Causal-Paced Deep Reinforcement Learning

Geonwoo Cho[1], Jaegyun Im[1], Doyoon Kim[1], Sundong Kim[1] †

[1]Gwangju Institute of Science and Technology

gwcho.public@gmail.com

Motivation & Contributions

- Causal-Aware Curriculum Learning: We propose a principled method to measure task novelty via estimated structural causal differences, without access to ground-truth SCMs.
- Balancing Novelty and Learnability: By combining causal misalignment with reward improvement, CP-DRL constructs curricula that guide agents toward causally unfamiliar yet learnable tasks.

CP-DRL: Causal-Paced Deep Reinforcement Learning

Core Idea

We estimate structural differences between tasks using observed trajectories, capturing novelty without access to true SCMs. This signal is used to guide a teacher policy in selecting tasks for a student agent.

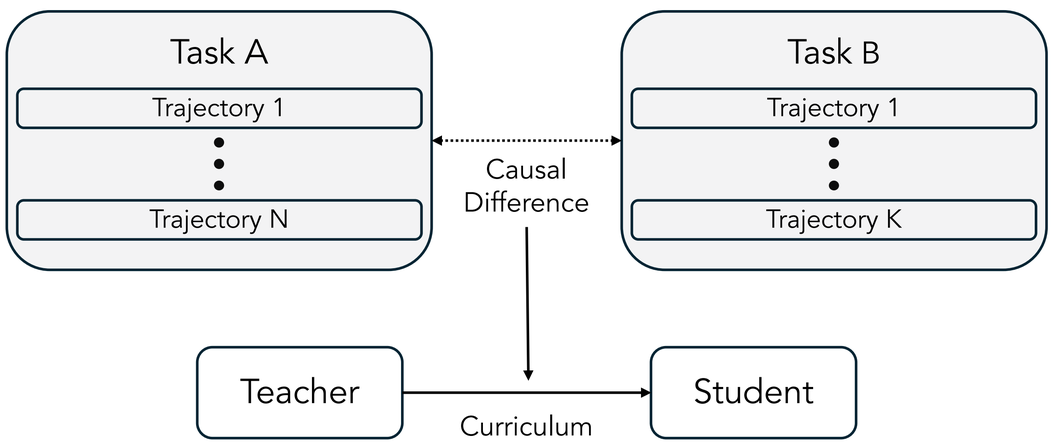


Figure 1. **Overview of CP-DRL.** Trajectory-based structural signals guide curriculum selection.

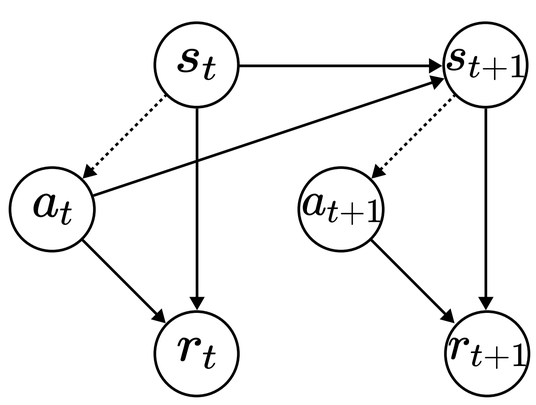


Figure 2. **Causal graph in our RL setting.** Solid arrows: environment-induced transitions and reward generation, Dotted arrows: denote policy-induced dependencies.

Prioritizing Causally Underexplored Tasks

Modular Component Models

$$z_t^s \sim q_{\phi_s}(z \mid s_t), \quad \hat{s}_t \sim p_{\theta_s}(s \mid z_t^s)$$
$$z_t^a \sim q_{\phi_a}(z \mid a_t), \quad \hat{a}_t \sim p_{\theta_a}(a \mid z_t^a)$$
$$\hat{s}_{t+1} \sim p_{\phi_r}(s_{t+1} \mid s_t, a_t), \quad \hat{r}_t \sim p_{\phi_r}(r \mid s_t, a_t)$$

