Journal of KIIT. Vol. 24, No. 8, pp. 89-104, Aug. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.8.89

컨테이너 터미널 적재계획 정보시스템을 위한 심층강화학습 알고리즘 비교 분석: PPO, DQN, SAC를 중심으로

유홍성*, 오태연**

Comparative Analysis of Deep Reinforcement Learning Algorithms for Container Terminal Stowage Planning Information Systems: PPO, DQN, and SAC

Hongsung Yoo*, Taeyeon Oh**

요 약

컨테이너 터미널의 적재계획은 운영 정보시스템의 핵심 의사결정 모듈이나, 심층강화학습 알고리즘의 이산 슬롯 배치 환경 적합성에 대한 근거는 충분히 규명되지 않았다. 본 논문은 단일 베이 적재계획 환경에서 PPO, DQN, SAC를 동일 학습 예산으로 비교하고, 표준 SAC의 성능 저하 원인을 세 가지 구조적 메커니즘으로 상정하여 이에 대응하는 재설계 요소의 기여를 누적 절제 실험으로 분석하였다. 단일 시드 실험에서 PPO가 최고 배치 품질과 가장 빠른 수렴을 보였고, 최대 규모에서 붕괴한 표준 SAC는 재설계를 통해 보상 기준 PPO의 약 82%까지 회복되었으며 학습신호 안정화 요소군의 기여가 가장 컸다. 실측 경험분포 재표본화 평가에서는 PPO의 성능 하락이 가장 작아, 운영 분포에 대한 강건성이 알고리즘 선택의 독립적 고려 기준이 될 수 있음을 시사한다.

Abstract

Stowage planning is a core decision module of terminal operating systems, yet little evidence exists on which deep reinforcement learning algorithm fits the discrete slot-assignment environment. We compare PPO, DQN, and the continuous-control SAC under identical budgets in a single-bay environment, hypothesize three structural failure mechanisms of standard SAC, and analyze corresponding redesign components via cumulative ablation. Under a single seed, PPO achieved the fastest convergence and highest stowage quality; standard SAC collapsed at the largest scale, and the redesign recovered it to about 82% of PPO's return, mostly via learning-signal stabilization. Episodes resampled from three months of empirical cargo data showed the smallest degradation for PPO, suggesting distributional robustness as an independent criterion.

Keywords

deep reinforcement learning, container stowage planning, soft actor-critic, proximal policy optimization, terminal operating system

* 서울과학종합대학원대학교 AI·빅데이터학과
- ORCID: https://orcid.org/0009-0004-6156-5061
** 서울과학종합대학원대학교 AI·빅데이터학과 조교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1867-5554

· Received: Jul. 13, 2026, Revised: Aug. 10, 2026, Accepted: Aug. 13, 2026

· Corresponding Author: Taeyeon Oh
Dept. of AI Bigdata, aSSIST University, Seoul, Korea
Tel.: +82-70-7012-2700, Email: tyoh@assist.ac.kr

Ⅰ. 서 론

컨테이너 터미널의 운영 경쟁력은 본선 작업의 효율성에 크게 좌우되며, 그 중심에는 적재계획(Stowage planning)이 있다. 적재계획은 선박의 안전성 제약과 하역 작업의 효율성을 동시에 고려하여 컨테이너의 선내 배치 위치를 결정하는 문제로, 터미널 운영시스템(TOS, Terminal Operating System)의 핵심 의사결정 모듈에 해당한다[1]. 컨테이너 적재계획 문제(CSPP, Container Stowage Planning Problem)는 NP-난해 조합최적화 문제로서 수리계획과 휴리스틱을 중심으로 오랜 기간 연구되어 왔으나, 문제 정식화의 과도한 단순화로 인해 산업 현장 적용성을 평가하기 어렵다는 한계가 지적되어 왔다[2]. 최근에는 터미널 운영 전반에 인공지능 기법을 도입하려는 시도가 활발해지면서, 적재계획 자동화를 위한 학습 기반 접근이 주목받고 있다[1].

이러한 흐름에서 심층강화학습(DRL, Deep Reinforcement Learning)을 적재계획에 적용하는 연구가 국내외에서 빠르게 증가하고 있다. 국외에서는 수요 불확실성 하의 마스터 적재계획에 제약 만족을 보장하는 DRL 모델이 제안되었고[3], 실선박 규모의 종단간(End-to-end) DRL 계획 모델이 확률계획 기반 기법을 상회하는 성능을 보였다[4]. 국내에서도 항만 야드의 컨테이너 다단 적재에 강화학습을 적용하여 재취급을 최소화하는 연구[5]와 DQN 기반 적재 순서 최적화 연구[6]가 보고되었다. 그러나 이들 연구는 대부분 특정 알고리즘 하나를 선택하여 적용 가능성을 보이는 데 집중하였고, 어떤 DRL 알고리즘이 적재계획이라는 이산(Discrete) 슬롯 배치 환경에 구조적으로 적합한지에 대한 근거는 충분히 제시되지 않았다. 최근 Huang et al.[7]이 적재계획 문제에 대해 다중 DRL 알고리즘 벤치마크를 수행하여 이 공백을 부분적으로 메웠으나, 해당 연구는 DQN(Deep Q-Network), PPO(Proximal Policy Optimization) 계열 알고리즘의 성능 순위 비교에 머물렀고, 연속 행동공간용으로 설계된 대표적 off-policy 알고리즘인 SAC(Soft Actor-Critic)는 비교 대상에 포함하지 않았다.

단순한 성능 우열 비교는 두 가지 한계를 갖는다. 첫째, SAC는 본래 연속 행동공간을 전제로 설계된 알고리즘이므로[8], 이산 슬롯 배치 문제에서 정책기반 알고리즘에 뒤처진다는 결과는 예상 가능한 결론에 그치기 쉽다. 둘째, 성능 순위는 왜 특정 알고리즘이 실패하는지, 그리고 어떤 개입으로 회복될 수 있는지에 대한 정보를 제공하지 못하므로, 정보시스템 설계자가 알고리즘을 선택하고 조정하는 데 필요한 실무적 지침으로 이어지지 않는다. 이에 본 연구는 관점을 전환하여, 표준 SAC가 이산 적재계획 환경에서 성능 저하를 일으키는 구조적 원인을 행동공간 불일치, 커리큘럼 전환에 따른 리플레이 버퍼 오염, 엔트로피 정규화 및 보상 스케일 왜곡의 세 가지 메커니즘으로 분해하고, 각 메커니즘에 대응하는 재설계 요소를 단계적으로 누적하는 절제(Ablation) 실험을 통해 그 기여를 실증하는 진단형(Diagnostic) 접근을 채택한다. 아울러 이산 행동공간 전용으로 설계된 off-policy 알고리즘인 DQN[9]을 대조군으로 포함함으로써, 관찰되는 성능 저하가 off-policy 학습 일반의 문제인지 SAC 고유의 문제인지를 분리하여 검증한다.

이산 문제에 SAC를 적용하는 것은 단순한 학술적 호기심이 아니라 실무적 요구에 기반한다. 터미널 운영 정보시스템은 적재계획과 같은 이산 의사결정과 크레인 속도·궤적 제어와 같은 연속 제어를 함께 포함하므로, 단일 알고리즘 스택으로 두 유형을 통합할 수 있다면 시스템 유지보수 측면에서 유리하다. 또한 off-policy 학습의 높은 샘플 효율은 시뮬레이션 비용이 큰 운영 환경에서 실질적 이점이 된다. 따라서 SAC의 이산 환경 적합성을 규명하고 회복 가능성을 확인하는 것은 응용 시스템 설계의 선결 과제라 할 수 있다.

본 논문의 기여는 다음과 같다. 첫째, 이산 컨테이너 적재계획 환경에서 표준 SAC의 실패 원인을 세 가지 구조적 메커니즘으로 분해하고, 각 메커니즘에 대응하는 재설계 요소(연속 행동 매핑, 관측·보상 정규화 및 엔트로피 계수 고정, 잠재함수 기반 보상 성형(Reward shaping)과 커리큘럼 버퍼 초기화)의 누적 투입 순서 기준 한계 기여를 절제 실험으로 측정하였다. 둘째, PPO[10]와 DQN을 포함한 동일 환경·동일 시드·동일 학습 예산의 공정 벤

치마크를 구성하고, 안정성과 효율성 지표를 분리한 5종 KPI로 알고리즘-환경 정합성을 다각도로 평가하였다. 셋째, 실제 컨테이너 터미널의 약 3개월치 적하 데이터로 시뮬레이션 환경의 화물 분포를 검증하고 실측 경험분포 재표본화 평가를 추가함으로써, 시뮬레이션 결과가 운영 분포 변화 하에서 갖는 강건성까지 검토하였다. 이를 통해 터미널 운영 정보시스템의 알고리즘 선택에 관한 실무적 지침을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 컨테이너 적재계획에 대한 강화학습 적용 연구와 이산 행동공간에서의 알고리즘 적용 문제를 정리한다. 3장에서는 적재계획 강화학습 환경과 SAC 실패 메커니즘의 진단 및 재설계 방법을 기술한다. 4장에서는 실험 설계와 결과 분석을 제시하고, 5장에서 결론과 향후 과제를 기술한다.

