Paper Review
Reinforcement Learning

RILS LAB

Robot Intelligence Learning & System

정선웅

2026. 08. 07

Published as a conference paper at ICLR 2026

01

DR-SAC: DISTRIBUTIONALLY ROBUST SOFT ACTOR-CRITIC FOR REINFORCEMENT LEARNING UNDER UNCERTAINTY

Mingxuan Cui*, **Duo Zhou***†
Yuxuan Han[1], **Grani A. Hanasusanto, Qiong Wang, Huan Zhang, Zhengyuan Zhou**[1]
University of Illinois Urbana-Champaign [1]New York University.

Preliminaries

DR-RL: Distributionally Robust Reinforcement Learning.

하나의 전이 분포를 가정하지 않고, 가능한 전이 분포들의 집합을 고려함.

$$P \in \mathcal{P}(\delta)$$

이 때 최악의 경우에도 좋은 정책을 선택하는 강화학습 알고리즘.

$$\max_{\pi} \min_{P \in \mathcal{P}(\delta)} J(\pi, P)$$

Motivation

Offline Deep RL의 한계: OOD에 약함

기존 DR-RL의 한계: Tabular 또는 Value-based 방식에 한정

- 연속 공간에 적용 가능한 유일한 offline DR-RL인 RFQI(Robust Fitted Q-Iteration)
- Total Variation(TV) 거리를 사용해 고차원 행동에서 어려움을 겪고, 학습 속도를 낮추는 등 성능에 한계가 있다.

$$D_{TV}(P,Q) = \frac{1}{2}\sum_{x}|P(x) - Q(x)|$$

Contribution

1. DR Maximum Entropy Framework
   - 불확실성 집합(Ambiguity set) 하에서 최악의 시나리오에도 이득을 최대화하는 DR Soft Policy Iteration을 수학적으로 정립하고 수렴함을 증명
2. Scalable Reformulation
   - 기존에는 모든 (s,a) 마다 내부 스칼라 최적화 문제(β)를 개별적으로 풀어야 해서 계산량이 많음
   - 단일 함수 최적화 문제 (g(s,a))로 전환해 계산량을 감소
   - RFQI 대비 학습 시간을 80% 이상 감소
3. VAE를 활용한 Nominal Distribution(공칭 분포) 추정
   - 오프라인 RL에서는 공칭 전이 분포(기준 분포)를 알 수 없음
   - VAE로 공칭 전이 분포를 추정하고, next-state sample을 생성함

Published as a conference paper at ICLR 2026

02

POLICYFLOW: POLICY OPTIMIZATION WITH CONTINUOUS NORMALIZING FLOW IN REINFORCEMENT LEARNING

Shunpeng Yang[1,4], **Ben Liu**[2] **& Hua Chen**[3,4†]
[1]Hong Kong University of Science and Technology, [2]Southern University of Science and Technology
[3]Zhejiang University-University of Illinois Urbana-Champaign Institute, [4]LimX Dynamics
[†] Corresponding author, huachen@intl.zju.edu.cn

