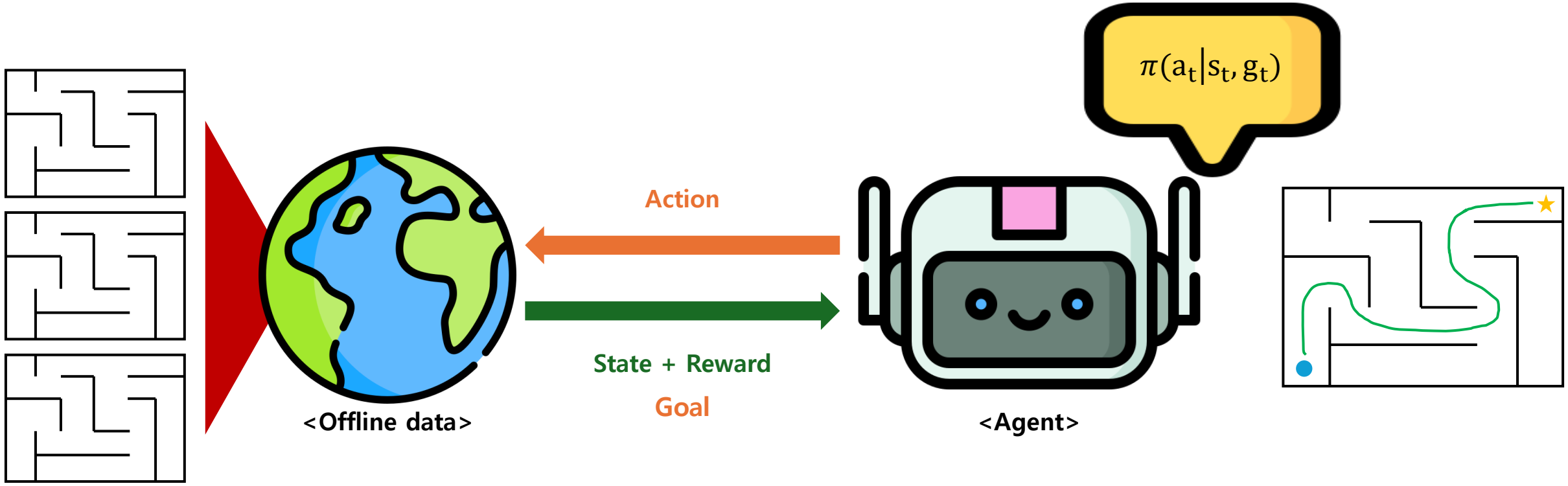


1. Introduction

Offline Goal-Conditioned Reinforcement Learning

❖ **Offline GCRL**

- **Online으로 학습 시 상호작용에 대한 비용과 안전성 문제 발생 가능성**
- **미리 수집된 데이터셋**을 통해 environment와 추가 상호작용 없이 임의의 goal에 달성할 수 있는 policy를 학습



2. Related Works

❖ **HIQL : Offline Goal-Conditioned RL with Latent States as Actions(NeurIPS 2023)**

- **Hierarchical RL**을 **offline GCRL**에 도입
- **High-level policy** → subgoal 제시, **Low-level policy** → subgoal에 도달하기 위한 action 수행

HIQL: Offline Goal-Conditioned RL with Latent States as Actions

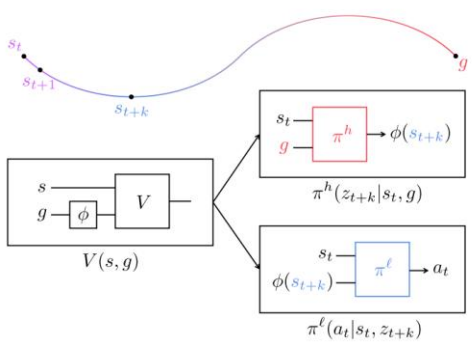
Seohong Park[1] **Dibya Ghosh**[1] **Benjamin Eysenbach**[2] **Sergey Levine**[1]
[1]University of California, Berkeley [2]Princeton University
seohong@berkeley.edu

Abstract

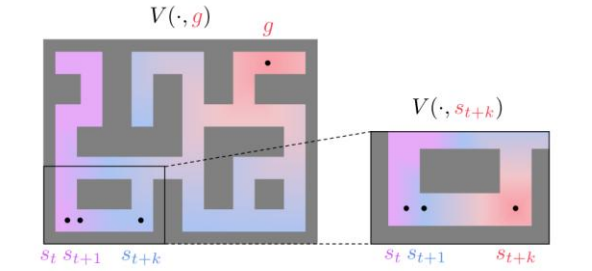
Unsupervised pre-training has recently become the bedrock for computer vision and natural language processing. In reinforcement learning (RL), goal-conditioned RL can potentially provide an analogous self-supervised approach for making use of large quantities of unlabeled (reward-free) data. However, building effective algorithms for goal-conditioned RL that can learn directly from diverse offline data is challenging, because it is hard to accurately estimate the exact value function for faraway goals. Nonetheless, goal-reaching problems exhibit structure, such that reaching distant goals entails first passing through closer subgoals. This structure can be very useful, as assessing the quality of actions for nearby goals is typically easier than for more distant goals. Based on this idea, we propose a hierarchical algorithm for goal-conditioned RL from offline data. Using one action-free value function, we learn two policies that allow us to exploit this structure: a high-level policy that treats states as actions and predicts (a latent representation of) a subgoal and a low-level policy that predicts the action for reaching this subgoal. Through analysis and didactic examples, we show how this hierarchical decomposition makes our method robust to noise in the estimated value function. We then apply our method to offline goal-reaching benchmarks, showing that our method can solve long-horizon tasks that stymie prior methods, can scale to high-dimensional image observations, and can readily make use of action-free data. Our code is available at https://seohong.me/projects/hiql/

1 Introduction

Many of the most successful machine learning systems for computer vision [15, 36] and natural language processing [10, 18] leverage large amounts of unlabeled or weakly-labeled data. In the reinforcement learning (RL) setting, offline goal-conditioned RL provides an analogous way to potentially leverage large amounts of multi-task data without reward labels or video data without action labels: offline learning [54, 55] enables leveraging previously collected and passively observed data, and goal-conditioned RL [44, 79] enables learning from unlabeled, reward-free data. However, offline goal-conditioned RL poses major challenges. First, learning an accurate goal-conditioned value function for any state and goal pair is challenging when considering very broad and long-horizon goal-reaching tasks. This often results in a noisy value function and thus potentially an erroneous policy. Second, while the offline setting unlocks the potential for using previously collected data, it



(a) **Three components of HIQL.**



Flat policy
π : max $V(s_{t+1},g) - V(s_t,g)$

Hierarchical policy
π^h : max $V(s_{t+k},g) - V(s_t,g)$
π^ℓ : max $V(s_{t+1},s_{t+k}) - V(s_t,s_{t+k})$

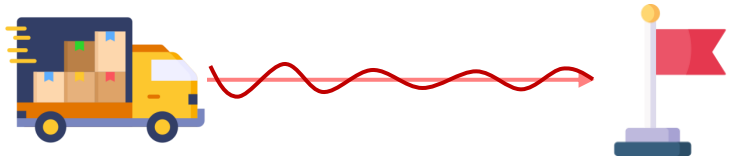
(b) **Hierarchical policies get clearer learning signals.**

2. Related Works

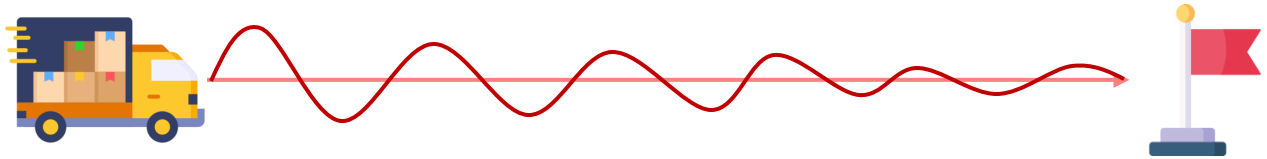
HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **Offline GCRL이 가지고 있는 문제점이 무엇일까?**

- **가까운 목표**는 정확히 추정 가능하지만, **먼 목표**일수록 **value function에 노이즈가 커짐**
- Offline에서는 **상호작용을 통한 교정이 불가능해 노이즈가 누적됨**