II. 관련 연구

2.1 컨테이너 적재계획과 강화학습

CSPP는 항로 전체의 기항 순서를 고려하는 다항만(Multi-port) 문제와 단일 항만에서의 슬롯 배치 문제로 구분되며, 실무에서는 베이 단위 배치를 결정하는 마스터 계획과 개별 슬롯을 결정하는 슬롯 계획으로 계층 구분하여 접근하는 것이 일반적이다[2]. 전통적으로는 정수계획, 제약계획 및 휴리스틱 기법이 주류를 이루었으나, 계산 시간과 문제 단순화의 한계로 인해 실운영 적용에는 제약이 있었다[2][11]. 국내에서도 재취급 최소화, 작업 효율성, 장비 운영의 관점에서 수리적 최적화, 휴리스틱, 강화학습 기반 기법을 아우르는 연구 동향이 정리된 바 있다[11].

강화학습 기반 접근은 적재계획을 순차적 의사결정 문제로 정식화하여 계획 생성 시간을 크게 단축할 수 있다는 장점이 있다. van Twiller et al.[3]은 수요 불확실성 하의 마스터 적재계획에 대해 인코더-디코더 구조와 타당성 사영(Feasibility projection) 계층을 결합한 DRL 모델을 제안하여 볼록 제약의 만족을 보장하였고, 후속 연구인 AI2STOW[4]는 실선박 규모 문제에서 강화학습 및 확률계획 베이스라인을 상회하는 종단간 계획 성능을 보였다. 국내에서는 Jang et al.[5]이 항만 야드 환경에서 출고 일자를 고려한 다단 적재 강화학습 모델로 재취급 최소화를 달성하였고, Cho et al.[6]은 DQN을 이용하여 재취급 수를 최소화하는 적재 순서 결정 모델을 제안 하였다. 또한 항만 물류의 배차 문제에 다중 에이전트 강화학습을 적용한 연구[12]와 3차원 적재 최적화에 PPO와 그래프 신경망을 결합한 연구[13] 등, 항만·물류 도메인 전반으로 강화학습 적용이 확산되고 있다.

이와 같이 개별 알고리즘의 적용 가능성은 충분히 입증되었으나, 알고리즘 간 비교를 통해 환경 적합성을 규명하려는 시도는 최근에야 시작되었다. Huang et al.[7]은 크레인 스케줄링을 포함한 적재계획 Gym 환경을 구축하고 DQN, QR-DQN, A2C, PPO, TRPO의 5개 알고리즘을 벤치마크하여, 문제 복잡도가 증가할수록 알고리즘 간 성능 격차가 뚜렷해짐을 보였다. 그러나 이 연구는 성능 순위의 보고에 초점을 두어 격차의 구조적 원인을 분석하지 않았고, 연속 행동공간 계열 알고리즘(SAC 등)은 비교에서 제외하였다. 본 연구는 이 부분에 초점을 둔다. 즉, 성능 순위가 아니라 실패 원인의 분해와 재설계에 의한 회복 가능성을 다루며, 실측 항만 데이터로 환경 분포를 검증한다는 점에서 기존 벤치마크 연구와 구별된다.

2.2 심층강화학습 알고리즘과 이산 행동공간 적용 문제

본 연구에서 다루는 세 알고리즘은 학습 방식과 행동공간 전제가 서로 다르다. PPO[10]는 클리핑된 대리 목적함수로 정책 갱신 폭을 제한하는 on-policy 정책기반 알고리즘으로, 이산·연속 행동공간 모두에 적용 가능하며 안정적인 수렴 특성으로 응용 연구에서 기준 알고리즘으로 널리 채택된다[3][7]. DQN[9]은 이산 행동공간을 전제로 행동가치 함수를 근사하는 off-policy 가치기반 알고리즘이다. 반면 SAC[8]는 최대 엔트로피 목적함수 하에서 확률적 정책과 소프트 Q-함수를 동시에 학습하는

off-policy 알고리즘으로, 쌍곡탄젠트(Hyperbolic tangent, tanh)로 압착된(Squashed) 가우시안 정책을 사용하여 연속 행동공간을 명시적으로 전제한다. SAC의 높은 샘플 효율과 탐색 성능은 로봇 제어 등 연속 제어 영역에서 입증되었으나, 이러한 설계 전제는 이산 환경 적용 시 세 가지 문제를 야기한다. 연속 출력을 이산 행동으로 변환하는 과정의 정보 손실과 경계 포화, 이산 보상 구조에서의 엔트로피 자동 조정 왜곡, 그리고 환경 비정상성(Non-stationarity) 하에서의 리플레이 버퍼 오염이 그것이다. 이들 메커니즘의 상세한 분석은 3.2절에서 다룬다.

SAC를 이산 행동공간으로 확장하려는 기존 연구로는 정책과 크리틱을 범주형 분포로 재유도한 SAC-Discrete[14]가 대표적이다. 이는 알고리즘 내부를 수정하는 접근이므로 두 가지 실무적 제약이 따른다. 검증된 표준 SAC 구현체를 그대로 활용할 수 없어 별도의 구현과 검증이 필요하고, 이산 문제와 연속 제어 문제에 서로 다른 알고리즘을 각각 운용해야 하므로 하나의 SAC 구현으로 두 영역을 함께 관리하는 이점이 사라진다. 이에 반해 본 연구는 표준 SAC 구현을 유지한 채 행동 인터페이스(연속-이산 매핑), 학습 신호(정규화·보상 성형), 버퍼 관리(커리큘럼 전환 시 초기화)를 재설계하는 환경 측 접근을 취하며, 각 요소의 기여를 절제 실험으로 분리 검증한다는 점에서 차별화된다. 세 접근을 대비하면 DQN과 PPO는 이산 행동을 직접 지원하고, SAC-Discrete는 알고리즘 내부를 재유도하며, 본 연구는 표준 SAC를 보존한 채 환경 인터페이스만 조정한다. 다만 SAC-Discrete와의 정량 비교는 본 실험 범위에 포함되지 않았으므로, 본 결과는 표준 SAC 대비 회복 가능성의 확인에 한정된다. 한편 잠재함수 기반 보상 성형(Reward shaping)이 최적 정책을 보존한다는 이론적 결과[15]와 커리큘럼 학습의 전이 설계에 관한 논의[16]는 본 재설계의 이론적 기반을 제공한다.

국내에서도 응용 도메인별 DRL 알고리즘 비교 연구가 보고되고 있다. Kim 등[17]은 사이버 전장 시뮬레이션 환경에서 DQN, PPO, SAC의 학습 성능을 비교하여 도메인 특성에 따른 알고리즘 선택의 중요성을 보였다. 그러나 해당 연구를 포함한 기존 비교 연구들은 관찰된 성능 차이를 보고하는 데 그치며, 특정 알고리즘의 부진을 구조적 원인 수준에서 진단하고 재설계로 회복시키는 접근은 시도되지 않았다. 본 연구는 이산 적재계획이라는 산업 응용 환경에서 이러한 진단-재설계-검증의 전 과정을 통합적으로 수행한 사례를 제시하며, 알고리즘-환경 적합성에 대한 실무 지침을 도출하고자 한다. 이상의 검토를 토대로 본 연구의 질문을 다음 세 가지로 정리한다. RQ1: 표준 SAC의 성능 저하는 off-policy 학습 일반의 문제인가, 연속 행동공간용 알고리즘을 이산 환경에 적용하는 데서 발생하는 구조적 문제인가. RQ2: 실패 메커니즘에 대응하는 세 재설계 요소군의 상대적 기여는 어떠한가. RQ3: 학습 분포에서의 성능은 실측 화물 분포에서도 유지되는가.

III. 적재계획 강화학습 환경과 SAC 재설계

3.1 단일 베이 적재계획 환경

본 연구의 환경은 컨테이너선의 단일 베이(Single bay)를 대상으로 하는 슬롯 계획 문제를 Gymnasium 인터페이스로 구현한 것이다. 베이는 행(Row)과 단(Tier)으로 구성된 격자이며, 에이전트는 선적 순서에 따라 도착하는 컨테이너 각각에 대해 적재할 행을 선택한다. 선택된 행의 최하단 빈 단에 컨테이너가 적재되며, 모든 컨테이너의 배치가 완료되면 에피소드가 종료된다. 즉 본 환경은 단을 직접 선택하지 않는 행 할당(Row-assignment) 축약 문제로, 실제 CSPP 전체가 아닌 그 부분 문제를 다룬다. 문제는 MDP (S, A, P, R, γ)로 정식화되며 상태공간 $S \subset [-2,2]^{79}$, 행동공간 $A = \{0,\cdots,9\}$, 할인율 $\gamma = 0.99$이다. 화물은 40ft 일반(GP) 컨테이너 단일 종류로 하고, 컨테이너 무게는 10~20MT의 균등분포에서, 양하항(POD, Port of Discharge)은 6개 항만 중에서 무작위로 생성된다. 선적순서는 실무 관행에 따라 POD가 먼 항만의 화물부터, 동일 POD 내에서는 무거운 화물부터 선적하는 것으로 정하였다.