Policy Flow

PPO에 Continuous Normalizing Flow, CNF 모델을 결합

복잡한 행동 분포를 가볍고 안정적으로 학습시키는 on-policy RL 알고리즘

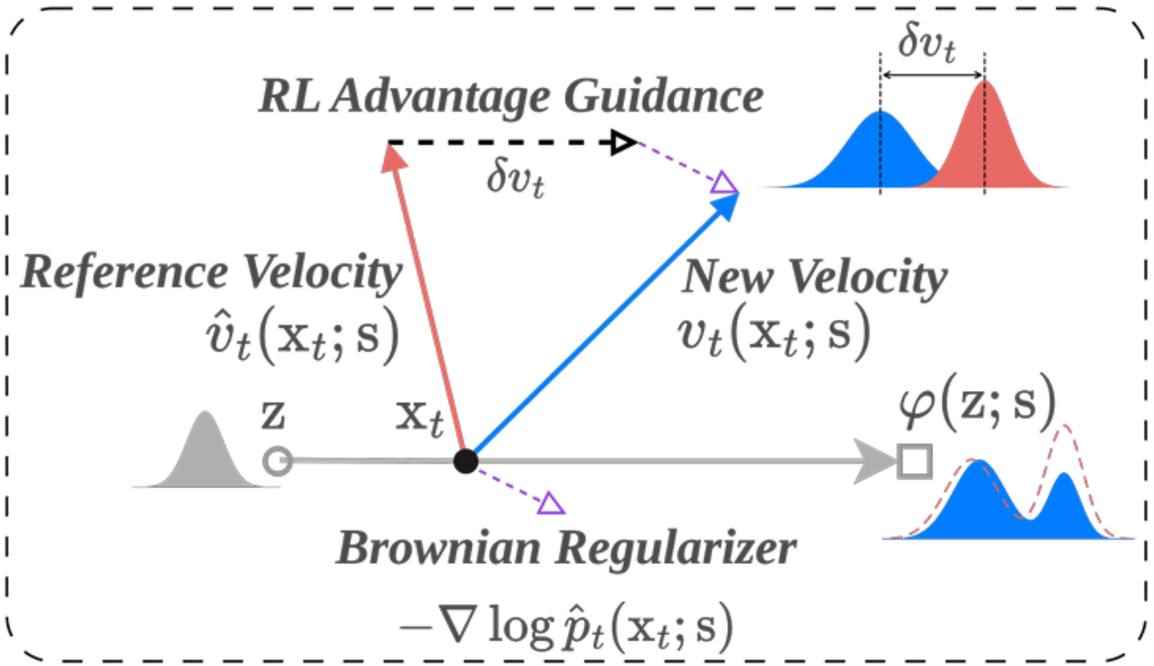
Motivation

Gaussian 정책의 한계. PPO는 분포를 gaussian 으로 모형화하기 때문에, multimodal의 행동 분포는 표현할 수 없음

생성 모델(CNF, Diffusion)을 PPO에 적용 시 발생하는 문제

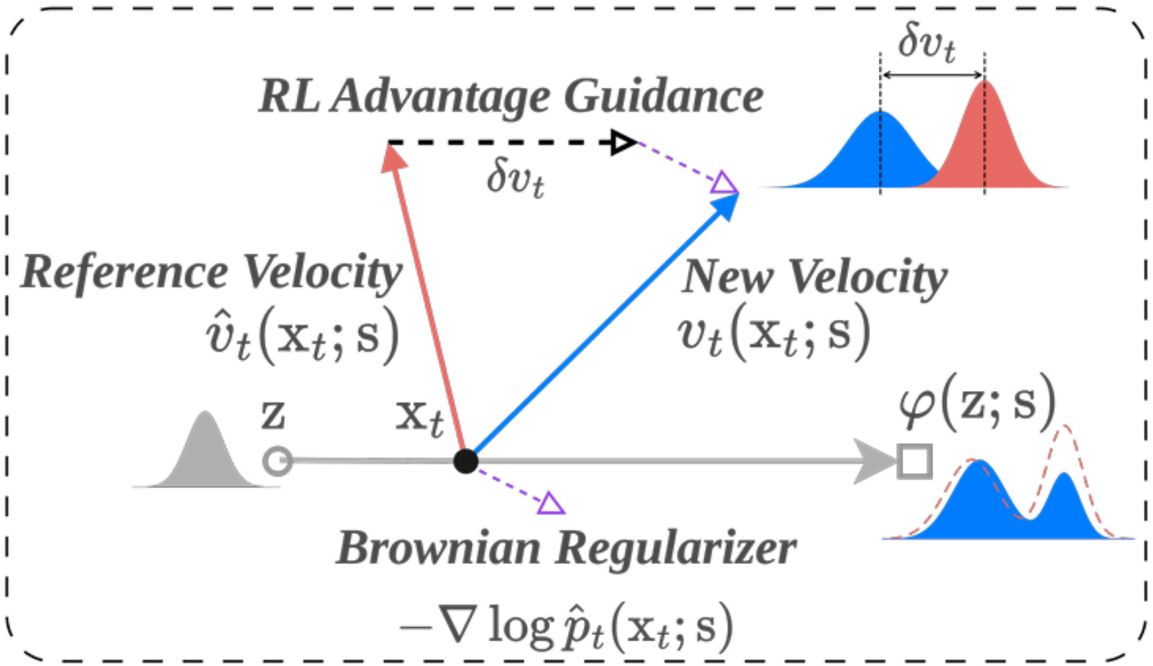
- ODE, SDE를 역전달해야 함 → 계산 비용 증가, gradient 폭발, 소실 문제 → 학습이 불안정