노이즈 영향이 적어 비교적 정확한 경로 학습



노이즈가 누적되어 실제 목표에서 크게 벗어난 경로 학습

Hierarchical policy 추출

2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **Hierarchical Policy**

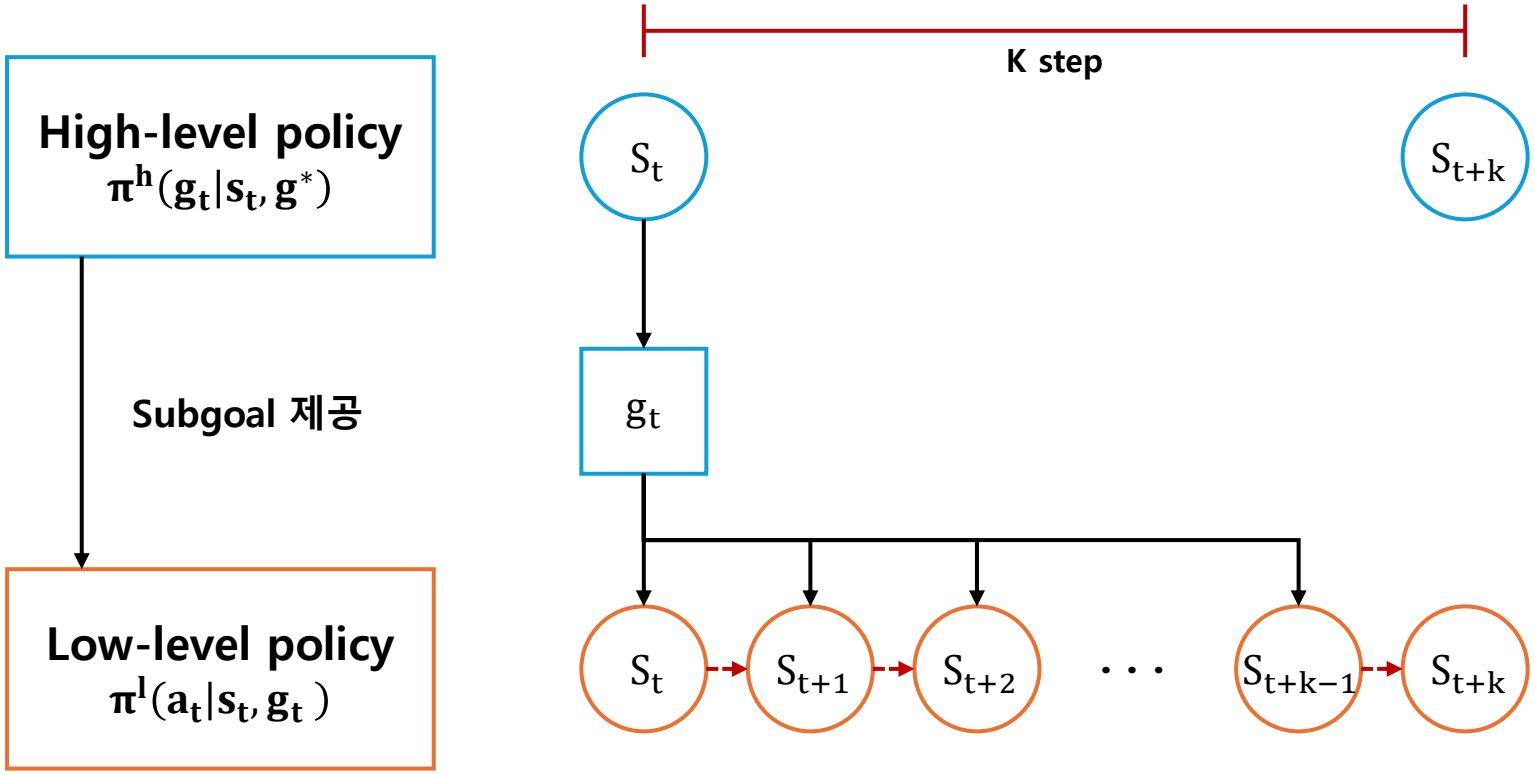
- Hierarchical policy는 **high/low-level policy 2가지로 분리**
- **High-level policy**는 **subgoal 예측**, **Low-level policy**는 subgoal까지의 **action 예측**

2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **Hierarchical Policy**

- Hierarchical policy는 **high/low-level policy 2가지로 분리**
- **High-level policy**는 **subgoal 예측**, **Low-level policy**는 subgoal까지의 **action 예측**

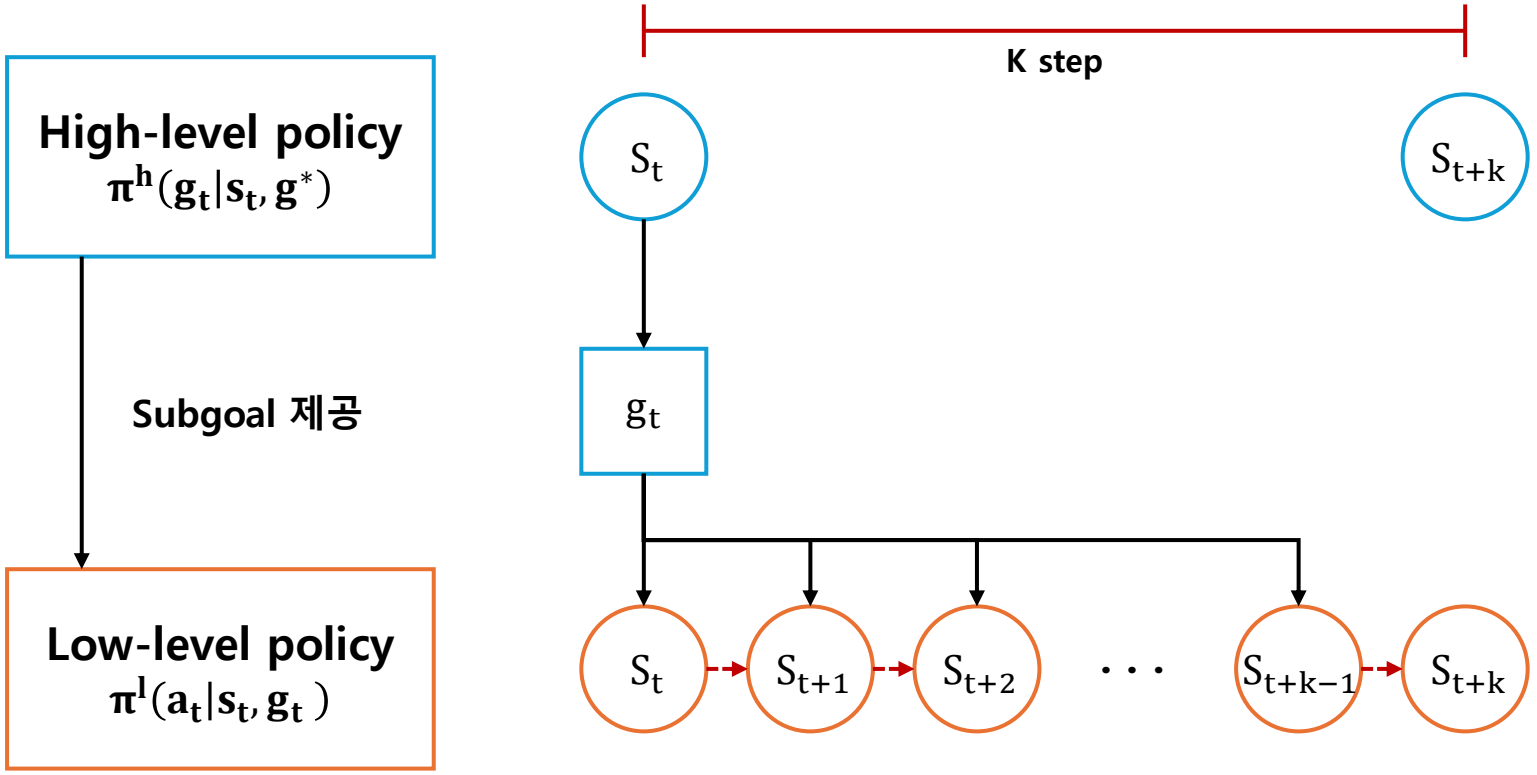


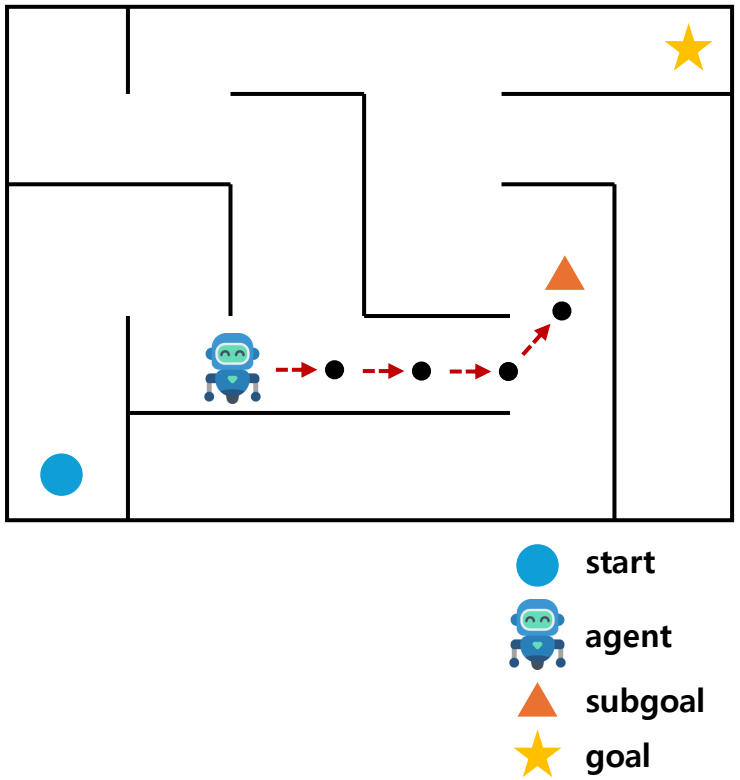
2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **Hierarchical Policy**

- Hierarchical policy는 **high/low-level policy 2가지로 분리**
- **High-level policy**는 **subgoal 예측**, **Low-level policy**는 subgoal까지의 **action 예측**