Losses

$$\mathcal{L}_{\text{state}} = \|\hat{s}_t - s_t\|^2 + \beta_1 \cdot \mathrm{KL}(q_{\phi_s} \| \mathcal{N})$$
$$\mathcal{L}_{\text{action}} = \|\hat{a}_t - a_t\|^2 + \beta_2 \cdot \mathrm{KL}(q_{\phi_a} \| \mathcal{N})$$
$$\mathcal{L}_{\text{transition}} = \|\hat{s}_{t+1} - s_{t+1}\|^2, \quad \mathcal{L}_{\text{reward}} = \|\hat{r}_t - r_t\|^2$$

Causal Metric (CM)

$$\mathrm{Dis}_i = \mathrm{std}\left(\{\hat{y}_i^{(k)}\}_{k=1}^K\right), \quad i \in \{\text{state}, \text{action}, \text{transition}, \text{reward}\}$$
$$\mathrm{CM}(c) = \sum_i w_i \cdot \mathrm{Dis}_i$$

We integrate $\mathrm{CM}(c)$ *into CURROT's optimal transport cost, encouraging selection of causally novel yet learnable tasks.*

Disagreement as a Proxy for Causal Difference

A Toy Example in CausalWorld

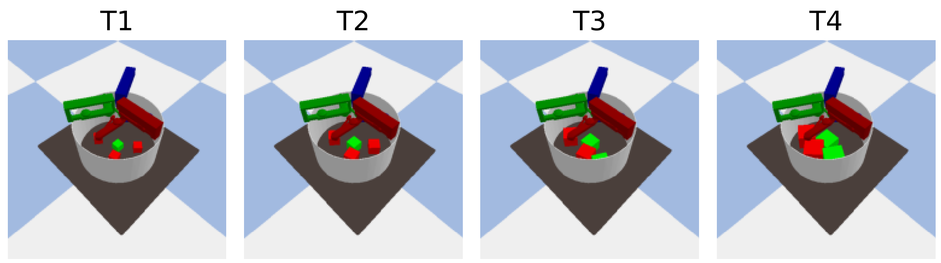


Figure 3. **Visualization of tasks T1–T4 from the CausalWorld General environment.** In this setup, agents receive rewards proportional to the intersection ratio between each block and the goal configuration. From T1 to T4, we progressively increase the block size, action magnitude, and reward scale, inducing increasing causal differences between tasks.

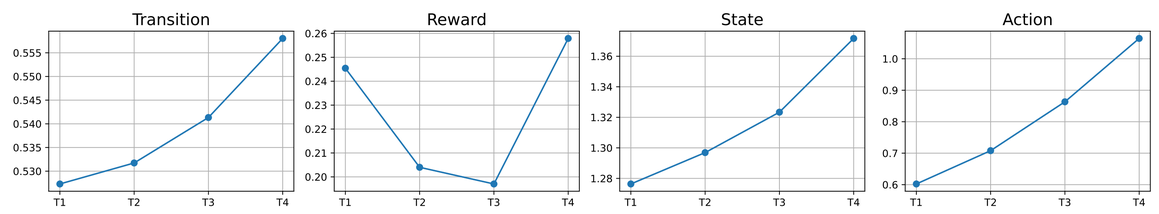
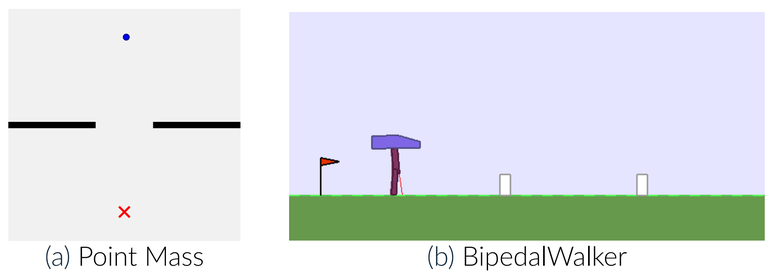


Figure 4. **Disagreement metrics across tasks T1–T4.** Each metric is computed after training on T1 for 10 episodes and sampling 5 episodes from each task. Transition, state, and action disagreements increase monotonically, suggesting they effectively reflect causal differences. Reward disagreement shows no clear trend due to reward sparsity and random trajectory collection.

