1em

Adversarial Robustness of Reinforcement Learning Systems

Ziqing Lu, Babak Hassibi, Lifeng Lai, Weiyu Xu

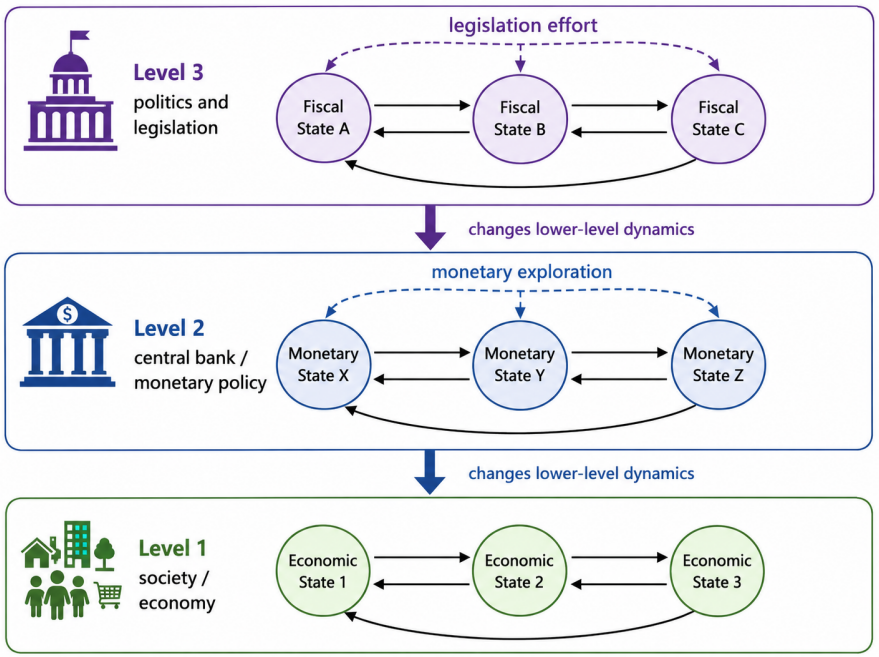
Program of Applied Mathematical and Computational Sciences,

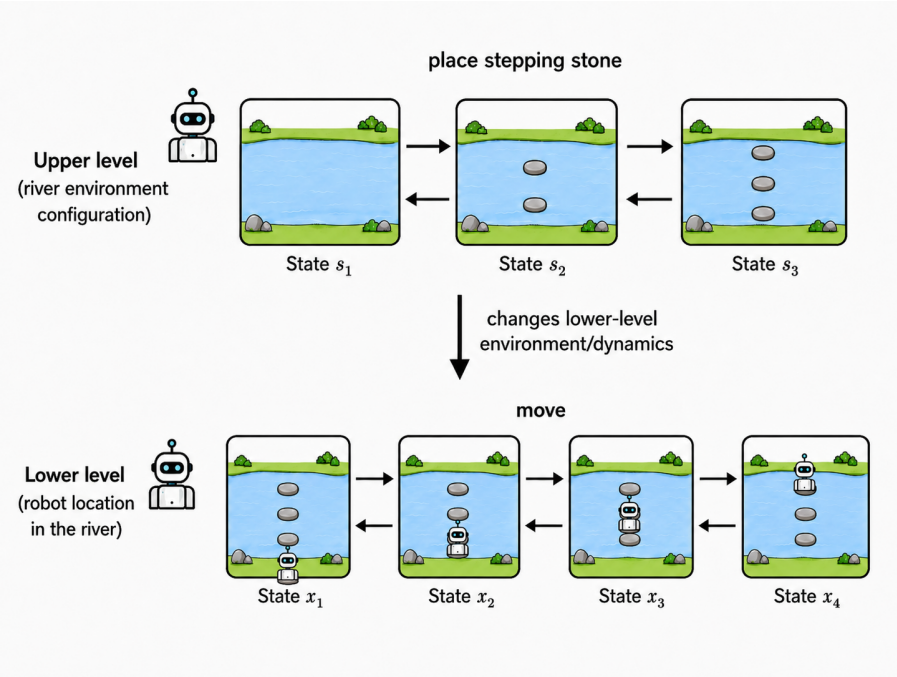
University of Iowa

ICML, July 2026

Motivation

- Passive adaptation in a fixed environment
- VS. **Active model-changing** on the environment