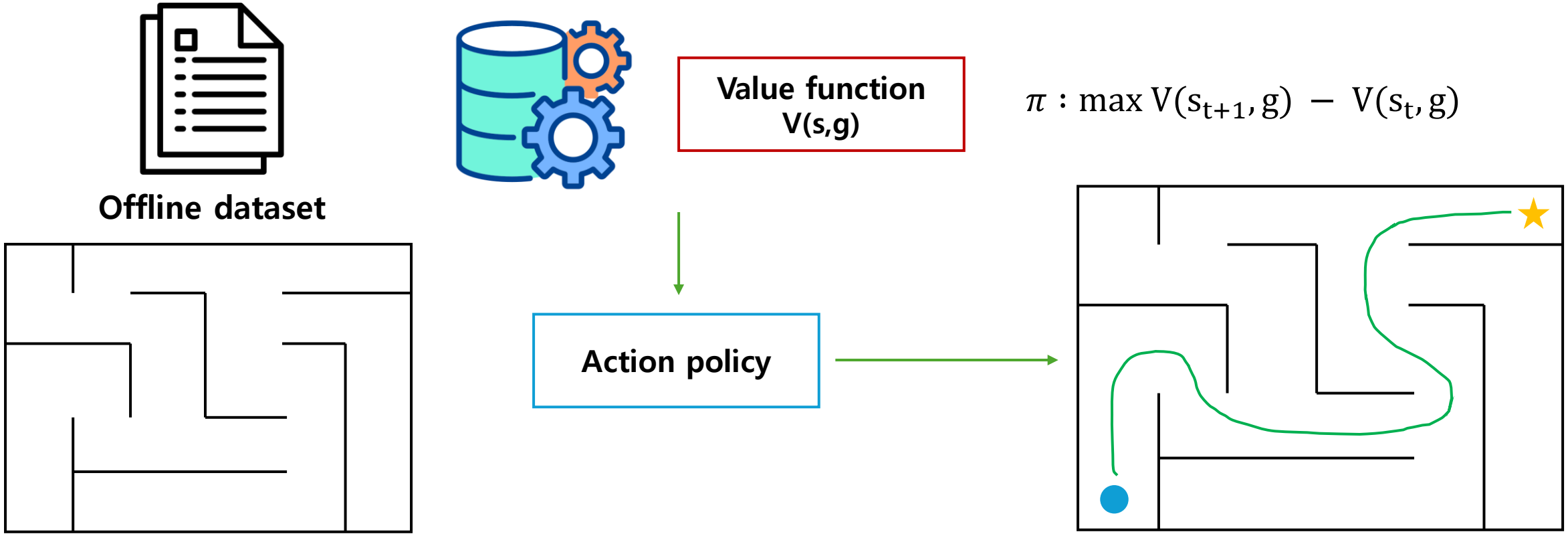


2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **HIQL**

- 기존 Offline GCRL : dataset으로부터 **value를 학습** → 주어진 상황에 **학습된 action**을 통해 **goal 도달**



2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

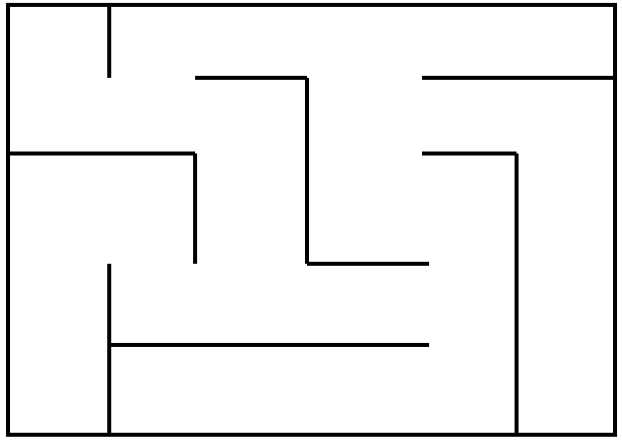
❖ **HIQL**

- 기존 Offline GCRL : dataset으로부터 value를 학습 → 주어진 상황에 학습된 action을 통해 goal 도달
- 방법론 : value function으로부터 **high/low-level policy** 학습

Offline dataset

Value function V(s,g)

$$\pi : \max V(s_{t+1}, g) - V(s_t, g)$$



High-level policy

$$\pi^h : \max V(s_{t+k}, g) - V(s_t, g)$$

Low-level policy

$$\pi^l : \max V(s_{t+1}, s_{t+k}) - V(s_t, s_{t+k})$$

2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

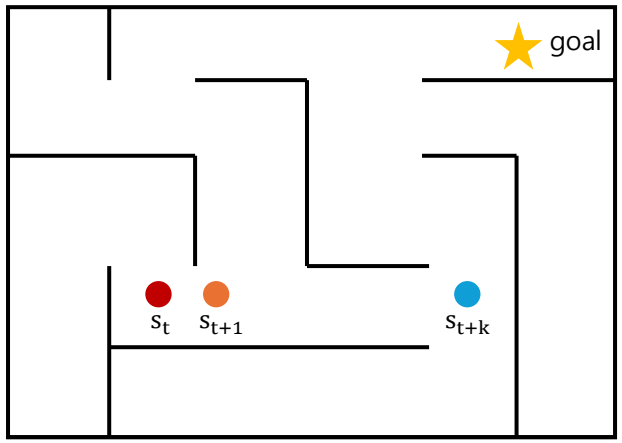
❖ **HIQL**

- 기존 Offline GCRL : dataset으로부터 value를 학습 → 주어진 상황에 학습된 action을 통해 goal 도달
- 방법론 : value function으로부터 **high/low-level policy** 학습

Offline dataset

Value function V(s,g)

$$\pi : \max V(s_{t+1}, g) - V(s_t, g)$$



High-level policy

$$\pi^h : \max V(s_{t+k}, g) - V(s_t, g)$$

Low-level policy

$$\pi^l : \max V(s_{t+1}, s_{t+k}) - V(s_t, s_{t+k})$$

2. Related Works

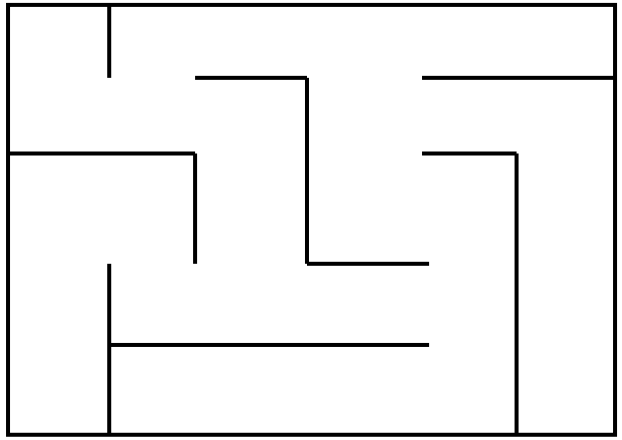
HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **HIQL**

- 기존 Offline GCRL : dataset으로부터 value를 학습 → 주어진 상황에 학습된 action을 통해 goal 도달
- 방법론 : value function으로부터 **high/low-level policy** 학습

Offline dataset

Value function
V(s,g)

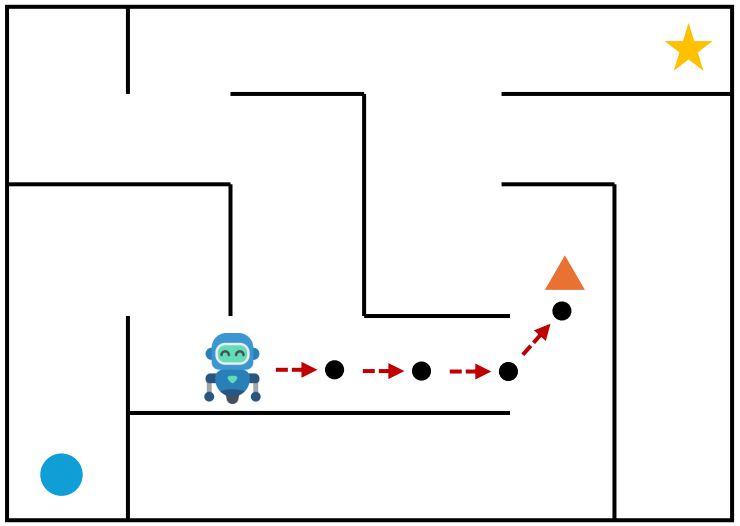


High-level policy

Subgoal

Low-level policy

Action for subgoal



2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ 실험 결과 : HIQL의 성능은 향상되었을까?

- State-based 3개, Pixel-based 3개로 총 6개 벤치마크에서 실험 진행
- HIQL은 기존 방법론들을 능가하며, long-horizon 과제에서 뛰어난 성능을 보여줌

Dataset	GCBC	HGCBC	GC-IQL	GC-POR	TAP	TT	HIQL (ours)	HIQL (w/o repr.)
antmaze-medium-diverse	67.3 ±10.1	71.6 ±8.9	63.5 ±14.6	74.8 ±11.9	85.0	**100.0**	86.8 ±4.6	89.9 ±3.5
antmaze-medium-play	71.9 ±16.2	66.3 ±9.2	70.9 ±11.2	71.4 ±10.9	78.0	**93.3**	84.1 ±10.8	87.0 ±8.4
antmaze-large-diverse	20.2 ±9.1	63.9 ±10.4	50.7 ±18.8	49.0 ±17.2	82.0	60.0	**88.2** ±5.3	87.3 ±3.7
antmaze-large-play	23.1 ±15.6	64.7 ±14.5	56.5 ±14.4	63.2 ±16.1	74.0	66.7	**86.1** ±7.5	81.2 ±6.6
antmaze-ultra-diverse	14.4 ±9.7	39.4 ±20.6	21.6 ±15.2	29.8 ±13.6	26.0	33.3	**52.9** ±17.4	52.6 ±8.7
antmaze-ultra-play	20.7 ±9.7	38.2 ±18.1	29.8 ±12.4	31.0 ±19.4	22.0	20.0	39.2 ±14.8	**56.0** ±12.4
kitchen-partial	38.5 ±11.8	32.0 ±16.7	39.2 ±13.5	18.4 ±14.3	-	-	**65.0** ±9.2	46.3 ±8.6
kitchen-mixed	46.7 ±20.1	46.8 ±17.6	51.3 ±12.8	27.9 ±17.9	-	-	**67.7** ±6.8	36.8 ±20.1
calvin	17.3 ±14.8	3.1 ±8.8	7.8 ±17.6	12.4 ±18.6	-	-	**43.8** ±39.5	23.4 ±27.1

Dataset	GCBC	HGCBC (+ repr.)	GC-IQL	GC-POR (+ repr.)	HIQL (ours)
procgen-maze-500-train	16.8 ±2.8	14.3 ±4.1	72.5 ±10.0	75.8 ±12.1	**82.5** ±6.0
procgen-maze-500-test	14.5 ±5.0	11.2 ±3.7	49.5 ±9.8	53.8 ±14.5	**64.5** ±13.2
procgen-maze-1000-train	27.2 ±8.9	15.0 ±5.7	78.2 ±7.2	82.0 ±6.5	**87.0** ±13.9
procgen-maze-1000-test	12.0 ±5.9	14.5 ±5.0	60.0 ±10.6	69.8 ±7.4	**78.2** ±17.9
visual-antmaze-diverse	71.4 ±6.0	35.1 ±12.0	72.6 ±5.9	47.4 ±17.6	**80.5** ±9.4
visual-antmaze-play	64.4 ±6.3	23.8 ±8.5	70.4 ±26.6	57.0 ±8.1	**78.4** ±4.6
visual-antmaze-navigate	33.2 ±7.9	21.4 ±4.6	22.1 ±14.1	16.1 ±15.2	**45.7** ±18.1
roboverse	26.2 ±4.5	26.4 ±6.4	31.2 ±8.7	46.6 ±7.4	**61.5** ±5.3

