# Discovering Creative Behaviors through DUPLEX: Diverse Universal Features for Policy Exploration

Borja G. Leon[†,1], Francesco Riccio[2], Kaushik Subramanian[2], Peter R. Wurman[2], Peter Stone[2,3]
[1]Iconic, [2]Sony AI, [3]The University of Texas at Austin
[†]internship project while at Sony AI.

Videos and more results!

**Sony AI**
https://ai.sony/joinus/

## 1. Introduction

**Motivation:** In canonical RL settings, agents are set to regress towards an optimal single policy. DUPLEX builds on previous work to generalize such a paradigm and train agents to find a diverse set of policies that can solve context-dependent tasks. In the modern gaming industry the capability to show **diverse behaviors is extremely important to create engaging experiences for users**.

**Problem:** Diversity learning increases complexity of the training problem for RL agents that now have to **trade-off performance and diversity** in order to show different competitive behaviors. DUPLEX makes training of diverse policies robust in hyper-dynamic, realistic environment.

**Research Question:** *Can we train a **multi-policy RL agent** where each policy solves context-dependent tasks while following **diverse trajectories for each context** and apply it to **complex hyper-realistic environments**?*

## 2. DUPLEX

### a) Definition 4.1 (Diversity)

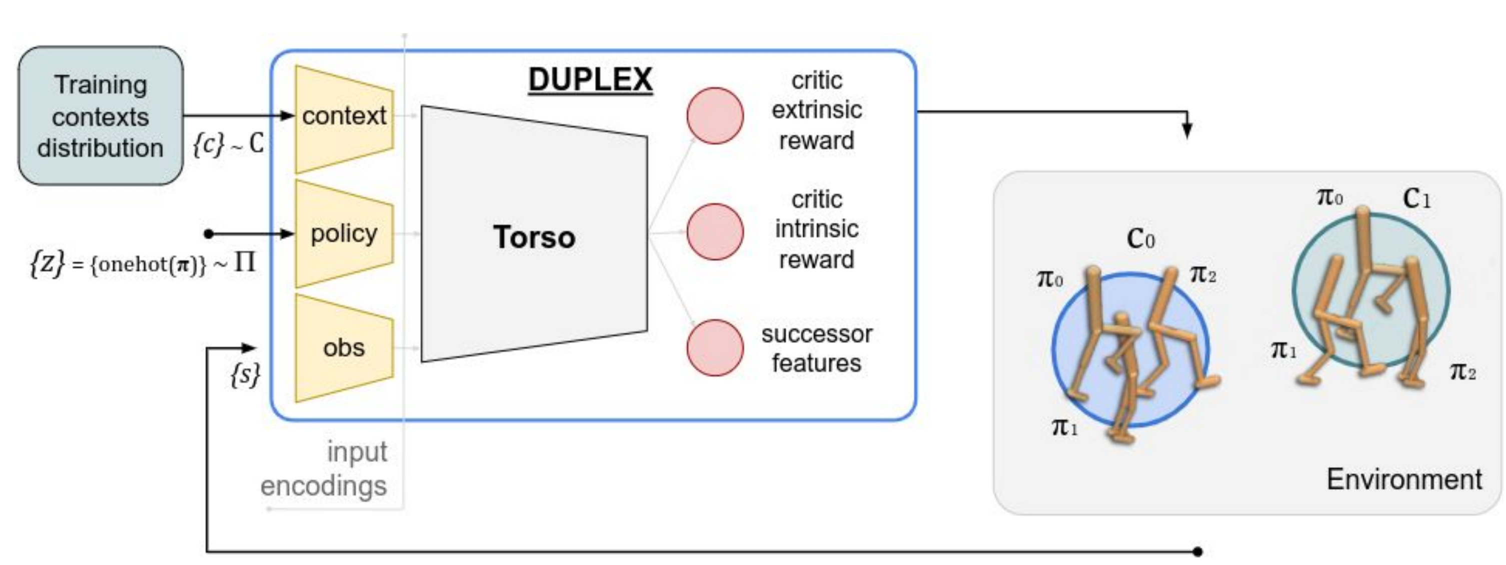
$$\text{Diversity}(\Pi) = \frac{1}{2\ size(\Pi)} \cdot \sum_{\substack{\forall \pi_i, \pi_j \in \Pi, \\ i != j}} \min ||\psi_{\pi_i} - \psi_{\pi_j}||_2^2$$