기존 연구(FPO, DPPO)는 비대칭적 근사 편향 or 엔트로피 규제 부족 → Mode Collapse



Contribution

1. Importance Ratio Approximation
   - 전체 ODE 경로로 gradient 전파 대신 단순 선형 보간 경로(Interpolation path) 설정.
2. Brownian Regularizer
   - 브라운 운동의 열 방정식(Heat Equation)에서 착안
   - 속도장이 Negative Score(엔트로피 증가 방향)를 따르도록 유도하는 경량화된 정규화 손실 → 우도 없이 모드 붕괴를 막고, 탐색을 촉진함



03

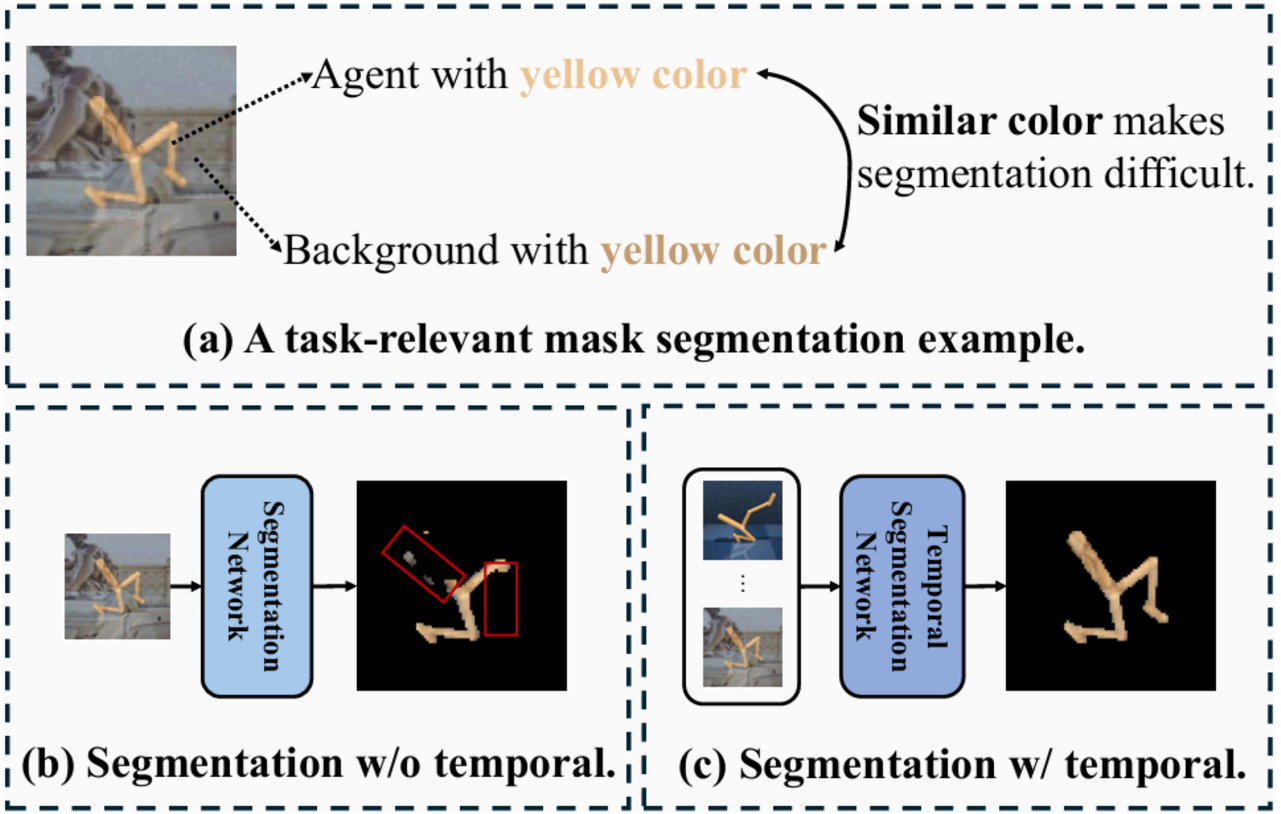
TSTM: Temporal Segmentation for Task-relevant Mask in Visual Reinforcement Learning Generalization