2. Related Works

HIQL : Offline Goal-Conditioned RL with Latent States as Actions

❖ **실험 결과 : Offline data에 레이블이 부족하더라도 성능이 뛰어날까?**

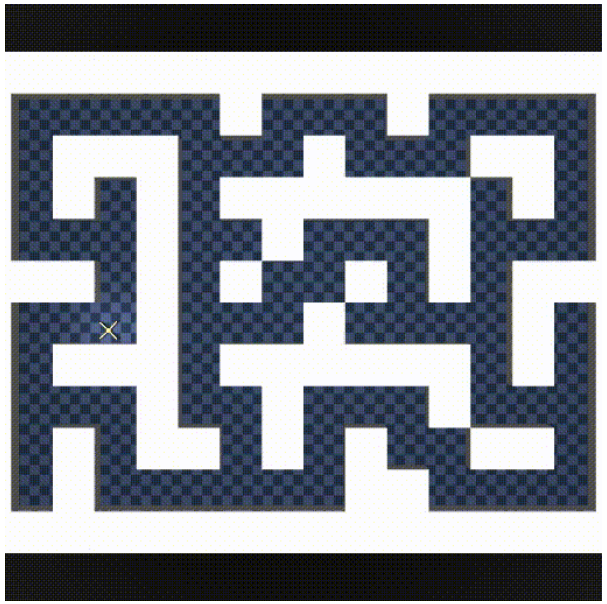
- Offline data 중 일부에만 액션 레이블이 있고 나머지는 상태 전이만 존재
- 일부만 제공되어도 성능이 높으며 다른 알고리즘보다 더 높은 성능을 기록

Dataset	GC-IQL (full)	GC-POR (full)	HIQL (full)	HIQL (action-limited)	vs. HIQL (full)	vs. Prev. best (full)
antmaze-large-diverse	50.7 ±18.8	49.0 ±17.2	88.2 ±5.3	88.9 ±6.4	+0.7	+38.2
antmaze-ultra-diverse	21.6 ±15.2	29.8 ±13.6	52.9 ±17.4	38.2 ±15.4	−14.7	+8.4
kitchen-mixed	51.3 ±12.8	27.9 ±17.9	67.7 ±6.8	59.1 ±9.6	−8.6	+7.8
calvin	7.8 ±17.6	12.4 ±18.6	43.8 ±39.5	35.8 ±30.7	−8.0	+23.4
procgen-maze-500-train	72.5 ±10.0	75.8 ±12.1	82.5 ±6.0	77.0 ±12.5	−5.5	+1.2
procgen-maze-500-test	49.5 ±9.8	53.8 ±14.5	64.5 ±13.2	65.5 ±16.4	+1.0	+11.7

2. Related Works

❖ **OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL(ICLR 2025)**

- Offline GCRL 알고리즘의 체계적인 평가를 위한 새로운 벤치마크



OGBench

2. Related Works

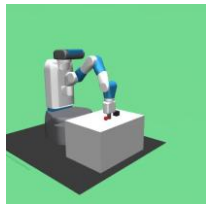
OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **새로운 벤치마크가 필요한 이유가 무엇일까?**

- 기존: Offline GCRL 알고리즘 능력 평가 벤치마크 부족, 재현성과 공정한 비교를 위한 환경 미흡

Offline GCRL 알고리즘

D4RL



Fetch

2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

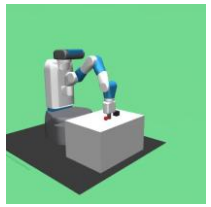
❖ **새로운 벤치마크가 필요한 이유가 무엇일까?**

- 기존: Offline GCRL 알고리즘 능력 평가 벤치마크 부족, 재현성과 공정한 비교를 위한 환경 미흡

Offline GCRL 알고리즘

D4RL

단일 목표 중심의 평가로 인한 다중 작업 능력 검증의 부재



Fetch

단순하고 단일 행동 위주의 과제 난이도

2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

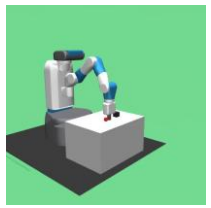
❖ **새로운 벤치마크가 필요한 이유가 무엇일까?**

- 기존: Offline GCRL 알고리즘 능력 평가 **벤치마크 부족, 재현성과 공정한 비교를 위한 환경 미흡**

Offline GCRL 알고리즘

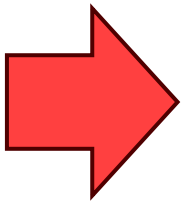
D4RL

단일 목표 중심의 평가로 인한 다중 작업 능력 검증의 부재



Fetch

단순하고 단일 행동 위주의 과제 난이도



New benchmark

2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **새로운 벤치마크가 필요한 이유가 무엇일까?**

- 기존: Offline GCRL 알고리즘 능력 평가 **벤치마크 부족, 재현성과 공정한 비교를 위한 환경 미흡**
- **OGBench**: 다양한 행동을 평가할 수 있는 **offline GCRL 체계적 평가 벤치마크**

Offline GCRL 알고리즘

NEW **OGBench**

2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **새로운 벤치마크가 필요한 이유가 무엇일까?**

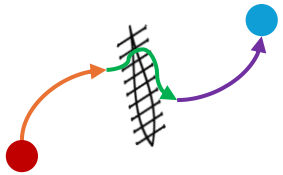
- 기존: Offline GCRL 알고리즘 능력 평가 **벤치마크 부족, 재현성과 공정한 비교를 위한 환경 미흡**
- **OGBench**: 다양한 행동을 평가할 수 있는 **offline GCRL 체계적 평가 벤치마크**

Offline GCRL 알고리즘

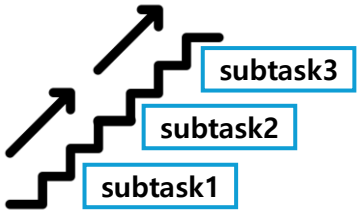
NEW **OGBench**

Challenges

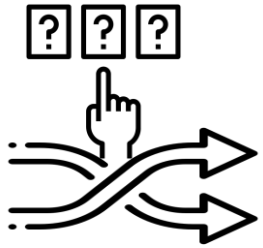
Suboptimal, unstructured data



Goal stitching



Long-horizon reasoning

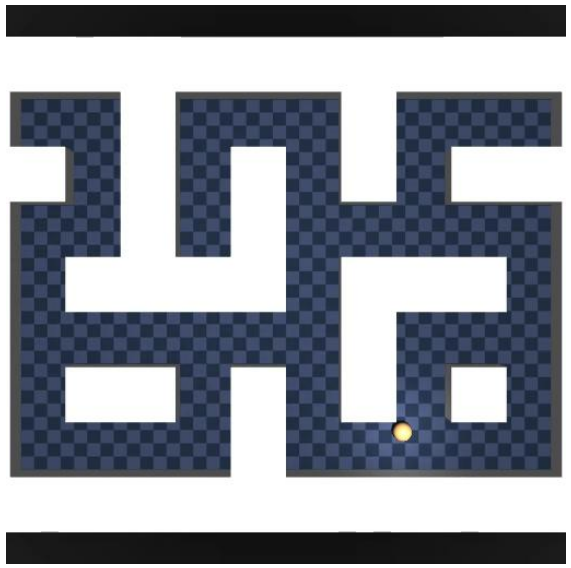


Handling stochasticity

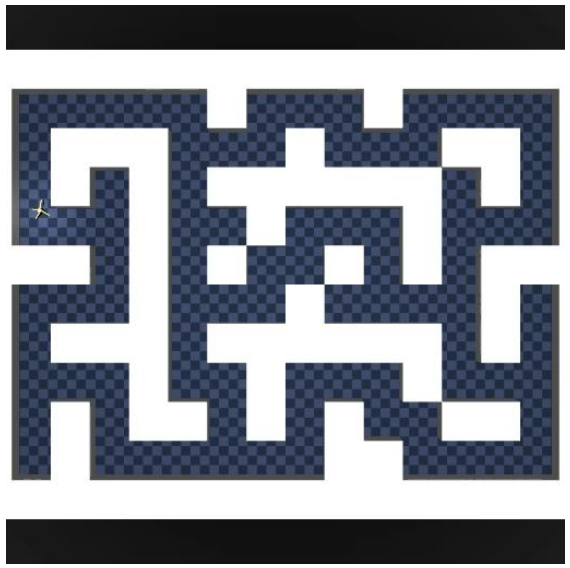
2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