문제 규모에 따른 학습 안정화를 위해 4단계 커

리큘럼을 적용한다. 격자 크기와 컨테이너 수는 4×4(16개), 6×6(36개), 8×8(64개), 10×10(100개)로 단계적으로 확대되며, 모든 단계에서 적재율은 100%로 유지된다. 상태 벡터는 커리큘럼 전 단계에서 차원이 79로 고정되며, 행별 특성 6종(적재율, 누적 무게, 평균 POD, 최상단 POD, 순서역전율, 열 무게비) 60차원, 현재 컨테이너 특성(무게, POD one-hot) 7차원, 전역 특성(진행률, 무게중심 편차, 무게 변동계수 등) 6차원, 잔여 화물의 POD 분포 6차원으로 구성되며, 미사용 행동의 특성은 0으로 패딩된다. 행동공간은 행 선택에 해당하는 Discrete(10)으로, 커리큘럼 전이 시에도 관측·행동 차원이 보존되어 정책의 전이 학습이 가능하다. 행동 마스킹은 사용하지 않으며, 해당 단계에 존재하지 않는 행(예: Lv1의 4~9번)을 선택하면 R2의 2배 페널티가 부과되고 그 컨테이너는 미배치로 건너뛴 채 에피소드가 계속된다. 가득 찬 행을 선택한 경우에도 R2 부과 후 동일하게 진행되므로, 한 번의 무효 선택은 해당 컨테이너의 최종 미배치로 직결되어 유효배치율에 그대로 반영된다. 이 전이 규칙은 세 알고리즘에 동일하게 적용된다. 한편 선적 순서가 POD 내림차순으로 고정되는 설계상 오버스토우는 구조적으로 발생하기 어려우며, 이는 본 환경이 오버스토우 회피 능력 자체를 변별하지 못함을 의미한다.

보상 함수는 총 15개 항목(R1~R15)을 필수, 안정성, 효율성의 3계층으로 구분하여 설계하였으며, 표 1은 항목별 목적, 보상값(수식), 적용 시점을 보인다. 필수 계층(R1, R2, R3, R7, R9, R11)은 물리적·운영적 제약 위반을 직접 억제하고, 안정성 계층(R5, R6)은 선체 안정성을, 효율성 계층(R4, R8, R10, R12, R13, R14, R15)은 양하 작업 효율을 유도한다. 안정성과 효율성 계층의 가중치는 계수 절댓값 합 기준 50:50으로 균형을 맞추었다. 열별 최대 누적 무게 제약은 160MT로 설정하였는데, 이는 실제 운항 중인 2,300TEU급 컨테이너선의 스택 하중 한계(갑판 100MT, 선창 210MT) 범위 내의 값으로, 최대 규모(10×10) 문제에서 전체 열 용량(1,600MT)이 화물 총중량 기대값(1,500MT)을 상회하여 제약 준수가 구조적으로 달성 가능하도록 한다.

표 1. 보상 함수 구성: 항목별 목적, 보상값, 적용시점
Table 1. Reward components: purpose, value, and timing

Item	Purpose	Value (formula)	Timing
R1	valid placement	+1.0	step
R2	full-stack penalty	−10.0 (invalid row −20.0)	step
R3	overstowage penalty	−15.0	step
R4	unloading-order bonus	+2.0 (+0.6 same POD/empty)	step
R5	row weight balance	+6.5×(1−min(CV,1))	step
R6	CoG deviation penalty	−8.5×|c−c*|/c*	step
R7	episode completion	+80×VPR	final
R8	POD band quality	+1.33×purity×tiers	final
R9	column weight limit	−5.0 (over 160MT)	step
R10	same-tier same-POD	+2.67×min(n,4)	step
R11	weight inversion penalty	−6.0	step
R12	perfect column order	+2.0×columns	final
R13	target-tier match	+1.33 / −1.0×dev. (−0.5 below)	step
R14	empty-row penalty	−2.0×rows	final
R15	vertical same-POD penalty	−2.67×min(run−1,3)	step

※ CV: coefficient of variation of row weights, c: center of gravity, c*: bay center, purity: ratio of the most frequent POD within a tier

시뮬레이션 환경의 화물 분포가 실무와 정합하는지 확인하기 위해, 실제 컨테이너 터미널의 3개월치 적하 데이터(38개 항차, 23,239건)를 분석하였다. 리퍼, 위험물, 규격외 화물을 제외한 40ft 일반 컨테이너 13,174건(57%)을 기준으로 할 때, 항차×베이 단위 화물 집합의 94%가 본 환경의 커리큘럼 규모(16~100개)에 대응하였고, 베이별 POD 종류 수(중앙값 1, 최대 4)는 환경의 6-POD 상한 내에 분포하였다. 다만 실측 무게 분포(평균 12.7MT, 표준편차 9.5MT)는 환경의 균등분포 가정보다 넓게 퍼져 있어(10~20MT 구간 비율 22.1%), 그림 1과 같이 두 분포의 차이가 확인된다. 이에 본 연구는 학습 분포는 유지하되, 실측 경험분포에서 평가 에피소드를 재표본화하여 생성하는 평가를 4.4절에 추가하여 분포 차이가 배치 성능에 미치는 영향을 검증한다.

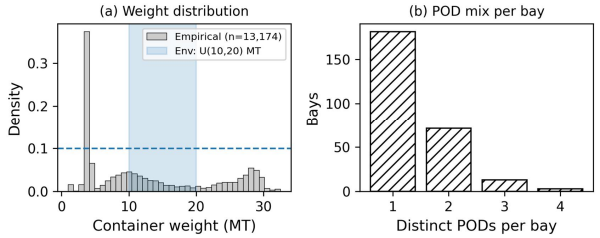

그림 1. 실측 화물 분포와 환경 가정의 비교
Fig. 1. Comparison of the empirical cargo distribution and environment assumptions

3.2 SAC 실패 메카니즘 진단

표준 SAC를 위 환경에 적용할 때 성능 저하를 일으킬 것으로 예상되는 구조적 원인을 세 가지 메커니즘으로 설명한다(이하 각각 메커니즘 ①, ②, ③으로 표기하며, 3.3절의 재설계 요소와 대응된다).

첫째, 행동공간 불일치이다(메커니즘 ①). SAC의 정책은 tanh로 압착된 가우시안 분포로부터 연속 행동을 샘플링하므로, 이산 행 선택에 적용하려면 연속 출력을 이산 인덱스로 변환해야 한다. 가장 단순한 방법인 1차원 출력의 균등 구간 반올림(Naive discretization)은 두 가지 문제를 갖는다. tanh 함수의 포화 특성상 양 끝 구간(0번, 9번 행)에 도달하려면 큰 크기의 사전 활성값이 필요하여 경계 행의 선택 빈도가 구조적으로 낮아지고(경계 포화), 반올림 경계 부근에서는 연속 출력의 미세한 변화가 서로 다른 행 선택으로 불연속하게 이어져 정책 경사 신호가 행동 결과와 정합하지 않는다. 아울러 off-policy 학습의 리플레이 버퍼에는 이산 변환 전의 연속 행동이 저장되어 크리틱이 연속 행동 공간에서 학습되므로, 동일한 행으로 매핑되는 서로 다른 연속 행동이 크리틱에는 구별되어 반영되는 불일치와, 내림 변환의 비미분성으로 정책 경사가 변환을 직접 통과하지 못하는 한계가 함께 존재한다.

둘째, 커리큘럼 전환에 따른 리플레이 버퍼 오염이다(메커니즘 ②). off-policy 알고리즘인 SAC는 과거 전이(transition)를 리플레이 버퍼에 저장하여 재사용하는데, 커리큘럼 단계가 전환되면 격자 크기, 컨테이너 수, 에피소드 길이가 모두 달라져 상태-행동 분포가 급변하는 비정상성이 발생한다[16]. 이전 단계의 전이가 버퍼에 잔존한 채 새 단계의 학습에 혼입되면 소프트 Q-함수의 타깃이 서로 다른 문제의 가치 척도로 오염되며, 이는 정책이 on-policy 데이터로만 갱신되는 PPO에는 존재하지 않는 SAC 고유의 취약점이다.

셋째, 엔트로피 정규화와 보상 스케일의 왜곡이다(메커니즘 ③). SAC의 자동 엔트로피 조정은 행동공간 차원에 기반한 목표 엔트로피를 기준으로 온도 계수를 적응시키는데, 연속-이산 매핑 하에서는 명목상의 연속 행동 차원과 실질적인 이산 선택지 수가 괴리되어 목표 엔트로피가 과대 설정되고 탐색이 과도하게 유지될 수 있다. 또한 본 환경의 보상은 스텝 보상(±15 내외)과 에피소드 종료 보상(최대 +80)의 스케일 차이가 크고 커리큘럼 단계에 따라 에피소드 누적 보상 범위가 수배씩 달라지므로, 관측·보상 정규화가 없는 경우 Q-함수 타깃의 스케일이 불안정해져 크리틱 학습이 저해된다.

3.3 SAC 재설계와 절제 실험 설계

3.2절에서 진단한 세 가지 메커니즘에 대응하는 재설계 요소를 정의하고, 이를 단계적으로 누적하는 4단계의 절제(Ablation) 구성을 설계한다. 기준 구성인 S0은 표준 SAC에 1차원 출력의 균등 구간 반올림 이산화만 적용한 것으로, 3.2절에서 진단한 세 메커니즘의 영향을 모두 받는다.