### b) Training objective

$$\max_{\Pi} \ \text{Diversity}\ (\Pi)\ \text{s.t}\ d_{\pi_c} \cdot r_e \geq \rho \hat{v}_e, \quad \forall \pi_c \in \Pi$$

$$r_d^i(s,a,c) = \phi(s,a,c) \cdot (\psi_{\pi_c^i} - \psi_{\bar{\pi}_c^i})$$

### c) DUPLEX data flow




### d) Dynamic intrinsic reward factor:

$$\chi_t = \alpha_\chi \chi_t' + (1 - \alpha_\chi)\chi_{(t-1)}$$

$$\chi' = |v_{e_{\text{avg}}} / v_{d_{\text{avg}}}|(1 - \rho)$$

χ scales intrinsics rewards proportionally to the sum of extrinsic values of policies in Π.

### e) Soft-lower bound:

$$\lambda = \left\{ \sigma_k \left( \frac{v_{e_{\text{avg}}}^i - \beta \hat{v}_{e_{\text{avg}}}}{|\hat{v}_{e_{\text{avg}}} + l|} \right) \right\}_{i=1}^n$$

λ to bound the near-optimal subspace for each policy using where σ is a sigmoid function and β ∈ [0, 1] indicates the reward region we are interested in exploring

### f) Adding entropy to the SFs estimation

$$\psi^{\gamma,i}(s_t, a_t, c) = \phi_t + \mathbb{E}_{\pi_c} \sum_{k=t+1}^{\infty} \gamma^{k-t} [\phi_k + \alpha_H H(\pi_c^i(\cdot|s,c))]$$

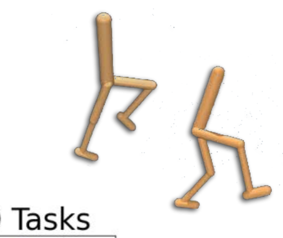
to support critic estimates when policy is uncertain