Weicheng Du[1] Wenjia Meng[1*] Zhengzhe Zhang[1] Yilong Yin[1] Xiankai Lu[1,2]
[1]School of Software, Shandong University, China
[2]School of Mathematics and Computer Science, Quanzhou Normal University, China
weichengdu@mail.sdu.edu.cn, wjmeng@sdu.edu.cn, zhengzhezhang@mail.sdu.edu.cn
ylyin@sdu.edu.cn, luxiankai@sdu.edu.cn

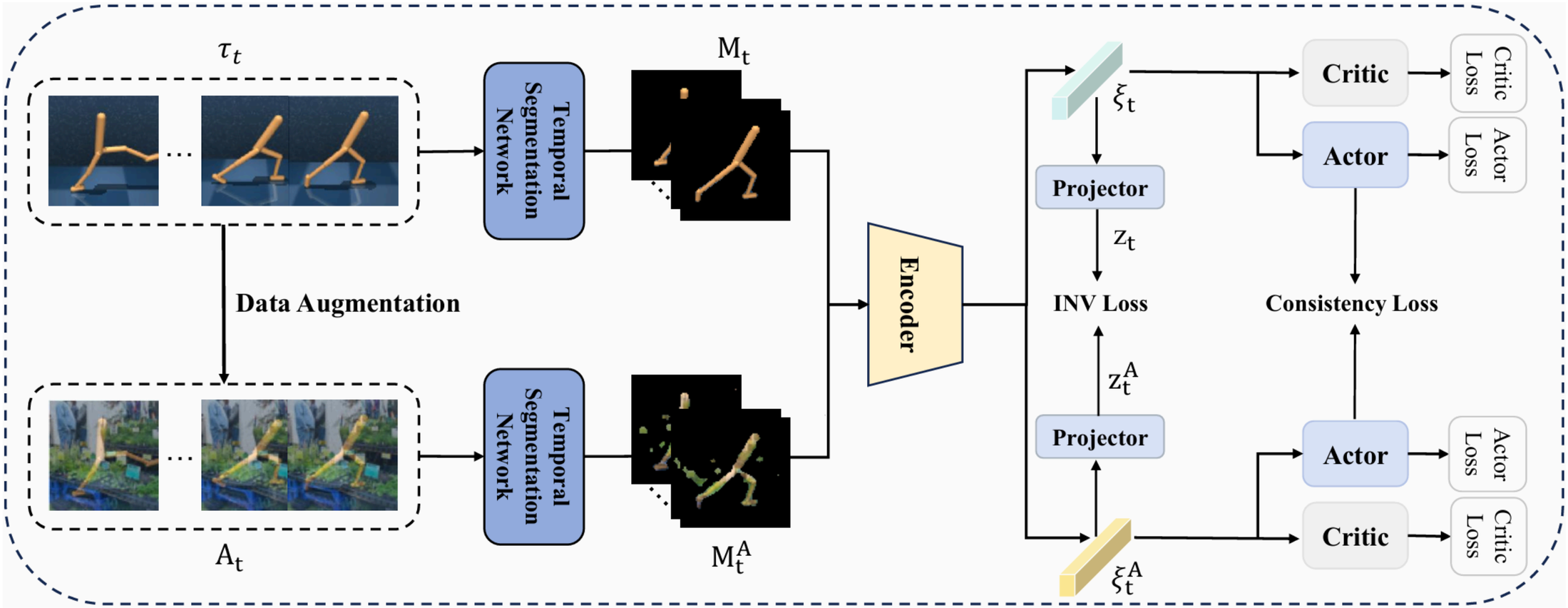
CVPR 2026

Motivation

1. Visual RL의 범용성 한계
   - 학습 환경과 실제 환경이 다르면 제어 성능이 급격히 저하됨
2. 기존 마스킹/세그멘테이션 방식의 한계
   - SGQN, MaDi 등은 작업에 필요한 영역만 추출하려 시도 → 단일 프레임만 사용함
   - 유사한 색이나 노이즈는 단일 이미지로 구분이 불가능함