New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- $\mathcal{B}$ is the set of model-changing actions

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- $\mathcal{B}$ is the set of model-changing actions
- Upper level state $P \in \mathcal{P}$ is the lower-level transition kernel P
- μ_0 is the uniform state distribution, reset to be uniform every episode

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- Primitive policy of episode k: $\pi^k : \mathcal{S} \to \mathcal{A}$

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- Primitive policy of episode k: $\pi^k : \mathcal{S} \to \mathcal{A}$
- Closed form lower-level state-value: $V^{\pi_k, P_k} = (I - \gamma P_k^{\pi_k})^{-1} r^{\pi_k}$

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- Primitive policy of episode k: $\pi^k : \mathcal{S} \to \mathcal{A}$
- Closed form lower-level state-value: $V^{\pi_k, P_k} = (I - \gamma P_k^{\pi_k})^{-1} r^{\pi_k}$
- Initial expected return: $J(\pi_k, P_k) = \mu_0^T V^{\pi_k, P_k}$

New Framework: Bi-level Configurable MDP

Bi-level Configurable MDP

The bi-level configurable MDP can be formulated as: a lower-level $\mathcal{M}_L = \{\mathcal{S}, \mathcal{A}, P, T, \mu_0, r, \gamma\}$ and an upper-level $\mathcal{M}_U = \{\mathcal{P}, \mathcal{B}, Q, K, R, \lambda\}$: $\mathcal{S}, \mathcal{P}$ are the state spaces, $\mathcal{A}, \mathcal{B}$ are the action spaces, P, Q are the transition functions, K is the number of episodes, T is the horizon, r and R are reward functions, γ and λ are the discount factors of each level environment.

- Upper-level transition function: $Q(P'|P, b)$
- Upper reward function: $R : \mathcal{P} \times \mathcal{B} \to \mathbb{R}$, $C(P_k, b_k)$ is the cost function

$$R(P_k|b_k) = \sum_{P_{k+1} \in \mathcal{P}} Q(P_{k+1}|P_k, b_k) J(\pi^*_{k+1}, P_{k+1}) - C(P_k, b_k) \quad (1)$$

- Configuration policy: Θ, and higher-level state value W^{Θ}

Bi-level Value Iteration: Lower Level

Algorithm: Bi-level Value Iteration, Part I

Require: Empirical lower-level MDPs $\hat{\mathcal{M}}_L = \{(\mathcal{S}, \mathcal{A}, \hat{P}, \mu_0, r, \gamma) : \hat{P} \in \hat{\mathcal{P}}\}$; lower-level iteration number $H_L > 0$; initial estimates $\{V^0_{\hat{P}}\}_{\hat{P}\in\hat{\mathcal{P}}}$.

Ensure: Lower-level values $\{V^{H_L}_{\hat{P}}\}_{\hat{P}\in\hat{\mathcal{P}}}$, policies $\{\pi^{H_L}_{\hat{P}}\}_{\hat{P}\in\hat{\mathcal{P}}}$, and returns $\{J_{\hat{P}}\}_{\hat{P}\in\hat{\mathcal{P}}}$.

1: **for** $h = 0, 1, \ldots, H_L$ **do**
2: **for** each $\hat{P} \in \hat{\mathcal{P}}$ **do**
3: Update, for all $s \in \mathcal{S}$,

$$V^{h+1}_{\hat{P}}(s) \leftarrow \max_{a\in\mathcal{A}} \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \hat{P}(s'|s,a) V^h_{\hat{P}}(s') \right).$$

4: **end for**
5: **end for**
6: **for** each $\hat{P} \in \hat{\mathcal{P}}$ **do**
7: Set

$$\pi^{H_L}_{\hat{P}} \leftarrow \text{greedy policy with respect to } V^{H_L}_{\hat{P}}, \qquad J_{\hat{P}} \leftarrow \mu_0^T V^{H_L}_{\hat{P}}.$$