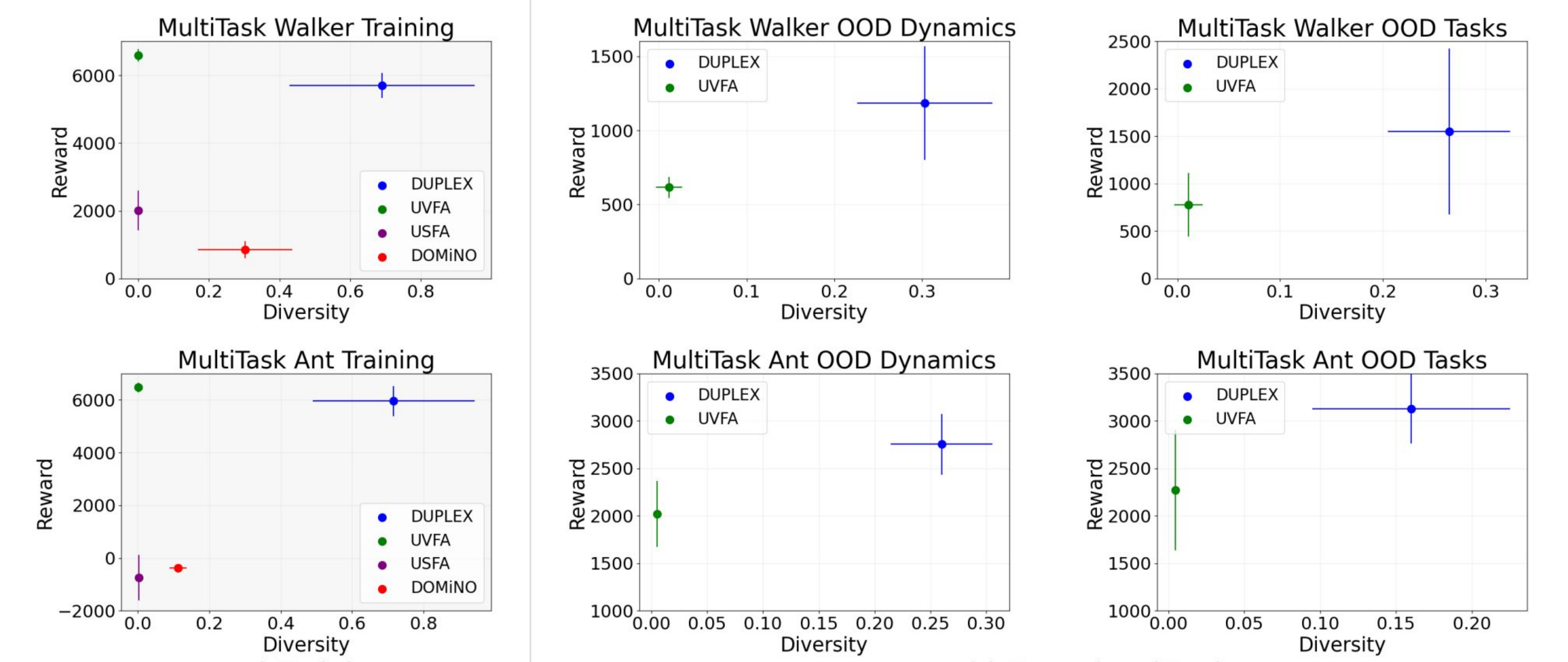
### g) Averaging critic networks

$$y(\phi, s', c, z) = \phi(t) + \gamma \left( \underset{j=1,2}{\text{avg}}\ \tilde{\psi}_{\theta_{\text{targ},j}}(s', \tilde{a}_z', c) - \alpha \log \pi_\omega^z(\tilde{a}_z'|s', c) \right)$$

## 3. Experimental Session

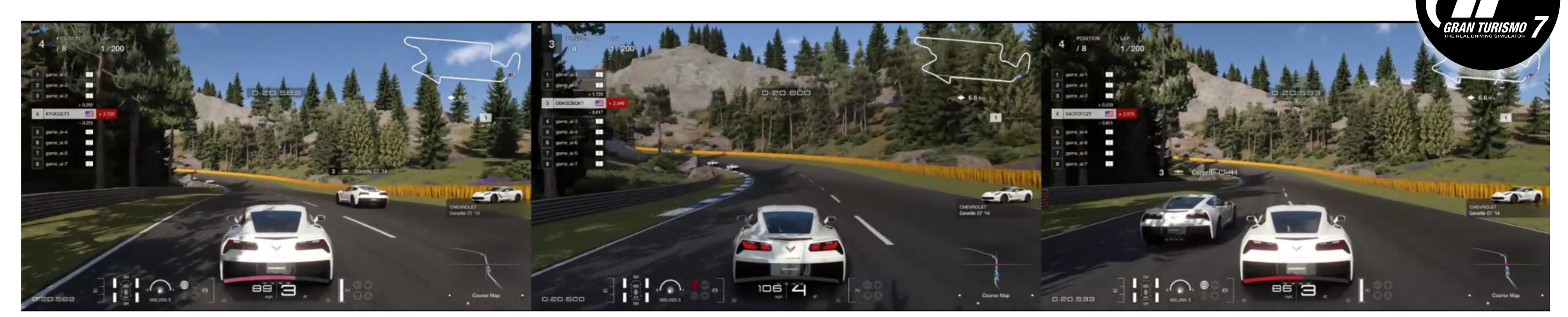
### a) MuJoCo Walker2D and Ant multitask environments