S1은 행동공간 불일치(메커니즘 ①)에 대응하여 행동 인터페이스를 2차원 연속 상자 공간 Box(-1,1)²로 확장한다. 정책이 출력한 두 연속 행동 a1과 a2는 식 (1)과 같이 주 차원과 보조 차원의 가중 결합을 통해 스칼라 u ∈ [0,1]로 변환된다.

$$u = 0.9 \cdot \frac{(a_1 + 1)}{2} + 0.1 \cdot \frac{(a_2 + 1)}{2} \tag{1}$$

식 (1)에서 a_1은 행 선택의 주 신호, a_2는 미세 조정 신호이며, 이러한 가중 결합 구조는 연속 프로토 행동(Proto-action)을 이산 행동으로 사상하는 대규모 이산 행동공간 접근[18]과, 기본 신호에 소폭의 잔차(Residual) 신호를 더해 미세 조정을 담당시키는 잔차 강화학습[19]의 설계 원리를 따른 것이다. 이에 따라 보조 차원의 가중치는 잔차 성분이 기본

신호를 지배하지 않아야 한다는 원리[19]에 근거하여 주 차원 대비 작은 값으로 유지하였으며, 구체적인 값(0.9/0.1)은 사전 학습 시도 결과를 토대로 경험적으로 설정하였다. 가중치에 대한 체계적 탐색이나 민감도 분석은 수행하지 않았다(5장 향후 과제 참조). 다만 학습된 최종 재설계 정책을 대상으로 추론 단계에서 가중치 변경의 영향을 점검하는 강건성 평가를 수행하였으며, 그 결과는 4.3절에서 제시한다. 적재 행 인덱스 k는 식 (2)와 같이 u의 균등 구획으로 결정된다.

$$k = \min(\lfloor u \cdot N \rfloor,\ N-1),\ N=10 \tag{2}$$

식 (2)에서 N은 선택 가능한 행의 수이고, $\lfloor \cdot \rfloor$은 내림(floor) 연산이며, min 연산은 u가 1일 때 행 인덱스가 최댓값 N−1을 넘지 않도록 절단하는 역할을 한다. 보조 차원 a2는 주 차원이 tanh 포화 영역에 있을 때에도 경계 행(k=0, 9)에 도달할 수 있는 우회 경로를 제공하여, 반올림 이산화의 경계 포화 문제를 완화한다. 다만 이는 포화의 근본적 해소가 아니라 우회 경로의 추가이며, 변환의 비미분성은 S1에서도 유지된다(4.3절에서 행동 선택 분포로 그 효과를 직접 확인한다).

S2는 학습신호 왜곡(메커니즘 ③)에 대응하여 S1에 관측·보상의 이동 통계 정규화(VecNormalize)를 추가하고, 엔트로피 계수를 0.01로 고정하여 자동 온도 조정에 의한 과잉 탐색을 차단한다. 이로써 커리큘럼 단계에 따라 수배씩 달라지는 보상 스케일이 표준화되어 소프트 Q-함수 타깃이 안정된다.

S3은 리플레이 버퍼 오염(메커니즘 ②)에 대응하는 최종 재설계 구성으로, S2에 잠재함수 기반 보상 성형과 커리큘럼 전환 시 버퍼 초기화를 추가한다. 잠재함수 Φ(s)는 식 (3)과 같이 상태 s에서의 누적 유효배치율 VPR(s), 오버스토우율 OSR(s), 열무게제한 위반율 CWVR(s), 적재 진행률 ρ(s)의 가중합으로 정의한다.

$$\Phi(s) = 10 \cdot \mathrm{VPR(s)} + 10 \cdot (1-\mathrm{OSR(s)}) + 5 \cdot (1-\mathrm{CWVR(s)}) + 5 \cdot \rho(\mathrm{s}) \tag{3}$$

식 (3)의 잠재함수는 배치가 진전되고 제약 위반이 없을수록 커지는 상태 가치의 근사로, 희소한 에피소드 종료 보상(R7)에 의존하는 학습 초기 구간에 밀도 높은 학습신호를 공급한다. 성형이 적용된 보상 r′는 식 (4)와 같다.

$$r' = r + \gamma\,\Phi(s') - \Phi(s) \tag{4}$$

식 (4)의 잠재함수 기반 성형은 최적 정책을 변화시키지 않음이 이론적으로 보장된다[15]. 여기서 γ는 할인율, s′는 다음 상태이다. 에피소드 종료 시에는 Φ(s′) = 0으로 처리하여 종료 상태의 잠재값을 소거하였고, 본 환경은 시간 제한에 의한 절단(Truncation) 없이 종료(Termination)만 존재하므로 [15]의 정책 불변 조건이 유지된다. 성형은 관측·보상 정규화 이전의 환경 단계에서 적용된다. 아울러 커리큘럼 전환 시에는 리플레이 버퍼를 초기화하고 학습 개시 스텝(Learning_starts)을 재설정하여, 이전 단계의 전이가 새 단계의 가치 학습에 혼입되는 것을 차단하고 새 전이가 충분히 축적된 후 갱신이 재개되도록 한다. 한편 행위자(actor) 256×256, 크리틱(Critic) 400×300의 분리 신경망 구조는 S0~S3 전 구성에 공통으로 적용하여 절제 축에서 제외하였다. 표 2는 기준 알고리즘 PPO와 최종 재설계 SAC(S3)의 설정을 비교한 것이다.

표 2. PPO와 재설계 SAC(S3)의 설정 비교
Table 2. Configuration comparison between PPO and redesigned SAC(S3)

Item	PPO (baseline)	SAC (redesigned, S3)
Action space	Discrete(10)	Box(−1,1)² → discrete map
Learning	on-policy	off-policy (buffer 5×10^5)
Normalization	none	VecNormalize (obs/reward)
Reward signal	R1~R15 original	R1~R15 + potential shaping
Curriculum shift	env replacement	env replacement + buffer reset
Network	shared 512-256-128	actor 256×256 / critic 400-300
Entropy coef.	0.01	0.01 (fixed)

재설계에 사용된 주요 계수의 선정 근거는 다음과 같다. 엔트로피 계수는 자동 온도 조정에 의한 과잉 탐색(메커니즘 ③)을 차단하기 위해 기준 알고리즘 PPO와 동일한 0.01로 고정하여 알고리즘 간 탐색 강도 조건을 일치시켰다. 잠재함수 계수(10, 10, 5, 5)는 보상 함수의 계층 설계 원칙(안정성·효율성 가중 균형)과 정합되도록, 성형 보상이 스텝 보상 스케일(±15 내외)을 과도하게 왜곡하지 않는 범위에서 설정하였다. 리플레이 버퍼 크기(5×10^5), Learning_starts(10^4) 등 나머지 하이퍼파라미터는 Stable-Baselines3 기본값을 기점으로 예비 실험에서 소폭 조정한 값이며, 특정 알고리즘에 유리한 대규모 탐색은 수행하지 않았다(표 3).

절제 구성을 완전 요인(Factorial) 설계가 아닌 누적 방식으로 구성한 이유는 재설계 요소들이 위계적 의존 관계를 갖기 때문이다. 행동 매핑(S1)이 없으면 표준 SAC는 이산 환경에서 유의미한 정책 학습 자체가 어렵고, 관측·보상 정규화가 없는 상태에서 잠재함수 기반 보상 성형을 적용하면 보상 스케일이 불안정해진다. 따라서 3.2절의 메커니즘 진단 순서에 따라 요소를 누적 투입하였으며, 각 구성 간 성능 차이는 투입 순서에 의존하는 누적 기여로 해석된다. 절제 구성별 추가 요소와 대응 메커니즘은 다음과 같다. 기준 구성 S0은 표준 SAC에 1차원 출력의 반올림 이산화만 적용한 것이다. S1은 S0에 연속 2차원 행동 매핑(메커니즘 ① 대응)을 추가하고, S2는 S1에 관측·보상 정규화와 엔트로피 계수 고정(메커니즘 ③ 대응)을 추가하며, S3은 S2에 잠재함수 기반 보상 성형과 커리큘럼 전환 시 버퍼 초기화(메커니즘 ② 대응)를 추가한 최종 재설계 구성이다. 인접 구성 간 성능 차이는 해당 재설계 요소의 한계 기여로 해석되며, 이를 통해 2장에서 제기한 연구 질문, 즉 각 실패 메커니즘의 상대적 비중을 정량적으로 규명한다.

3.4 평가 지표

알고리즘-환경 적합성을 다각도로 진단하기 위해 배치 품질 지표를 효율성 계열과 안정성 계열로 구분하여 정의한다. 효율성 계열은 유효배치율(VPR, Valid Placement Rate; 전체 배치 시도 중 유효 적재 비율), 오버스토우율(OSR, Overstow Rate; 먼저 양하할 컨테이너 위에 늦게 양하할 컨테이너가 적재된 비율), 수직 단일 POD 적재율(PSR, Pure Single-POD Rate; 1개 이상 적재된 행 중 단일 POD로만 구성된 행의 비율이며 빈 행은 분모에서 제외)로 구성된다. PSR은 동일 POD를 수평 밴드로 분산 배치하는 설계 원칙(R13, R15)에 위배되는 수직 적재를 포착하는 지표로, OSR·CWVR과 마찬가지로 낮을수록 우수하다. 안정성 계열은 무게균형지수 WBI = max(0, 1−CV)와 열 무게제약 위반율 CWVR로 구성된다. 여기서 CV는 1개 이상 적재된 열별 누적 무게의 변동계수이며(WBI가 1에 가까울수록 균형), CWVR은 전체 배치 시도 중 배치 직후 해당 열의 누적 무게가 160MT를 초과한 배치의 비율이다. 학습 동역학 측면에서는 에피소드 보상 이동평균이 최종 안정 구간의 95%에 최초 도달하는 수렴 스텝 수와 평가 에피소드 간 표준편차를 함께 보고한다. 모든 지표는 학습 종료 후 결정론적 정책으로 산출하며(본문 보고 기준인 Lv4는 300개, 중간 레벨 진단은 30개 평가 에피소드), 평가 에피소드의 난수 시드를 알고리즘 간 공유하여 동일 화물 목록에 대해 비교한다. 한편 본 환경은 선적 순서가 POD 내림차순으로 고정되어 학습된 정책 하에서 오버스토우가 구조적으로 발생하기 어려우므로, OSR은 예외 상황 점검용 보조 지표로 사용한다.