8: **end for**

Bi-level Value Iteration: Upper Level

Algorithm: Bi-level Value Iteration, Part II

Require: Empirical upper-level MDP $\hat{\mathcal{M}}_U = (\hat{\mathcal{P}}, \mathcal{B}, \hat{Q}, \hat{R}, \lambda)$; upper-level iteration number $H_U > 0$; lower-level returns $\{J_{\hat{P}}\}_{\hat{P} \in \hat{\mathcal{P}}}$; initial estimate W^0.

Ensure: Upper-level estimate $W^{H_U} \in \mathbb{R}^m$ and higher-order policy Θ^{H_U}.

1: **for** $h = 0, 1, \ldots, H_U$ **do**
2: **for** each $\hat{P} \in \hat{\mathcal{P}}$ **do**
3: Update

$$W^{h+1}(\hat{P}) \leftarrow \max_{b \in \mathcal{B}} \left(\sum_{\hat{P}' \in \hat{\mathcal{P}}} \hat{Q}(\hat{P}' | \hat{P}, b) \left(J_{\hat{P}'} + \lambda W^h(\hat{P}') \right) \right).$$

4: **end for**
5: **end for**
6: Set

$$\Theta^{H_U} \leftarrow \text{greedy policy with respect to } W^{H_U}.$$

7: **return** Θ^{H_U}, W^{H_U}.

Physical limitation and estimation error

- **Physical discrepancy** between the ideal configured environment and the real environment

$$TV(P_c(\cdot|s,a), P(\cdot|s,a)) = 1/2\|P_c(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \delta_c$$

Physical limitation and estimation error

- **Physical discrepancy** between the ideal configured environment and the real environment

  $TV(P_c(\cdot|s,a), P(\cdot|s,a)) = 1/2\|P_c(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \delta_c$

- **Estimation error** between the real environment and the measured environment
  - $TV(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \delta_g$
  - $TV(Q(\cdot|P,b), \hat{Q}(\cdot|P,b)) \leq \Delta$

Performance Analysis

Lemma (estimation error Lemma)

If $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, *the total variance distance between the empirical kernel* $\hat{P}$ *and the ground-truth* P *is bounded by* $TV(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \delta_g$, *then for any policy* $\pi : \mathcal{S} \to \mathcal{A}$, *we have*

$$\|V_P^\pi - V_{\hat{P}}^\pi\|_\infty \leq \frac{\gamma \delta_g V_{\max}}{(1-\gamma)},$$

where $V_{\max} := \frac{r_{\max}}{1-\gamma}$. *This result also applies to the optimal policies:*

$$\|V_P^{\pi^*(P)} - V_{\hat{P}}^{\pi^*(\hat{P})}\|_\infty \leq \frac{\gamma \delta_g V_{\max}}{(1-\gamma)}.$$

Performance Analysis

Lemma (Propagated Reward Error Bound)

We assume $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, *the total variance distance between the empirical kernel* $\hat{P}$ *and the ideally configured kernel* P_c *is bounded by* $TV(P_c(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \delta_g + \delta_c$. *Then for any higher-order deterministic policy* $\Theta : \mathcal{P}_c \to \mathcal{B}$, *we have the error bound for the upper-level reward function:*

$$\|R_Q^\Theta - \hat{R}_Q^\Theta\|_\infty \leq \frac{\gamma(\delta_g + \delta_c)V_{\max}}{1-\gamma}. \tag{2}$$

where the elements of $\hat{R}_Q^\Theta, R_Q^\Theta \in \mathbb{R}^m$ *are of the form as upper-level reward function definition (1), and* m *is the number of states in the upper-level MDP.*

Performance Analysis

Lemma (Error bound Lemma of Bi-level MDPs)