Framework


τ_t
Temporal Segmentation Network
M_t
Data Augmentation
A_t
Temporal Segmentation Network
M_t^A
Encoder
ξ_t
Projector
z_t
INV Loss
z_t^A
Projector
ξ_t^A
Critic
Actor
Critic Loss
Actor Loss
Consistency Loss
Actor
Critic
Actor Loss
Critic Loss

Contribution

1. Temporal Segmentation Network
   - Teacher Model: Encoder - ConvLSTM - Decoder 구조
     - 연속된 k개의 프레임 관측 시퀀스 → ConvLSTM으로 temporal dependencies 포착, 정확한 마스크 생성
   - Lightweight Student Model: KD로 구축
2. Invariant Representation Learning
   - 데이터 증강 → 마스크 → CNN 인코더 및 프로젝터에 통과
   - VICReg 기반 정규화 손실(MSE + Variance + Covariance Loss) 적용
     - Task-relevant Invariant Representation 추출
3. Actor-Critic
   - 추출된 불변표상을 state로 입력받아 SAC 기반으로 학습
   - KL Loss로 안정적인 행동 유도

04

IN-CONTEXT COMPOSITIONAL Q-LEARNING FOR OFFLINE REINFORCEMENT LEARNING

Qiushui Xu[1], **Yuhao Huang**[2], **Yushu Jiang**[3], **Lei Song**[4], **Jinyu Wang**[4], **Wenliang Zheng**[1], **Jiang Bian**[4]
[1] Penn State University, [2] Nanjing University, [3] University of Toronto, [4] Microsoft Research
{qjx5019,wmz5132}@psu.edu, huangyh@smail.nju.edu.cn,
barryy.jiang@mail.utoronto.ca,
{Lei.Song,Wang.Jinyu,Jiang.Bian}@microsoft.com

Motivation

Offline RL은 $\mathcal{D} = \{s_i, a_i, r_i, s_i'\}$ 만으로 policy 학습

일반적으로는 전체 data → 하나의 Q function 생성

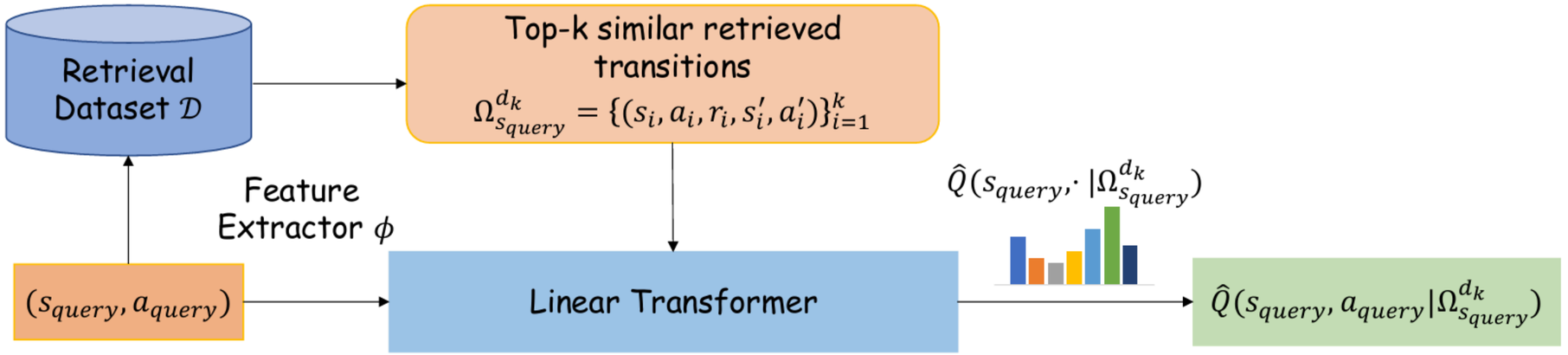
복잡한 task → 여러 상황, subtask가 있어 불리함

ex. Walker2d의 경우: 정상적으로 걷기, 균형을 잃고 회복, 넘어지기 직전, 속도를 높임 등의 상황이 있음.