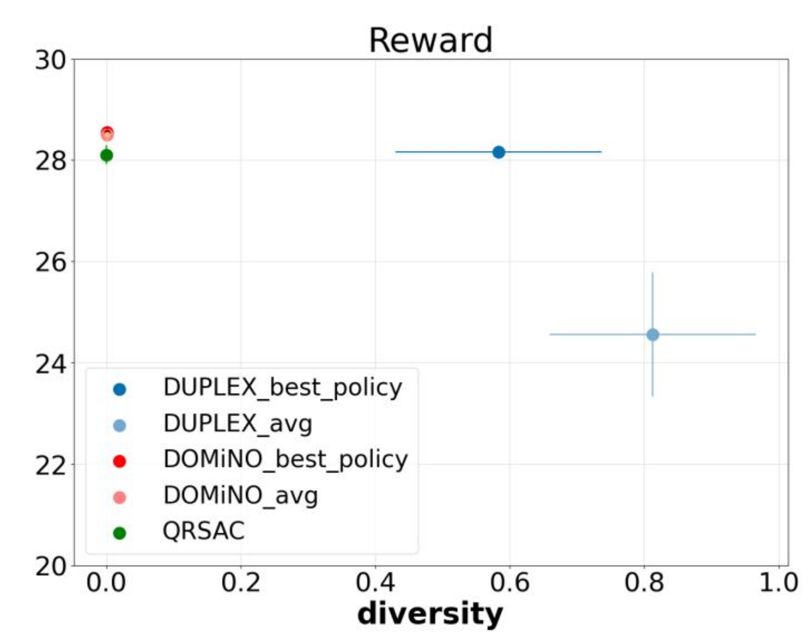


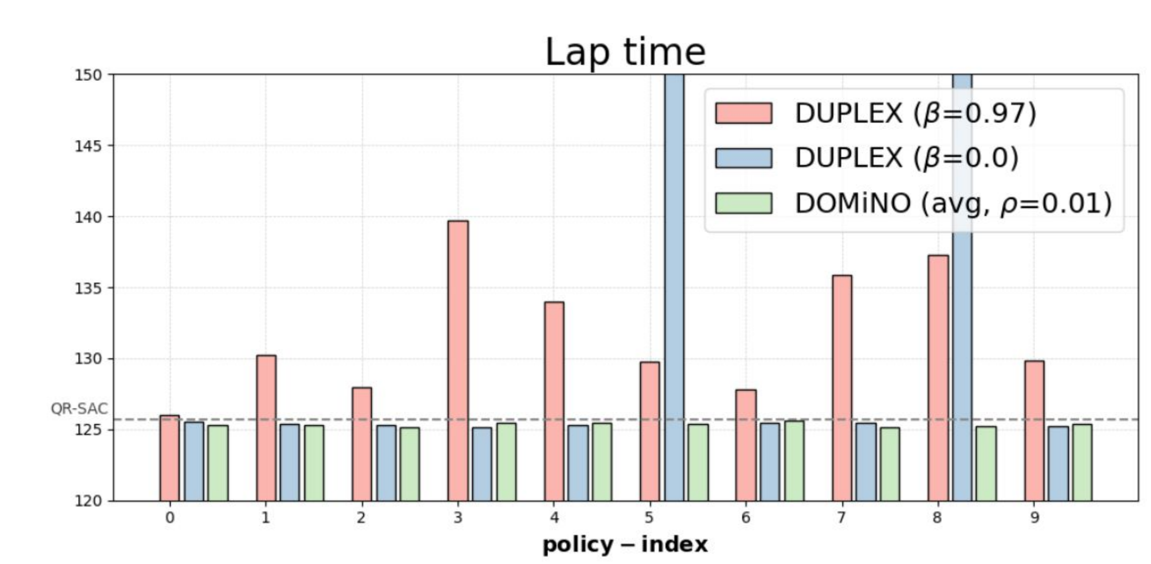


a) Training b) Zero-shot (Test)

### b) GranTurismo 7 environment






**References**
[1] Zahavy, T. et al.. Discovering policies with domino: Diversity optimization maintaining near optimality. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.
[2] Wurman, Peter R., et al. "Outracing champion Gran Turismo drivers with deep reinforcement learning." Nature 602.7896 (2022): 223-228.