If $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$, *the total variance distance between the empirical kernel* $\hat{P}$ *and the ideally configured kernel* P_c *is bounded by* $TV(P_c(\cdot|s,a), \hat{P}(\cdot|s,a)) \leq \delta_g + \delta_c$, *and if for* $\forall(P,b) \in \mathcal{P} \times \mathcal{B}$, *the total variance distance between the ground-truth higher-order kernel* Q *and the empirical higher-order kernel* $\hat{Q}$ *is bounded by* $TV(Q(\cdot|P,b), \hat{Q}(\cdot|P,b)) \leq \Delta$, *then for any higher-order policy* $\Theta : \mathcal{P} \to \mathcal{B}$, *we have that*

$$\|W_Q^\Theta - W_{\hat{Q}}^\Theta\|_\infty \leq \frac{\gamma(\delta_g + \delta_c)V_{\max}}{(1-\gamma)(1-\lambda)} + \frac{2\Delta \cdot \|\mu_0\|_\infty V_{\max} + 2\lambda\Delta \cdot W_{\max}}{1-\lambda}, \tag{3}$$

where $W_{\max} = \frac{R_{\max}}{1-\lambda}$.

Numerical Experiment: Configure the MuJoCo-Reacher's Environment

- **Goal** Move the robot's fingertip close to a target
- **Upper-level state space** target coordinates $(x, y) \in [-0.25, 0.25]^2$
- **Model-changing actions** (δ_x, δ_y)
- **Upper-level reward** $R = \sum_{k=0}^{K} r_{k,L} - \lambda_{cost} \|\Delta\|_2$, $\Delta^2 = \delta_x^2 + \delta_y^2$