각 경우 해야 할 행동이 다를 수 있으나, global Q-function은 다른 지역의 정보가 섞여 Q-value가 부정확해짐

Contribution

1. 하나의 고정된 Q-function이 아니라, 현재 상태의 문맥에 따라 달라지는 Q-function을 사용
2. Subtask label 없이도 Local Q-function 추론
   - 상태를 지정하지 않아도 retrieval과 in-context learning으로 local structure를 찾아냄
3. Linear Transformer로 In-Context TD-Learning 구현



05

From Reward-Free Representations to Preferences: Rethinking Offline Preference-Based Reinforcement Learning

Jun-Jie Yang [1] **Chia-Heng Hsu** [1] **Kui-Yuan Chen** [1] **Ping-Chun Hsieh** [1]

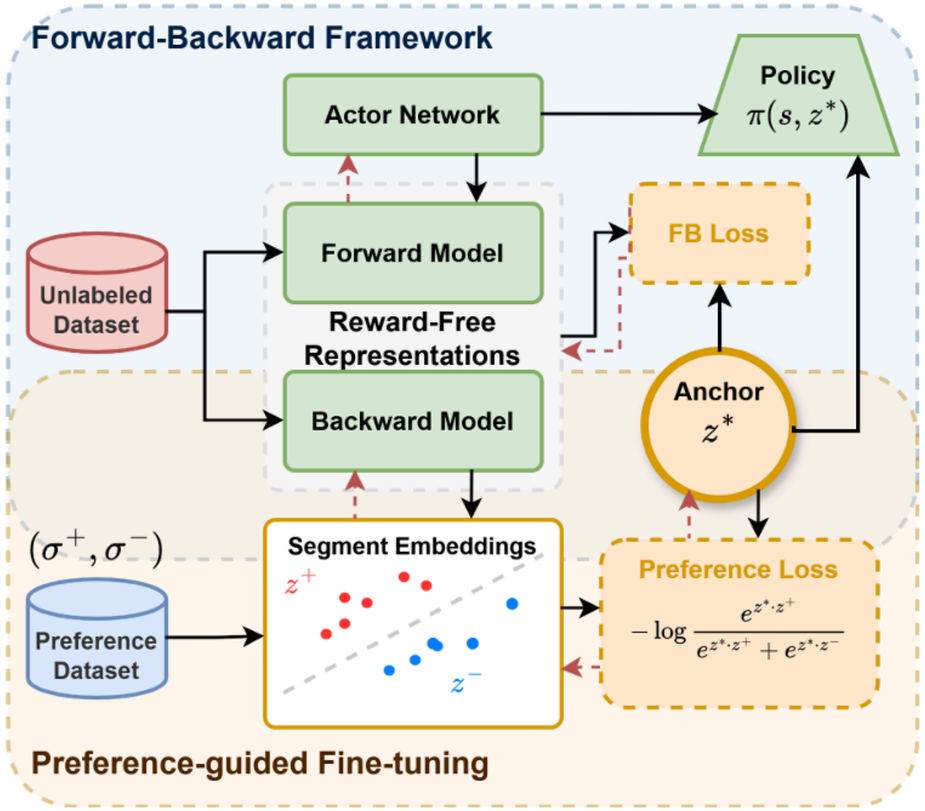
Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Motivation

기존 PbRL: Preference pair → Reward model 학습 → 전체 offline dataset에 reward 부여 → Offline RL로 policy 학습
Prefernce label이 적으면 reward model이 부정확해진다.

> 이미 reward-free 데이터에서 학습한 representation을 활용할 수 없을까?

Contribution



Contribution

1. PbRL과 RFRL의 연결
   - 선호 학습 문제를 reward model fitting 문제가 아니라, 이미 학습된 latent space에서 적절한 task direction z를 찾는 문제로 봄
2. Preference loss를 contrastive learning으로 재해석
   - 선호 피드백을 reward값으로 변환하지 않고, latent representation 간의 positive/negative contrast로 직접 사용 가능
3. Reward/Preference model이 필요없는 FB-PbRL 제안
   - Reward-Free Pre-training + Preference-guided Search and Fine-tuning
4. 높은 Preference Efficiency와 Robustness
   - 기존 offline PbRL baseline보다 preference 효율이 높으면서 강한 성능 달성.

Thank you

Question & Disscusion

RILS LAB