Ⅳ. 실험 설계 및 결과 분석

4.1 실험 설계

실험은 3.3절의 누적 절제 설계에 따른 SAC 4개 구성(S0~S3)과 기준 알고리즘 PPO, 진단 대조군 DQN의 총 6개 구성으로 수행하였다. 모든 구성은 동일한 커리큘럼 에피소드 예산(단계별 20,000/ 25,000/ 32,000/ 48,000 에피소드, 총 8,068,000 스텝)과 동일한 환경 난수 시드로 학습하였으며, Python 3.10, PyTorch 2.3.1(CUDA 12.1), Gymnasium 0.29, Stable-Baselines3 2.3[20] 기반으로 구현하여 VESSL 클라우드의 NVIDIA A100 SXM(80GB) 및

L40S(48GB) 인스턴스에서 실행하였다. 본 연구에 사용한 환경·학습·평가 코드와 설정 파일은 요청 시 제공 가능하다. 6개 구성의 학습에는 인스턴스 가동 기준 총 약 57시간(A100 약 34시간, L40S 약 23시간)이 소요되었다. 복수 구성을 동일 인스턴스에서 병렬 실행하였으므로 구성별 학습 시간의 분리 산출은 어려워 총 소요 시간만 참고 정보로 보고하며, 알고리즘 간 계산 비용의 비교는 구성별 네트워크 파라미터 수와 컨테이너 1건당 추론 시간을 기준으로 4.4절에서 제시한다. 표 3은 세 알고리즘의 주요 하이퍼파라미터를 비교한 것이다. 본 실험의 모든 결과는 단일 시드 기준이므로 이하의 비교는 통계적 검정을 수반하지 않는 경향적 관찰로 해석되어야 하며, 복수 시드 기반 검증은 향후 과제로 한다. 구성당 8,068,000 스텝의 학습이 요구되어 전 구성에 대한 복수 시드 반복 학습에는 상당한 계산 자원이 소요되므로, 본 연구에서는 학습 시드를 단일로 고정하는 대신 평가 단계의 변동성을 300개 평가 에피소드에 대한 10,000회 재표본 bootstrap 95% 신뢰구간으로 정량화하였으며, 주요 비교는 신뢰구간의 중첩 여부를 기준으로 해석하였다. 학습 완료 후 결정론적 정책으로 Lv4 기준 구성별 300개 평가 에피소드를 수행하되(중간 레벨 진단은 30개), 평가 시드를 모든 구성이 공유하여 동일한 화물 목록에 대해 비교하였고, 주요 지표에는 10,000회 재표본 bootstrap 95% 신뢰구간을 함께 보고한다. 평가는 최종 시점 모델을 사용하였다. 하이퍼파라미터는 Stable-Baselines3 기본값을 기점으로 예비 실험에서 소폭 조정한 값이며(표 3), 알고리즘별 대규모 탐색은 수행하지 않았다. 한편 PPO, DQN, SAC는 학습 구조와 갱신 방식, 하이퍼파라미터가 서로 다르므로 동일한 환경 절차만으로 모든 학습 조건이 동등해지는 것은 아니다. 본 연구의 비교는 각 알고리즘의 최적 성능 경쟁이 아니라, 동일 학습 예산(총 8,068,000 스텝), 동일 커리큘럼·환경 난수 시드, 동일 보상 함수, 동일 평가 프로토콜이라는 공통 운영 조건 하에서의 환경 적합성 진단을 목적으로 한다. 갱신 빈도, 배치 크기, 네트워크 용량의 차이는 알고리즘 고유 설계의 일부로 보존하였으며, 동일 환경 스텝이 동일한 정책 갱신 기회를 의미하지는 않으므로 결과 해석 시 이 점을 유의해야 한다. 이러한 구조적 차이에 따른 교란을 줄이기 위해 동일한 off-policy 학습과 리플레이 버퍼 구조를 갖는 DQN을 대조군으로 두어, 관찰된 성능 저하가 off-policy 학습 일반의 문제인지 연속 행동 표현에 기인한 SAC 고유의 문제인지를 분리하고자 하였다. 보상 성형을 사용하는 S3의 학습곡선 기록과 평가 보상은 모두 성형 전 원 보상 기준으로 산출하여 비교의 공정성을 확보하였다. 비학습 기준선으로는 균등 랜덤 배치와 휴리스틱 2종을 동일 평가 프로토콜로 측정하였다. 최근접 휴리스틱은 오버스토우를 만들지 않는 가장 왼쪽 행을 선택하고, 무게중심 휴리스틱은 열 무게제한을 고려하면서 배치 후 무게중심 편차가 최소가 되는 행을 선택한다. 이하의 결과는 최대 규모인 Lv4 (10×10, 100개)를 기준으로 보고한다.

표 3. 세 알고리즘별 주요 하이퍼파라미터
Table 3. Hyperparameters of the three algorithms

Item	PPO	DQN	SAC
Learning rate	3e-4	1e-4	3e-4
Batch size	64	64	256
Discount γ	0.99	0.99	0.99
Buffer size	-	5×10^5	5×10^5
Rollout/GAE λ	2048/0.95	-	-
Epochs/clip	10/0.2	-	-
Train freq.	-	4 steps	1 step
Target update	-	10^4 steps	τ=0.005
Exploration	ent 0.01	ε 1→0.05	auto (S0-S1) /0.01 (S2-S3)
Learning starts	-	10^4	10^4
Network	512-256-128	512-256-128	π256^2, Q400-300
Optimizer/steps	Adam/8.068M	Adam/8.068M	Adam/8.068M

본 실험의 결과는 단일 베이, 40ft 일반 컨테이너, 10~20MT 무게 범위, 6-POD, 최대 10×10 규모의 행 할당 축소 문제라는 조건 하에서 관찰된 것으로, 리퍼·위험물·규격외 화물, 다중 베이 간 상호작용, 크레인 스케줄링 등 실제 운영 제약은 포함하지 않는다.

4.2 알고리즘별 학습 성능 비교

그림 2는 6개 구성의 커리큘럼 학습곡선을, 표 4는 Lv4 최종 KPI를 보여준다. 이산 행동공간을 직접 지원하는 두 알고리즘인 PPO와 DQN은 모두 안정적으로 수렴하였다. PPO는 에피소드 보상 1377.0 (95% CI [1372.4, 1381.5])로 최고 성능을 기록하였고, DQN은 VPR 0.997, 보상 1351.4로 근소한 차이의 차선 성능을 보였다. 수렴 속도는 PPO가 3.83M 스텝으로 DQN(6.31M)보다 약 1.6배 빨랐다. 한편 두 휴리스틱은 유효배치율 1.000을 달성했음에도 총 보상이 낮은데, 이는 유효 배치 외의 다목적 항목(중량 균형, POD 배치, 무게 역전 등)을 반영하지 못하기 때문이다. 특히 최근접 휴리스틱은 같은 POD를 같은 행에 연속 적재하는 규칙 특성상 수직 단일 POD 적재(PSR 0.552)에 따른 감점(R15)과 열 무게제한 위반(CWVR 0.034)이 중첩되어 음의 보상을 기록하였다. 한편 30개 에피소드 예비 평가에서 관찰되었던 랜덤 배치와 무게중심 휴리스틱 간 순위 역전은 300개 평가에서 소멸하여 두 기준선이 사실상 동일한 보상(374.8 대 374.7)으로 수렴하였는데, 이는 소표본 평가의 불안정성을 보여주는 동시에 단일 규칙 기반 기법이 다목적 보상 구조에서 갖는 한계를 재확인한다. 학습 기반 정책이 이들 기준선(최대 374.8)을 3배 이상 상회한 결과는 다목적 보상 구조 하에서 강화학습이 갖는 이점을 보여준다.

반면 표준 SAC(S0)는 소규모 단계에서는 정상적으로 학습하였으나(Lv1 VPR 0.992, Lv2 0.984) 문제 규모가 커질수록 성능이 저하되어(Lv3 0.928), 최대 규모인 Lv4에서 VPR 0.274, 보상 -560.1(95% CI [-577.9, -541.7])로 사실상 붕괴하였다. 평가 에피소드 간 표준편차도 S0 158.2, S1 201.6으로 정상 수렴 구성(38.1~66.5)의 2배를 넘어 정책 불안정성이 뚜렷하였다. 여기서 DQN 대조군의 역할이 확인된다. 동일한 off-policy 학습과 리플레이 버퍼를 사용하는 DQN이 전 단계에서 정상 작동했으므로, S0의 붕괴는 off-policy 학습 일반의 문제가 아니라 연속 행동공간용 알고리즘을 이산 환경에 적용하는 데서 발생하는 문제로 좁혀진다. 즉 S0의 붕괴는 off-policy 학습 자체만으로는 설명되기 어려우며, 연속 행동 표현과 그 학습 구조가 주요 원인일 가능성이 시사된다. 다만 DQN은 가치·정책 구조, 탐색 방식, 갱신 규칙 등에서도 SAC와 다르므로 완전한 통제 대조군은 아니며, 이 귀속은 보수적으로 해석되어야 한다.