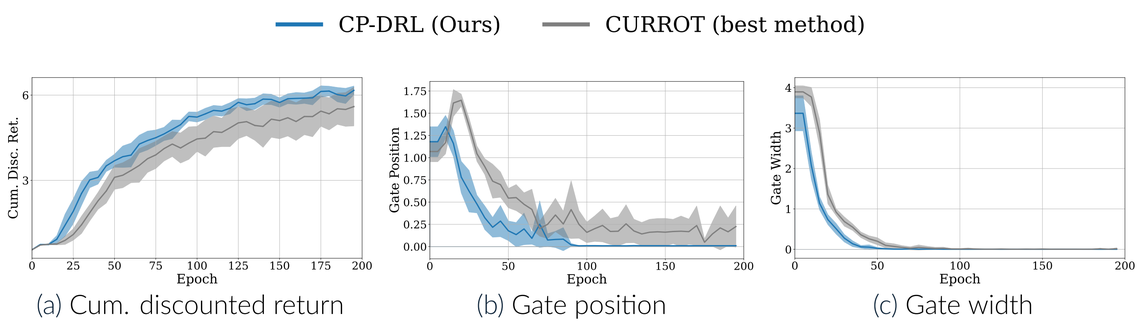
Performance

Performance in Point Mass and BipedalWalker environment



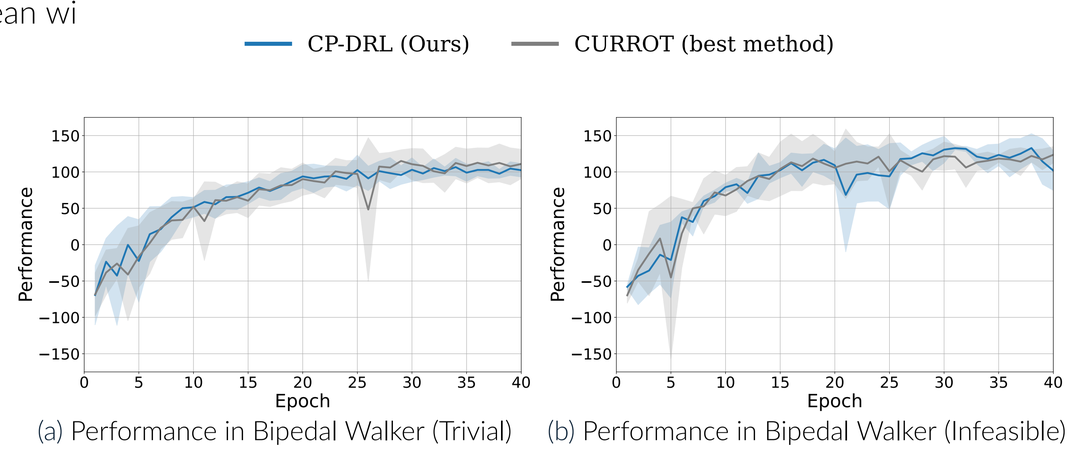
(a) Point Mass (b) BipedalWalker

Figure 5. **Illustrations of the environments used in our experiments.** (a) Point Mass. (b) BipedalWalker.



(a) Cum. discounted return (b) Gate position (c) Gate width

Figure 6. **Performance comparison between CP-DRL and CURROT in Point Mass environment.** (a) Cumulative discounted return. (b) Median distance to the target gate position. (c) Median distance to the target gate width. All curves show the mean wi



(a) Performance in Bipedal Walker (Trivial) (b) Performance in Bipedal Walker (Infeasible)

Figure 7. **Performance comparison between CP-DRL and CURROT in BipedalWalker.** (a) **Trivial**. (b) **Infeasible**. All curves show the mean with 95% confidence intervals.

Conclusions

- **Causal signals improve curriculum structure:** CP-DRL quantifies structural novelty via modular disagreement, enabling curriculum learning that prioritizes causally unfamiliar but learnable tasks.
- **Stable and efficient learning:** Across Point Mass and BipedalWalker environments, CP-DRL achieves faster convergence and more stable returns compared to CURROT and other baselines.