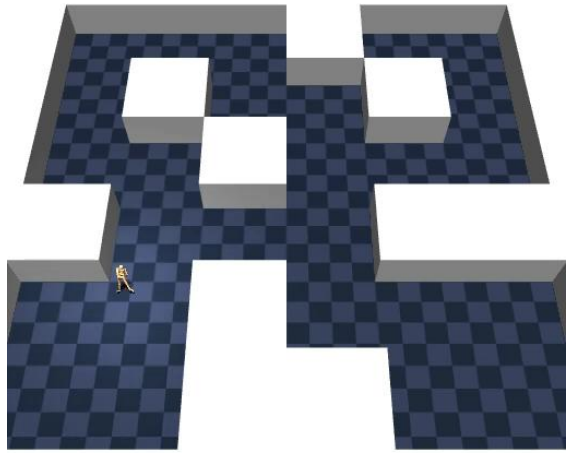
❖ **OGBench task : Locomotion task**



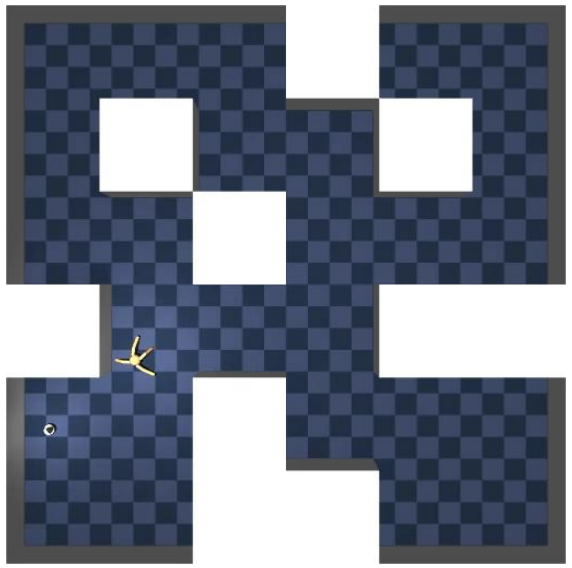
PointMaze



AntMaze



HumanoidMaze

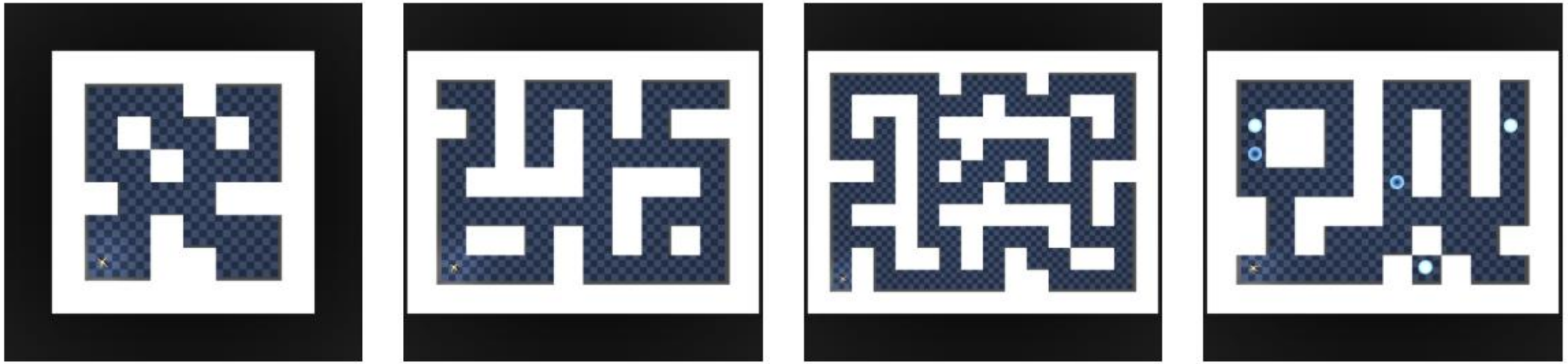


AntSoccer

2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **OGBench task : Locomotion task**

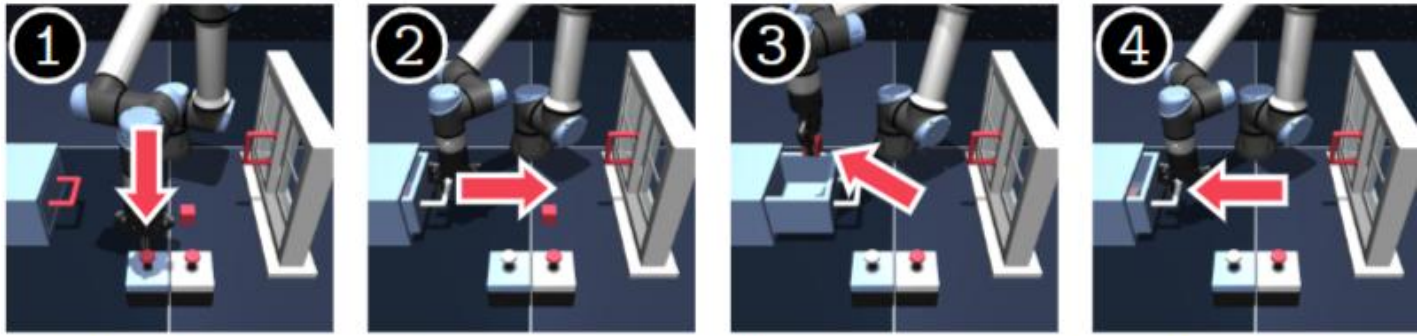


2. Related Works

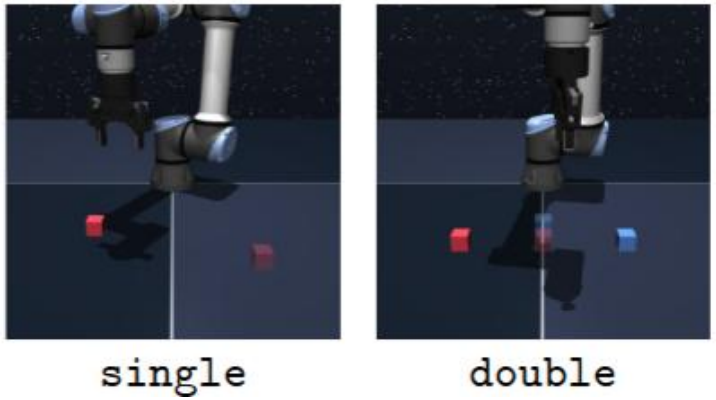
OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **OGBench task : Manipulation task**

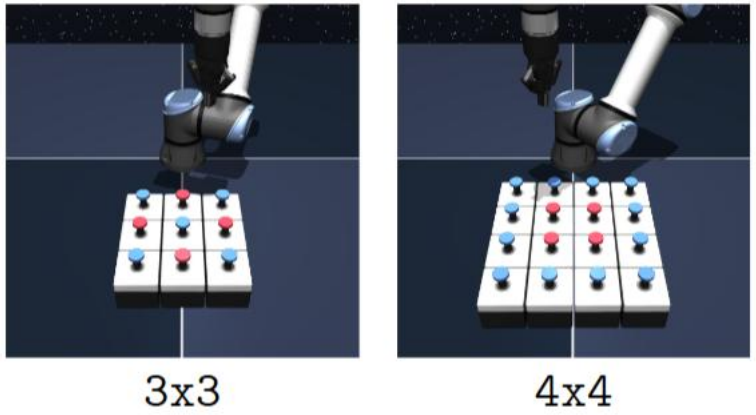
Scene



Cube



Puzzle



2. Related Works

OGBENCH: BENCHMARKING OFFLINE GOAL-CONDITIONED RL

❖ **OGBench에서의 Offline GCRL 성능 실험**

- **HIQL이 가장 높은 성능을 기록**
- **Long-horizon에서 높았던 실험이 새로운 벤치마크에서 성능이 낮아짐**

Environment	Dataset Type	Dataset	GCBC	GCIVL	GCIQL	QRL	CRL	HIQL
antmaze	navigate	antmaze-medium-navigate-v0	29 ±4	72 ±8	71 ±4	88 ±3	**95** ±1	**96** ±1
		antmaze-large-navigate-v0	24 ±2	16 ±5	34 ±4	75 ±6	83 ±4	**91** ±2
		antmaze-giant-navigate-v0	0 ±0	0 ±0	0 ±0	14 ±3	16 ±3	**65** ±5
		antmaze-teleport-navigate-v0	26 ±3	39 ±3	35 ±5	35 ±5	**53** ±2	42 ±3
	stitch	antmaze-medium-stitch-v0	45 ±11	44 ±6	29 ±6	59 ±7	53 ±6	**94** ±1
		antmaze-large-stitch-v0	3 ±3	18 ±2	7 ±2	18 ±2	11 ±2	**67** ±5
		antmaze-giant-stitch-v0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	**2** ±2
		antmaze-teleport-stitch-v0	31 ±6	**39** ±3	17 ±2	24 ±5	31 ±4	36 ±2
	explore	antmaze-medium-explore-v0	2 ±1	19 ±3	13 ±2	1 ±1	3 ±2	**37** ±10
		antmaze-large-explore-v0	0 ±0	**10** ±3	0 ±0	0 ±0	0 ±0	4 ±5
		antmaze-teleport-explore-v0	2 ±1	32 ±2	7 ±3	2 ±2	20 ±2	**34** ±15
humanoidmaze	navigate	humanoidmaze-medium-navigate-v0	8 ±2	24 ±2	27 ±2	21 ±8	60 ±4	**89** ±2
		humanoidmaze-large-navigate-v0	1 ±0	2 ±1	2 ±1	5 ±1	24 ±4	**49** ±4
		humanoidmaze-giant-navigate-v0	0 ±0	0 ±0	0 ±0	1 ±0	3 ±2	**12** ±4
	stitch	humanoidmaze-medium-stitch-v0	29 ±5	12 ±2	12 ±3	18 ±2	36 ±2	**88** ±2
		humanoidmaze-large-stitch-v0	6 ±3	1 ±1	0 ±0	3 ±1	4 ±1	**28** ±3
		humanoidmaze-giant-stitch-v0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	**3** ±2

2. Related Works

❖ **Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning(NeurIPS 2025)**

- Long-horizon에서 **HIQL의 성능 하락의 원인을 high-level policy**라는 것을 확인
- **OTA**는 **option 기반 시간적 추상화로 문제를 해결하여 long-horizon 상황에서의 성능 향상**

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

Hongjoon Ahn[1*] **Heewoong Choi**[1*] **Jisu Han**[2*] **Taesup Moon**[1,2,3†]