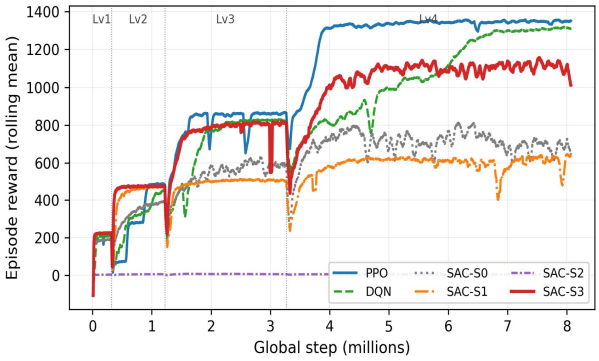

그림 2. 커리큘럼 학습곡선 비교(6개 구성)
Fig. 2. Curriculum learning curves of six configurations

표 4. Lv4 최종 성능: 학습 분포 및 실측 경험분포 재표본화 평가
Table 4. Final performance at Lv4 under training and empirically resampled distributions

Config	VPR	WBI	PSR	CWVR	Return	Emp. R
PPO	0.999	0.959	0.000	0.007	1377.0	920.9
DQN	0.997	0.952	0.000	0.010	1351.4	533.1
SAC-S0	0.274	0.096	0.609	0.004	-560.1	-545.4
SAC-S1	0.649	0.606	0.085	0.007	-157.4	-381.1
SAC-S2	0.971	0.926	0.111	0.008	860.3	-318.9
SAC-S3	0.962	0.920	0.000	0.011	1131.6	-617.3
Random	0.880	0.801	0.000	0.010	374.8	-
Nearest heur.	1.000	0.897	0.552	0.034	-275.5	-
CoG heur.	1.000	0.953	0.121	0.004	374.7	-

※ OSR omitted (0.000 for all configs; see Section 3.4). PSR: vertical single-POD stacking ratio (lower is better). Emp. R (empirical-resampling return): mean raw return over episodes resampled from the empirical cargo distribution (Lv4, 300 episodes)

4.3 SAC 재설계 절제 실험 분석

그림 3과 표 5는 누적 절제 구성에 따른 단계별 성능 변화를 보여준다. Lv4 보상 기준으로 S0

(-560.1)에서 S1(-157.4), S2(860.3), S3(1131.6)으로 점진적으로 증가하여, 세 재설계 요소군이 모두 양의 기여를 가짐이 관찰된다. 구성 간 누적 기여는 연속 2차원 행동 매핑(S0→S1)이 +402.7, 학습신호 안정화 요소군인 관측·보상 정규화와 엔트로피 계수 고정(S1→S2)이 +1017.7, 잠재함수 보상 성형과 커리큘럼 버퍼 초기화(S2→S3)가 +271.3이었다. 즉 학습신호 안정화 요소군(메커니즘 ③ 대응)이 전체 회복분의 약 60%를 차지하여, SAC 실패의 주원인이 행동공간 변환 그 자체보다 이산 다목적 보상 구조에서의 신호 스케일 왜곡에 있을 가능성을 시사한다. 다만 누적 절제 설계의 특성상 각 수치는 투입 순서에 의존하는 누적 기여이며 S2에는 세 요소(관측 정규화, 보상 정규화, 엔트로피 고정)가 동시에 포함되므로, 요소별 독립 기여의 분리는 향후 과제이다. 따라서 본 실험의 결과는 전체적인 성능 회복 효과와 요소군 단위의 누적 기여를 보여주는 것으로 해석해야 하며, 개별 요소의 독립적 효과를 규명하기 위해서는 요소 단위의 절제 실험이나 요인(Factorial) 설계에 기반한 추가 검증이 필요하다. 행동 매핑만 적용한 S1이 소규모 단계에서는 완전한 성능(Lv1~2 VPR 1.000)을 보이다가 Lv4에서 붕괴(0.649)하는 양상은, 신호 왜곡의 영향이 문제 규모와 함께 증폭됨을 보여준다.

부가적으로 세 가지 관찰이 주목된다. 첫째, S2는 0.56M 스텝 만에 최종 플래토의 95%에 도달하여 전 구성 중 가장 빠른 수렴을 보였다(PPO 3.83M, DQN 6.31M). 정규화와 엔트로피 고정만으로 off-policy 학습의 샘플 효율이 회복됨을 의미한다. 둘째, S2→S3 구간의 개선(+271.3)은 KPI상 VPR과 WBI가 거의 동일한 가운데 수직 단일 POD 적재율(PSR 0.111 → 0.000)에서 발생하여, 보상 성형과 버퍼 초기화의 이득이 주로 수평 밴드 배치의 완성에 기여했음을 보여준다. 셋째, 최종 재설계 S3은 보상 기준 PPO의 약 82% 수준까지 회복하였으나(bootstrap 95% CI 기준 S3 [1124.0, 1139.0]과 PPO [1372.4, 1381.5]는 겹치지 않음) VPR(0.962)과 WBI(0.920)에서는 여전히 PPO에 미치지 못하였다. 요컨대 본 실험 환경에서 재설계된 SAC는 경쟁력 있는 대안이 되지만, 이산 배치 문제에서 이산 전용 알고리즘의 우위를 역전시키지는 못하였다. 이상으로 2장의 연구 질문 RQ1(실패 원인의 구조적 분해)과 RQ2(재설계 요소군별 기여)에 대한 정량적 관찰이 확보되었다. 나아가 행동 선택 분포는 메커니즘 ①(행동공간 불일치)의 직접 증거를 제공한다(그림 4). Lv4 평가 300 에피소드(구성당 30,000회 선택) 기준으로 S0은 전체 행동의 81.7%가 중앙 행(4번)에 고착되고 경계 행(0번, 9번) 선택률이 3.6%에 그쳐(균등 기대치 20%), tanh 압착 정책의 중앙 편향과 경계 포화가 실제 행동 분포로 확인된다. 2차원 매핑을 도입한 S1은 경계 선택률이 13.2%로 회복되나 특정 행(1번, 39.1%)으로의 쏠림이 남고, 최종 재설계 S3은 행별 8.7~11.9%의 준균등 분포로 PPO의 균등 분포(행별 10.0%)에 근접한다. 즉 재설계의 진전에 따라 행동 분포의 왜곡이 단계적으로 해소되는 양상이 KPI 개선과 병행하여 관찰된다. 한편 행동 매핑 가중치(0.9/0.1) 선택의 영향을 사후 점검하기 위해, 학습된 최종 재설계(S3) 정책을 추론 단계에서 가중치 0.95/0.05와 0.8/0.2로 변경하여 동일 평가 프로토콜(Lv4, 300개 평가 에피소드)로 재평가하였다. 그 결과 보상은 기준 구성(1131.6) 대비 약 6% 낮은 1065.5와 1066.7, 유효배치율은 0.952와 0.948로 성능 붕괴 없이 유지되었고, 수직 단일 POD 적재율(PSR 0.000)과 경계 행 선택률(0.179~0.201)의 행동 특성도 보존되었다. 이는 절제 실험의 결과가 특정 가중치 값의 선택에 결정적으로 의존하지 않음을 시사한다. 다만 이는 추론 단계의 점검으로, 학습 단계에서의 가중치 민감도 분석은 향후 과제로 남는다.

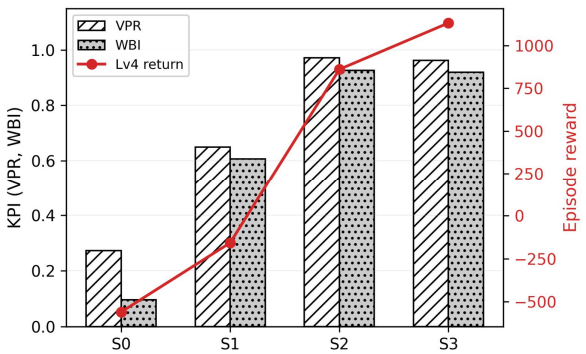

그림 3. SAC 재설계 절제 실험: 구성별 KPI 및 보상변화
Fig. 3. Ablation study of SAC redesign: KPI and return by configuration

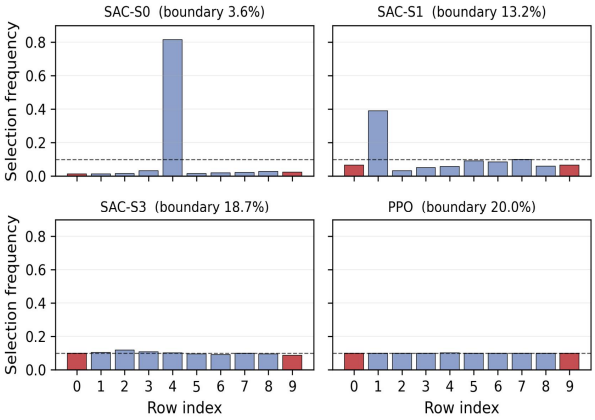

그림 4. 구성별 행 선택 분포와 경계 행 선택률(Lv4)
Fig. 4. Row-selection distributions and boundary-row rates (Lv4)