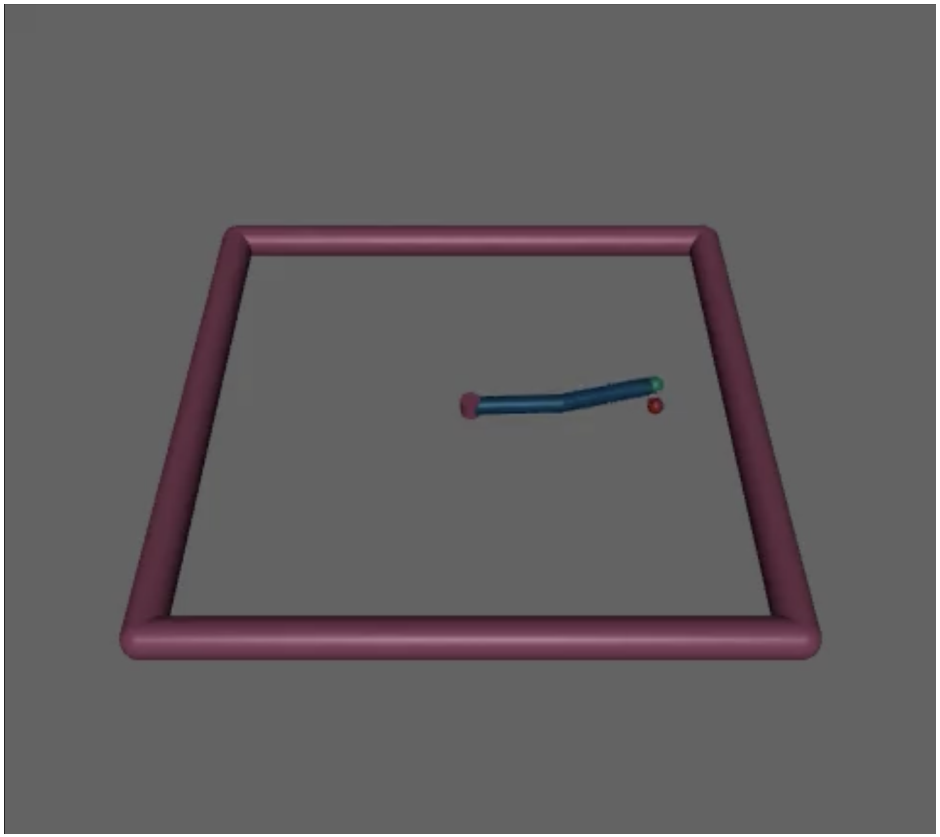


Figure: MuJoCo-Reacher environment

Numerical Experiment

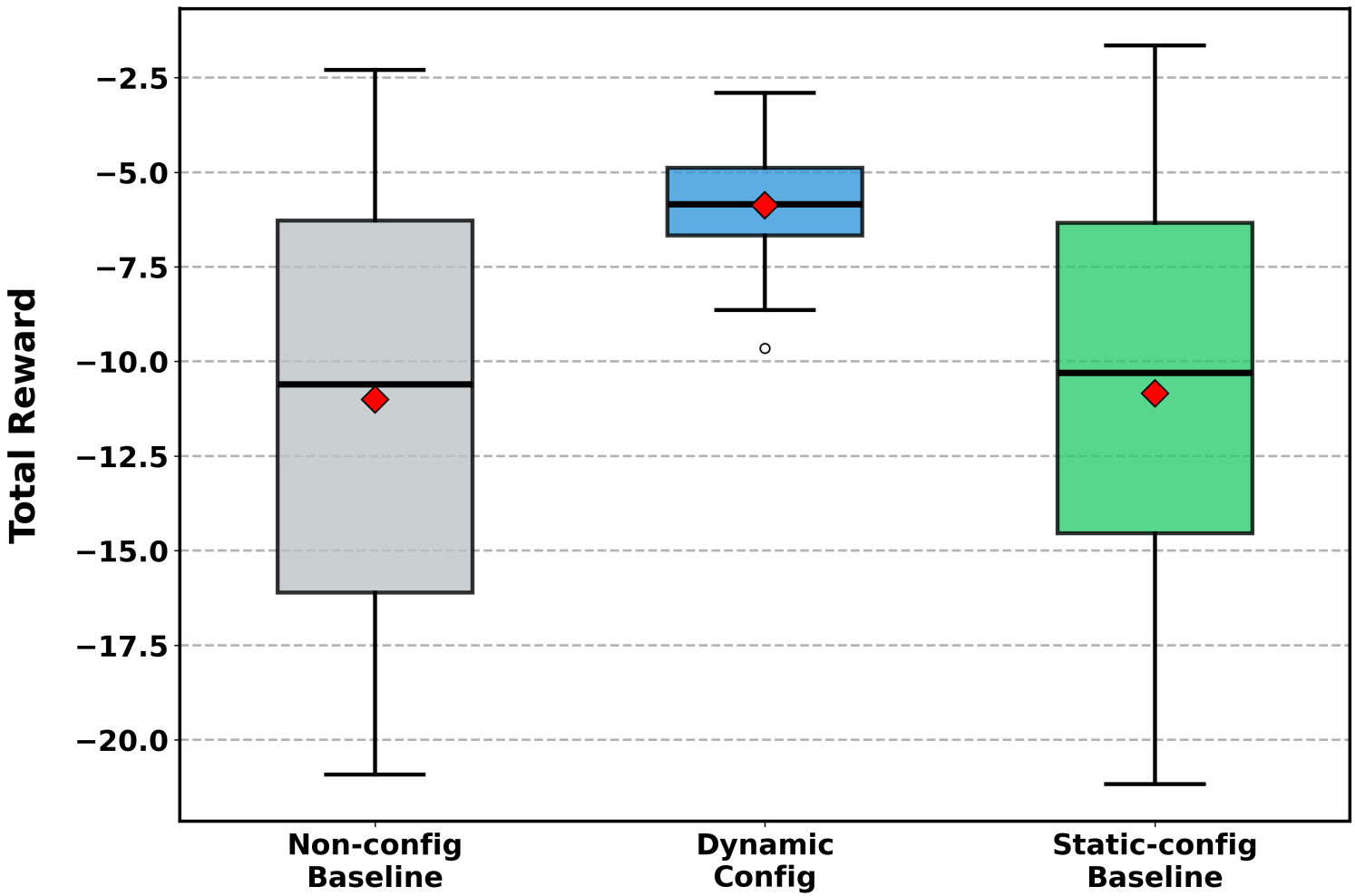


Figure: Performance Comparison of Configurable Reacher