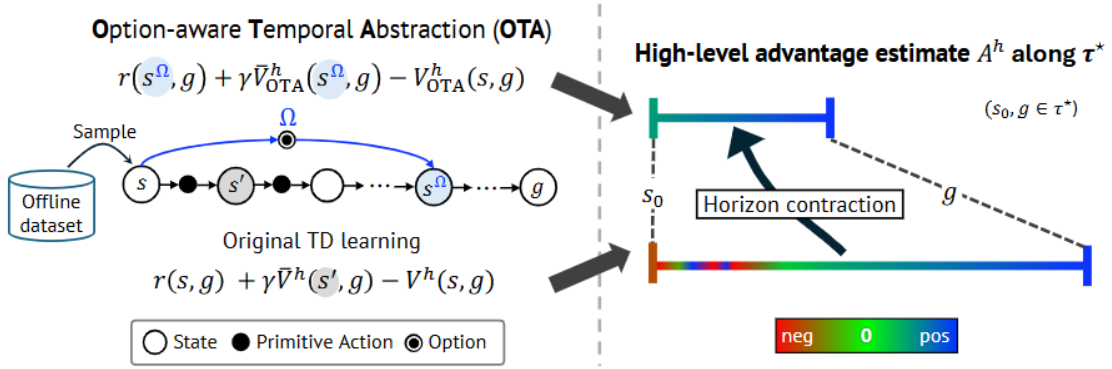
[1] Department of Electrical and Computer Engineering (ECE), Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence (IPAI), Seoul National University
[3] ASRI / INMC, Seoul National University
{hong0805, chw0501, jshcdi, tsmoon}@snu.ac.kr

Abstract

Offline goal-conditioned reinforcement learning (GCRL) offers a practical learning paradigm in which goal-reaching policies are trained from abundant state–action trajectory datasets without additional environment interaction. However, offline GCRL still struggles with long-horizon tasks, even with recent advances that employ hierarchical policy structures, such as HIQL [33]. Identifying the root cause of this challenge, we observe the following insight. Firstly, performance bottlenecks mainly stem from the high-level policy's inability to generate appropriate subgoals. Secondly, when learning the high-level policy in the long-horizon regime, the sign of the advantage estimate frequently becomes incorrect. Thus, we argue that improving the value function to produce a clear advantage estimate for learning the high-level policy is essential. In this paper, we propose a simple yet effective solution: ***Option-aware Temporally Abstracted*** value learning, dubbed **OTA**, which incorporates temporal abstraction into the temporal-difference learning process. By modifying the value update to be *option-aware*, our approach contracts the effective horizon length, enabling better advantage estimates even in long-horizon regimes. We experimentally show that the high-level policy learned using the OTA value function achieves strong performance on complex tasks from OGBench [32], a recently proposed offline GCRL benchmark, including maze navigation and visual robotic manipulation environments. Our code is available at https://github.com/ota-v/ota-v

1 Introduction

Offline goal-conditioned reinforcement learning (GCRL) has emerged as a practical framework for real-world applications by leveraging pre-collected datasets to train goal-reaching policies without requiring additional environment interaction [23, 32]. However, learning an accurate goal-conditioned value function in long-horizon settings remains a major challenge, as naively training the value function often leads to noisy estimates and erroneous policies [35, 20, 33]. To mitigate the learning of an erroneous policy, Hierarchical Implicit Q-Learning (HIQL) [33], one of the state-of-the-art

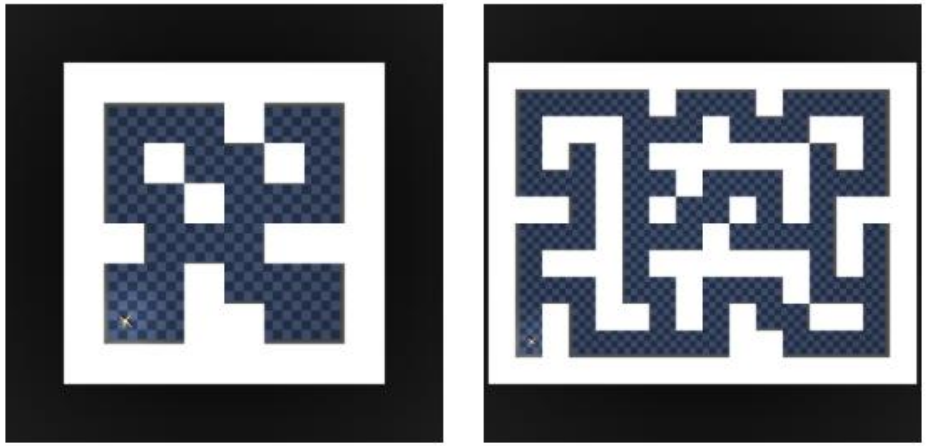


2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **HIQL은 long-horizon 상황에서의 GCRL을 잘 수행할 수 있을까?**

- High-level policy와 Low-level policy를 통해서 **long-horizon task를 해결하는 HIQL** → OGBench를 통해 평가
- Giant의 미로 크기에서 **HIQL의 성능이 낮게 평가됨**



	medium	giant
Antmaze	**96 , 94**	**65 , 2**
Humanoidmaze	**89 , 88**	**12 , 3**

*미로 및 환경에서의 success rate. 결과값은 각각 (navigate,stitch) dataset

High-level policy **Low-level policy**

2가지 policy 중 하나가
문제가 있을 것이다

2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **Subgoal을 offline dataset에서 가장 좋은 값을 제공했을 때 Low-level policy는 괜찮을까?**

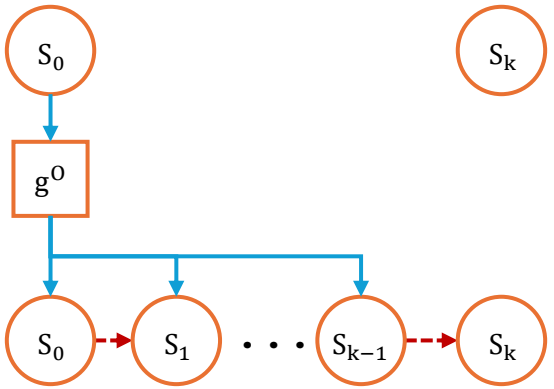
- **Oracle high-level policy** : 항상 최적의 subgoal 제시, high-level policy가 완벽하다고 가정
- Oracle에서 높은 성공률을 통해 low-level policy는 문제가 없고, high-level policy가 문제 상황으로 파악

Value function

High-level policy

Oracle High-level policy

Low-level policy



2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **Subgoal을 offline dataset에서 가장 좋은 값을 제공했을 때 Low-level policy는 괜찮을까?**

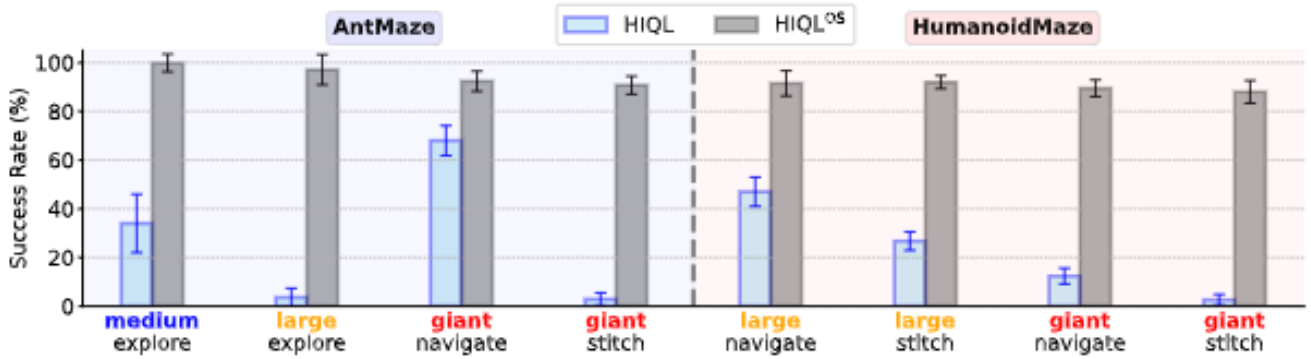
- **Oracle high-level policy** : 항상 최적의 subgoal 제시, high-level policy가 완벽하다고 가정
- Oracle에서 높은 성공률을 통해 low-level policy는 문제가 없고, high-level policy가 문제 상황으로 파악

Value function

High-level policy

Oracle High-level policy

Low-level policy

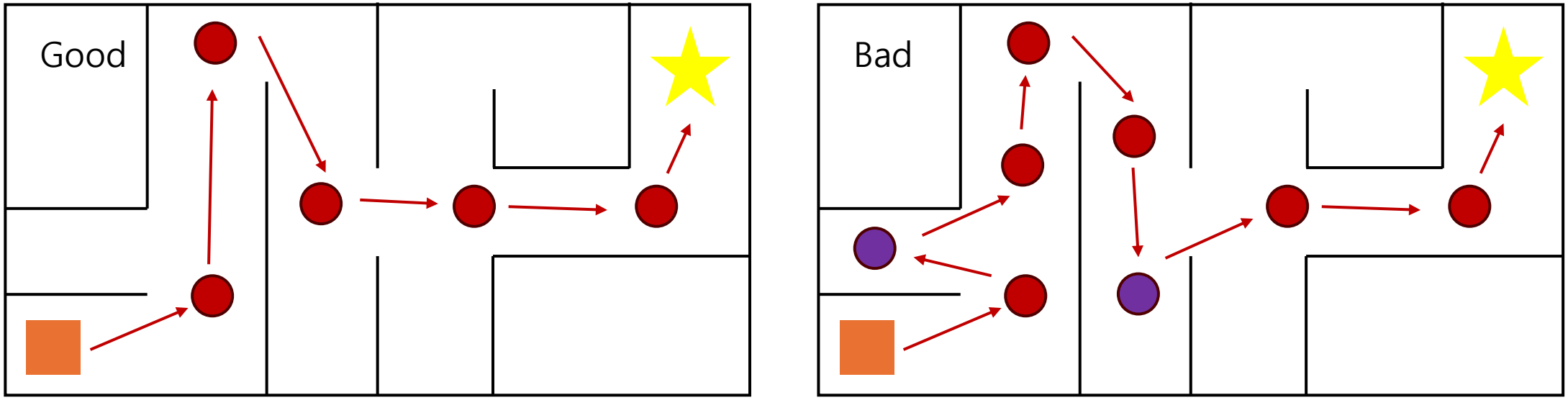


2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **High-level policy가 subgoal을 잘 제시하는가?**