표 5. 절제 구성별 보상, 누적 기여, 수렴 스텝
Table 5. Reward, cumulative contribution, and convergence by configuration

Config	Lv4 return	Cum. gain	Conv. steps (95%)
S0 (standard+rounding)	-560.1	-	3.56M
S1 (+2D mapping)	-157.4	+402.7	4.11M
S2 (+norm., fixed ent)	860.3	+1017.7	0.56M
S3 (+shaping, buf. reset)	1131.6	+271.3	3.85M

※ Convergence steps for S0 and S1 indicate arrival at a low-performance plateau and are not directly comparable to converging configurations

4.4 안정성-효율성 특성과 실측 경험분포 재표본화 평가

그림 5는 표 4의 주요 구성별 5개 KPI를 시각화한 것으로, 알고리즘별 배치 전략의 특성 차이를 보여준다. PSR은 수직 단일 POD 적재율로 낮을수록 우수한 지표이다. PPO, DQN과 재설계 S3은 PSR 0.000으로 수직 단일 POD 적재 없이 무게 균형(WBI)과 유효 배치를 동시에 달성하여, R13과 R15가 유도하는 수평 밴드 배치 원칙이 학습 정책에 그대로 관철되었음을 보여준다. 반면 붕괴한 S0은 PSR 0.609로 적재 행의 6할이 수직 단일 POD 적재에 고착되어, 다목적 보상 신호를 통합적으로 학습하지 못한 채 국소적 배치 패턴에 수렴하는 실패 양상을 드러낸다. 최근접 휴리스틱의 PSR 0.552 역시 같은 POD를 연속하여 같은 행에 쌓는 순차 규칙이 수평 밴드 원칙과 충돌함을 보여준다. 이는 다목적 보상 하에서 학습 알고리즘과 규칙 기반 기법이 서로 다른 배치 패턴으로 수렴함을 시사하며, 운영 목표에 따라 보상 가중을 재조정할 수 있는 실무적 여지를 보여준다.

실측 경험분포 재표본화 평가(이하 재표본화 평가)는 다음 절차로 수행하였다. 3.1절의 필터 기준(리퍼, 위험물, 규격외 화물 제외)을 통과한 40ft 일반 컨테이너 13,174건의 실측 무게 집합에서 에피소드별 화물 무게를 복원추출로 재샘플링하고, POD는 8개 이상의 화물을 갖는 항차×베이 268개에서 관측된 POD 구성비 벡터 중 하나를 에피소드마다 무작위로 선택하여 6개 POD 라벨에 배정한 뒤 해당 비율로 생성하였다. 즉 실측 항차를 그대로 재생하는 것이 아니라 실측 경험분포에서 화물 목록을 재표본화하는 방식이므로, 무게-POD 결합분포나 항차별 특성 같은 실제 데이터의 상관구조는 보존되지 않으며 결과는 주변분포 변화에 대한 강건성으로 한정하여 해석해야 한다. 즉 본 평가는 실제 터미널 운영 환경 전반에 대한 일반화 성능이 아니라, 무게 분포의 변화에 대한 정책의 민감도를 검증하는 것으로 그 범위가 한정된다. 격자 크기와 학습된 정책, 정규화 통계는 변경하지 않았고, S2·S3의 관측 정규화는 학습 분포에서 고정된 통계로 평가 모드에서만 적용되며(관측 클리핑 ±10) 보상 정규화는 비활성화하여 원 보상을 보고한다. 평가는 Lv4에서 300개 에피소드로 수행하였고 평가 시드는 학습 분포 평가와 동일하게 전 구성이 공유하였다. 학습 분포와 실측 분포의 차이는 무게 기준으로 정량화된다(학습: 10~20MT 균등 / 실측: 평균 12.7MT, 표준편차 9.5MT, 범위 1.0~32.5MT, 10~20MT 구간 비율 22.1%). 실측 데이터는 운영사 비식별 처리 후 분석에만 사용하였으며, 운영사와의 협약에 따라 원자료 공개는 불가하고 익명 통계량 형태로만 제공 가능하다.

평가 결과 알고리즘 간 강건성 격차가 뚜렷하게 나타났다. PPO는 보상 1377.0에서 920.9(95% CI [911.4, 929.8])로 33% 하락에 그치며 VPR 0.938을

유지한 반면, DQN은 533.1로 61% 급락하였고(VPR 0.861, WBI 0.952→0.744), SAC 계열은 재설계 여부와 무관하게 전 구성이 음의 보상으로 하락하였다(S0 -545.4, S1 -381.1, S2 -318.9, S3 -617.3). 주목할 점은 SAC 계열의 성능 저하가 정규화 적용 여부와 무관하게 나타났다는 것으로(정규화가 없는 S0·S1도 마찬가지로 낮은 보상에 머물렀음), 이는 고정된 정규화 통계만으로는 설명되지 않으며 연속-이산 매핑 정책이 관측 특징의 스케일 변화에 민감하게 반응하는 것으로 추정되나 정확한 원인 규명은 향후 과제로 남긴다. 확인되는 사실은 on-policy로 학습한 PPO의 성능 하락이 가장 작았다는 점이다. 이 결과는 두 가지 실무적 함의를 갖는다. 첫째, 학습 분포 성능만으로 알고리즘을 선택하는 것은 위험하며, 실측 분포 강건성이 독립적인 선택 기준으로 평가되어야 한다. 둘째, 시뮬레이션 기반 학습 시스템은 실측 화물 분포로 환경을 보정하거나 분포 변화에 대한 재학습 체계를 갖출 필요가 있다. 정보시스템 적용 관점의 추론 비용도 측정하였다. 표 6은 구성별 네트워크 파라미터 수와 컨테이너 1건당 추론 시간을 보인다. 추론 시간은 단일 CPU 스레드에서 워밍업 50회 후 2,000회의 단건 추론으로 측정한 평균과 95백분위 값으로, 평균 기준 PPO 0.44ms, DQN 0.37ms, 재설계 SAC 0.56ms(95백분위 1.1ms 이내)이며, 네트워크 파라미터 수는 PPO·DQN 약 41만 개, SAC 계열 약 70만 개이다. 100개 컨테이너의 배치 계획 1건 생성에 세 알고리즘 모두 0.1초가 걸리지 않는다. 따라서 추론 비용은 알고리즘 선택의 변별 요인이 되지 못하며, 배치 품질과 분포 강건성이 지배적 기준이 된다.

표 6. 알고리즘·구성별 계산 비용: 파라미터 수와 단건 추론 시간

Table 6. Computational cost by configuration: parameters and per-decision inference latency

Config	Parameters	Mean(ms)	p95(ms)
PPO	411,787	0.44	0.81
DQN	412,948	0.37	0.73
SAC-S0	698,790	0.34	0.70
SAC-S1	700,904	0.55	1.03
SAC-S2	700,904	0.56	0.95
SAC-S3	700,904	0.56	1.09

이상을 종합하면 RQ3에 대한 답으로, 본 실험 환경에서는 PPO가 배치 품질, 수렴 속도, 분포 강건성 모두에서 가장 안정적인 기준 알고리즘으로 관찰되었고, DQN은 학습-운영 분포가 고정된 환경에서 유효한 대안, 재설계 SAC는 연속 제어 모듈과의 단일 스택 통합이 요구되고 분포 정합이 관리되는 조건에서 고려할 수 있는 선택지로 정리된다.

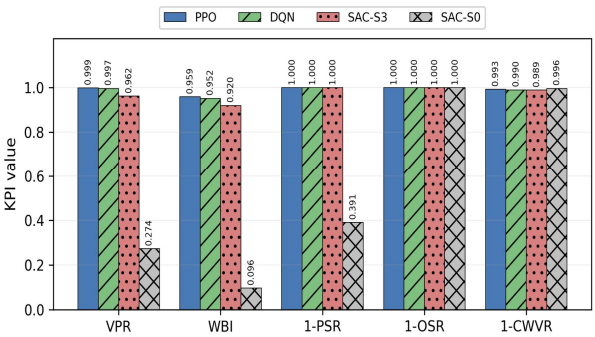

그림 5. 알고리즘별 KPI 비교: 안정성-효율성 특성(Lv4)

Fig. 5. KPI comparison across algorithms: stability-efficiency characteristics (Lv4)

V. 결론 및 향후 과제

본 논문은 컨테이너 터미널 적재계획 정보시스템의 관점에서 심층강화학습 알고리즘의 이산 슬롯 배치 환경 정합성을 분석하였다. 단순한 성능 순위 비교 대신, 표준 SAC의 성능 저하를 세 가지 구조적 메커니즘으로 분해하는 진단형 접근을 채택하고, 각 메커니즘에 대응하는 재설계 요소를 누적하는 절제 실험과 이산 전용 off-policy 대조군(DQN), 그리고 실측 적하 데이터 기반 분포 평가를 결합하여 원인-처방-검증의 전 과정을 수행하였다.