- **Order Consistency** : environment가 deterministic할 때 다음 state는 goal과의 거리가 줄어듦
- High-level policy의 subgoal이 goal에 가까워지는가? → goal과 subgoal 사이 거리 변동

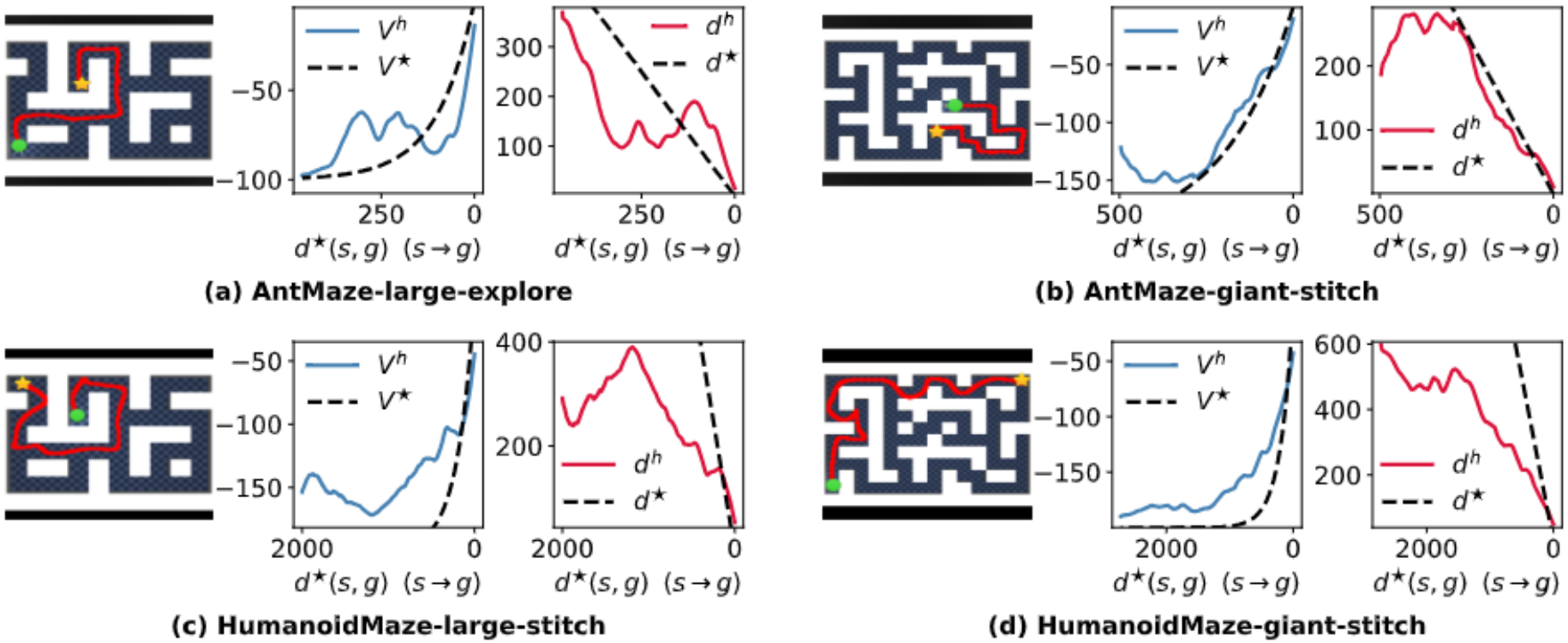


2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **High-level policy가 subgoal을 잘 제시하는가?**

- **Order Consistency** : environment가 deterministic할 때 다음 state는 goal과의 거리가 줄어듦
- High-level policy의 subgoal이 goal에 가까워지는가? → goal과 subgoal 사이 거리 변동



(a) AntMaze-large-explore

(b) AntMaze-giant-stitch

(c) HumanoidMaze-large-stitch

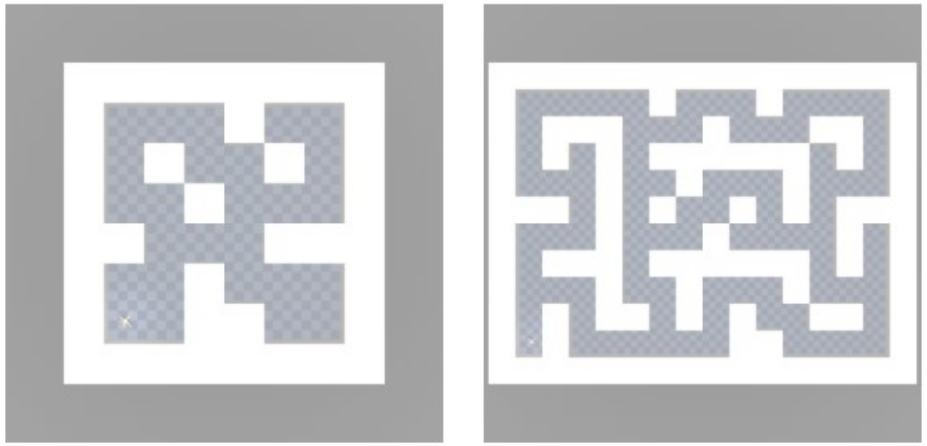
(d) HumanoidMaze-giant-stitch

2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **HIQL은 long-horizon 상황에서의 GCRL을 잘 수행할 수 있을까?**

- High-level policy와 Low-level policy를 통해서 long-horizon task를 해결하는 HIQL → OGBench를 통해 평가
- Giant의 미로 크기에서 HIQL의 성능이 낮게 평가됨



	medium	giant
Antmaze	96 , 94	65 , 2
Humanoidmaze	89 , 88	12 , 3

*미로 및 환경에서의 success rate. 결과값은 각각 (navigate,stitch) dataset

High-level policy **Low-level policy**

High-level policy
학습 방법을 바꿔보자

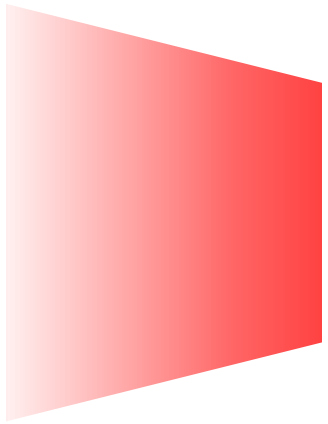
2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **Option**

- 여러 개의 **primitive action**을 하나로 묶은 확장된 행동 단위
- **Primitive action** : **매 스텝** 에이전트가 직접 취하는 **가장 세밀한 행동 단위**

각각의 행동 단위



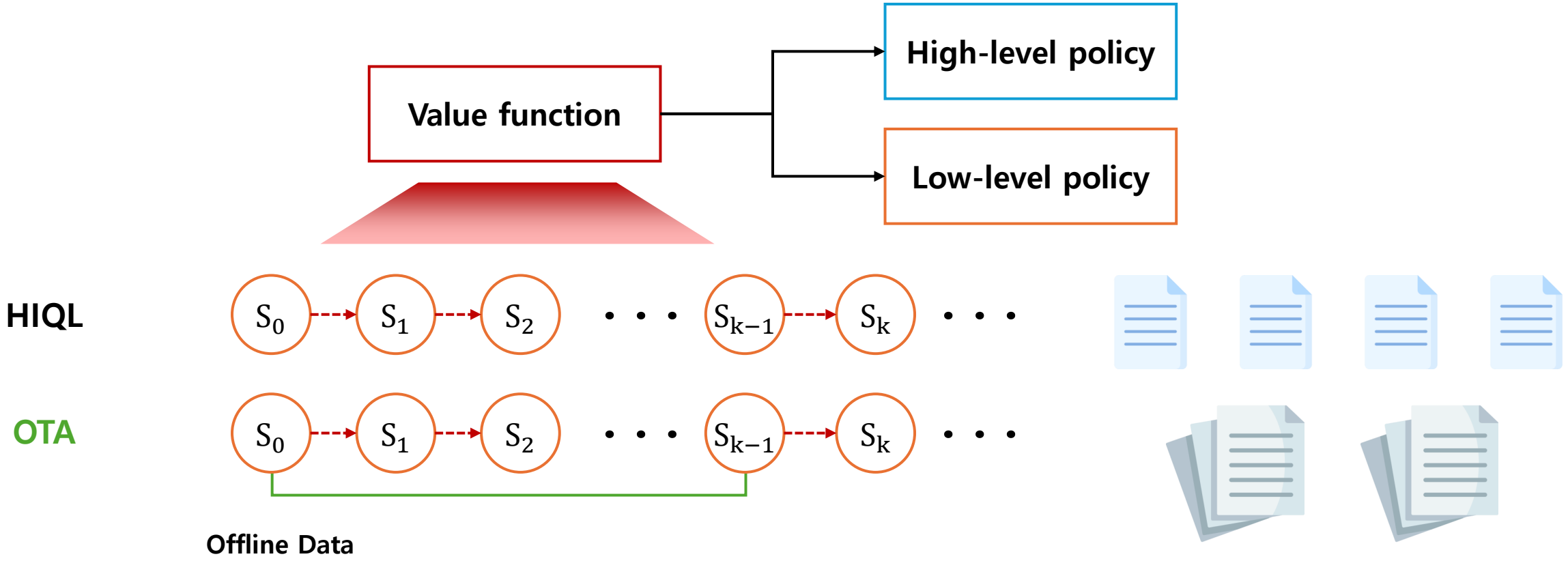
연속 행동 단위

2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **OTA**

- 한 스텝씩 value를 예측하는 대신, **temporal abstraction**으로 예측
- Value function의 노이즈를 억제를 통해 high-level policy가 정확한 subgoal 제시