연구 질문에 대한 관찰 결과는 다음과 같다. 첫째(RQ1: 표준 SAC 성능 저하의 원인), 표준 SAC는 최대 규모 문제(10×10, 100개)에서 유효배치율 0.274로 붕괴한 반면 동일한 off-policy 구조의 DQN은 0.997로 정상 작동하여, 본 실험 환경에서 실패는 off-policy 학습 자체만으로는 설명되기 어려우며 연속 행동 표현과 그 학습 구조가 주요 원인일 가능성이 시사되었다. 둘째(RQ2: 재설계 요소군의 상대적 기여), 절제 실험에서 학습신호 안정화 요소군(관측·보상 정규화와 엔트로피 계수 고정)의 누적

기여가 전체 회복분의 약 60%로 가장 컸고, 행동 매핑과 버퍼 관리가 이를 보완하여 최종 재설계 SAC는 보상 기준 PPO의 약 82%(유효배치율 0.962)까지 회복하는 경향을 보였다. 셋째(RQ3: 실측 화물 분포에서의 성능 유지), 실측 경험분포 재표본화 평가에서 PPO는 성능 하락이 -33%에 그친 반면 DQN은 -61%로 급락하고 SAC 계열은 전 구성이 크게 하락하여, 운영 분포에 대한 강건성이 학습 분포 성능과 독립적인 선택 기준이 될 수 있음이 시사한다. 이를 종합하면, 본 실험 환경에서는 PPO가 배치 품질, 수렴 속도, 분포 강건성 모두에서 가장 안정적인 기준 알고리즘으로 관찰되었고, DQN은 학습-운영 분포가 고정된 환경에서 유효한 대안이며, SAC는 연속 제어 모듈과의 단일 알고리즘 스택 통합이 요구되는 시스템에서 본 논문의 재설계와 분포 정합 관리를 전제로 고려할 수 있다. 다만 이 지침은 단일 시드·단일 베이 조건의 관찰에 기반한 경향적 제안임을 밝혀 둔다.

본 연구의 실무적 함의는 알고리즘 선택이 벤치마크 성능 순위가 아니라 환경 적합성, 분포 강건성, 시스템 통합성의 세 축에서 판단되어야 한다는 점이다. 특히 실측 데이터 기반 환경 캘리브레이션은 시뮬레이션 학습 결과의 현장 이전 가능성을 좌우하는 요소로, 본 논문이 수행한 3개월치 적하 데이터 검증은 그 실행 가능한 절차를 예시한다.

본 연구의 한계와 향후 과제는 다음과 같다. 첫째, 실험은 단일 베이, 40ft 일반 컨테이너, 단일 학습 시드 조건에서 수행되었으므로, 본 논문의 결론은 해당 조건 하의 경향적 관찰로 한정된다. 강화학습의 확률적 변동성을 반영하기 위한 복수의 독립 학습 시드 기반 평균·표준편차 및 통계적 검증, 그리고 다중 베이, 다양한 컨테이너 유형(리퍼·위험물·규격외), 크레인 스케줄링 등 실제 운영 제약을 포함하는 환경으로의 확장 검증이 필요하다. 둘째, 누적 절제 설계는 요소 투입 순서에 의존하므로 요소별 독립 기여의 분리(요인 설계)와 행동 매핑 가중치(0.9/0.1)의 민감도 분석이 요구된다. 셋째, 재표본화 평가에서의 SAC 계열 성능 저하 원인이 완전히 규명되지 않았으며, 정규화 통계 재적합과 분포 변화 적응 기법을 결합한 후속 분석이 필요하다. 넷째, 실제 터미널의 적하 이력을 활용한 오프라인 강화학습과 본 연구의 온라인 학습 정책 간 비교, 그리고 적재계획 품질지표와 터미널 생산성 간의 실증적 연계 분석을 향후 연구과제로 제시한다.

References

[1] Y. S. Choi, "A study on the application of artificial intelligence logic in port operating systems", Journal of Shipping and Logistics, Vol. 38, No. 1, pp. 101-118, Mar. 2022. https://doi.org/10.37059/tjosal.2022.38.1.101.

[2] J. van Twiller, A. Sivertsen, D. Pacino, and R. M. Jensen, "Literature survey on the container stowage planning problem", European Journal of Operational Research, Vol. 317, No. 3, pp. 841-857, Sep. 2024. https://doi.org/10.1016/j.ejor.2023.12.018.

[3] J. van Twiller, Y. Adulyasak, E. Delage, D. Grbic, and R. M. Jensen, "Navigating demand uncertainty in container shipping: Deep reinforcement learning for enabling adaptive and feasible master stowage planning", arXiv preprint, arXiv:2502.12756, Feb. 2025. https://doi.org/10.48550/arXiv.2502.12756.

[4] J. van Twiller, D. Grbic, and R. M. Jensen, "AI2STOW: End-to-end deep reinforcement learning to construct master stowage plans under demand uncertainty", arXiv preprint, arXiv:2504.04469, Apr. 2025. https://doi.org/10.48550/arXiv.2504.04469.

[5] W. S. Jang, H. J. Lee, M. S. Seo, S. J. Lee, and D. G. Kim, "Reinforcement learning-based container multi-stage loading modelling method in port logistics", Journal of KIIT, Vol. 21, No. 4, pp. 117-124, Apr. 2023. https://doi.org/10.14801/jkiit.2023.21.4.117.

[6] J. H. Cho, J. K. Kim, J. Y. Lee, B. S. Yoo, and N. K. Ku, "Container loading planning to minimize the number of rehandling using DQN",

Korean Journal of Computational Design and Engineering, Vol. 29, No. 4, pp. 310-322, Dec. 2024. https://doi.org/10.7315/CDE.2024.310.

[7] Y. Huang, N. Chennakeshava, A. Carras, V. Neverov, W. Liu, A. Plaat, and Y. Fan, "A benchmark study of deep reinforcement learning algorithms for the container stowage planning problem", arXiv preprint, arXiv:2510.02589, Oct. 2025. https://doi.org/10.48550/arXiv.2510.02589.

[8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor", Proc. of ICML 2018, Stockholm, Sweden, pp. 1861-1870, Jul. 2018.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement learning", Nature, Vol. 518, No. 7540, pp. 529-533, Feb. 2015. https://doi.org/10.1038/nature 14236.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms", arXiv preprint, arXiv:1707.06347, Jul. 2017. https://doi.org/10.48550/arXiv.1707.06347.

[11] H. Choi, B. W. On, and D. Jeong, "A survey on container loading strategies at ports", Proc. of 2025 KIIT Summer Conference, Jeju, Korea, pp. 700-704, Jun. 2025.

[12] H. J. Lee, W. S. Jang, S. J. Lee, and D. G. Kim, "Optimal dispatch modeling method based on multi-agent reinforcement learning in port logistics", Journal of KIIT, Vol. 21, No. 6, pp. 1-7, Jun. 2023. https://doi.org/10.14801/jkiit.2023.21.6.1.

[13] W. Kang, D. Seo, and M. Sa, "Reinforcement learning-based AI model for loading optimization", Journal of the Korea Institute of Information and Communication Engineering, Vol. 28, No. 1, pp. 33-39, Jan. 2024. https://doi.org/10.6109/jkiice. 2024.28.1.33.

[14] P. Christodoulou, "Soft actor-critic for discrete action settings", arXiv preprint, arXiv:1910.07207, Oct. 2019. https://doi.org/10.48550/arXiv.1910.07207.

[15] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping", Proc. of ICML 1999, Bled, Slovenia, pp. 278-287, Jun. 1999.

[16] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey", Journal of Machine Learning Research, Vol. 21, No. 181, pp. 1-50, Jul. 2020. https://doi.org/10.48550/arXiv.2003.04960.

[17] J. H. Kim, M. S. Kim, and M. S. Kim, "Research on attack techniques for cyber-range simulations based on deep reinforcement learning", Journal of KIIT, Vol. 24, No. 2, pp. 43-53, Feb. 2026. https://doi.org/10.14801/jkiit.2026.24.2.43.

[18] G. Dulac-Arnold, R. Evans, H. van Hasselt, et al., "Deep reinforcement learning in large discrete action spaces", arXiv preprint, arXiv:1512.07679, Dec. 2015. https://doi.org/10.48550/arXiv.1512.07679.

[19] T. Johannink, S. Bahl, A. Nair, et al., "Residual reinforcement learning for robot control", Proc. of IEEE International Conference on Robotics and Automation (ICRA), Montreal, Canada, pp. 6023-6029, May 2019. https://doi.org/10.1109/ ICRA.2019.8794127.

[20] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations", Journal of Machine Learning Research, Vol. 22, No. 268, pp. 1-8, Nov. 2021.

저자소개

유 홍 성 (Hongsung Yoo)

1997년 2월 : 인하대학교
경영학과(경영학석사)
2006년 2월 : 인하대학교
회계학과(경영학박사)
2026년 8월 : 서울과학종합대학원
대학교 AI·빅데이터학과
(공학석사)
2025년 9월 ~ 현재 : 인하대학교 경영대학 IBS
국제화센터 산학연구교수
관심분야 : 강화학습, 빅데이터, 설명가능한 AI, Edge AI

오 태 연 (Taeyeon Oh)

2008년 2월 : 서강대학교
수학과(이학사)
2026년 2월 :
서울과학종합대학원대학교
AI전문대학원(공학박사)
2022년 3월 ~ 현재 :
서울과학종합대학원대학교
AI·빅데이터학과 조교수
관심분야 : 데이터사이언스, 설명가능한 AI, Agentic AI