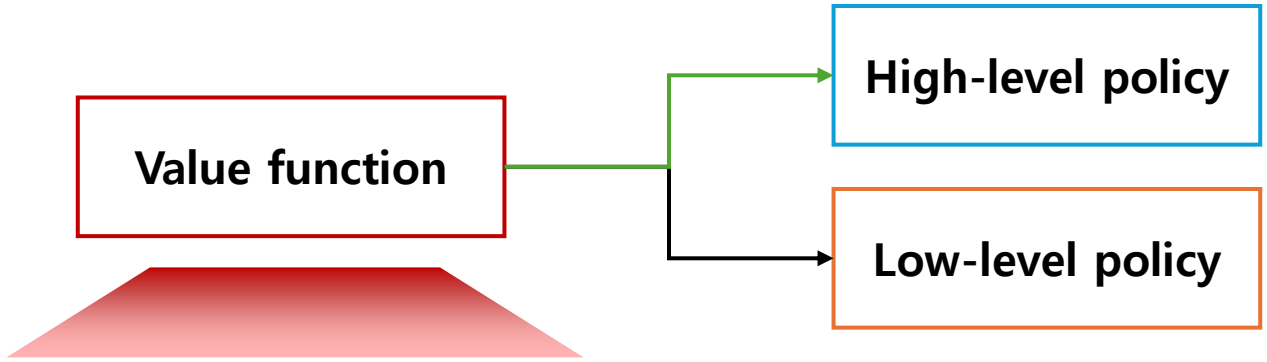


2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **OTA**

- 한 스텝씩 value를 예측하는 대신, **temporal abstraction**으로 예측
- Value function의 노이즈를 억제를 통해 high-level policy가 정확한 subgoal 제시



HIQL

$$\mathcal{L}(V) = \mathbb{E}_{(s,s')\sim D, g\sim p(g)}\left[L_2^\tau\left(r(s,g) + \gamma \bar{V}(s',g) - V(s,g)\right)\right]$$

OTA

$$\mathcal{L}\left(V_{\mathrm{OTA}}^h, n\right) = \mathbb{E}_{(s,s^\Omega)\sim\mathcal{D}, g\sim p(g)}\left[L_2^\tau\left(r\left(s^\Omega, g\right) + \gamma \bar{V}_{\mathrm{OTA}}^h\left(s^\Omega, g\right) - V_{\mathrm{OTA}}^h(s,g)\right)\right]$$

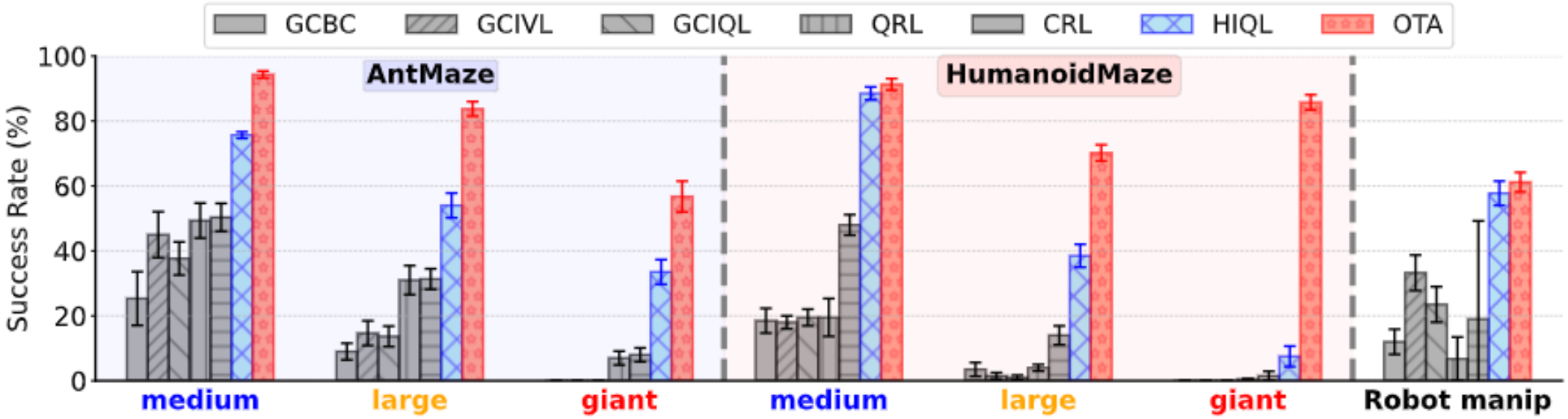
$$s^\Omega = s'\left(\Omega_{n,g}, s_t\right), \Omega_{n,g} = \begin{cases} s_{t+n} & \text{if } t+n < T \text{ and goal } g \text{ not reached} \\ g & \text{if goal } g \text{ is reached within } n \text{ steps} \end{cases}$$

2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **실험 결과 : OTA의 성능은 향상 되었는가?**

- 벤치마크 OGBench에서 **3가지 환경에서 long-horizon 상황**에서의 실험 진행
- OTA는 유의미한 **성능 향상**을 보였으며 미로 크기가 커질수록 **기존과의 격차가 벌어짐**

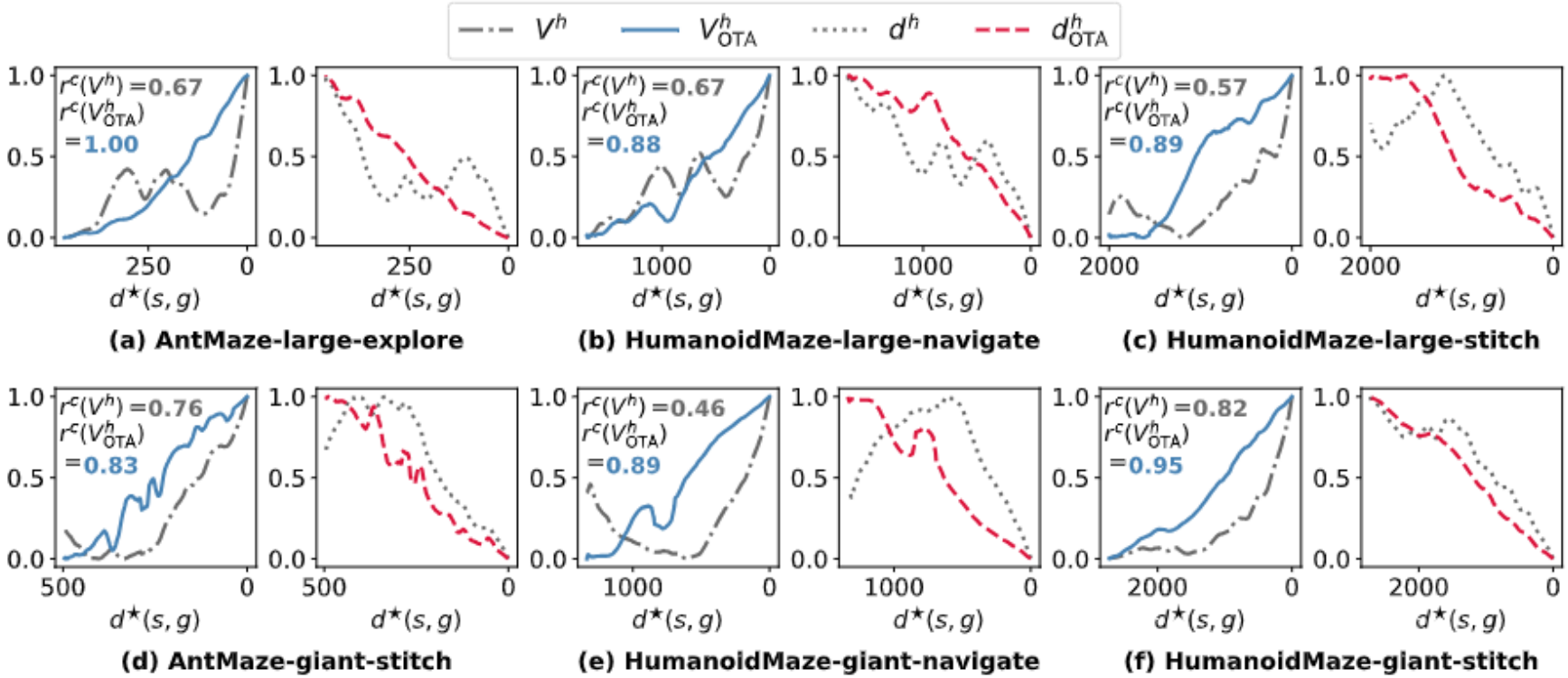


2. Related Works

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

❖ **실험 결과 : OTA는 high-level policy에서 subgoal을 잘 제시하는가?**

- HIQL과 비교하였을 때 **order consistency** 경향 파악
- OTA가 HIQL보다 **더 안정적으로 subgoal을 제공**



3. Conclusion

1. **Offline GCRL**
   - **하나의 goal을 선정**하여 도달하는 policy를 학습
   - Offline을 통해 추가적인 환경 상호작용 없이 사전 수집 데이터만으로 학습하여 **비용 위험 부담 없이 학습**

2. **HIQL**
   - **High-level policy**와 **Low-level policy**를 추출하여 **계층적 구조**를 사용

3. **OGBench**
   - Offline GCRL 알고리즘을 평가할 수 있는 **표준 벤치마크**

4. **OTA**
   - HIQL에서 high-level policy의 문제점을 해결하기 위해 **option 단위로 데이터를 묶어** 학습을 진행

4. Reference

[1] Park, S., Ghosh, D., Eysenbach, B., & Levine, S. (2023). Hiql: Offline goal-conditioned rl with latent states as actions. Advances in Neural Information Processing Systems, 36, 34866-34891.
[2] Park, S., Frans, K., Eysenbach, B., & Levine, S. (2025, May). Ogbench: Benchmarking offline goal-conditioned rl. In International Conference on Learning Representations (Vol. 2025, pp. 94937-94982).
[3] Ahn, H., Choi, H., Han, J., & Moon, T. (2026). Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. Advances in Neural Information Processing Systems, 38, 99833-99861.

고맙습니다