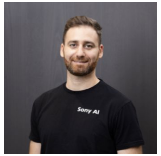
Discovering Creative Behaviors through **DUPLEX**: **D**iverse **U**niversal Features for **P**olicy **Ex**ploration

Borja G. León[†,1]



Francesco Riccio[2]

Kaushik Subramanian[2]

Peter R. Wurman[2]

Peter Stone[2,3]

Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024

[†]internship project while at Sony AI.

1 ICONIC 2 Sony AI 3 TEXAS The University of Texas at Austin

Superhuman **but** FUN?

Superhuman and **FUN**

DUPLEX contributes to diversity learning in RL by improving on previous work to better **preserve the diversity vs. near-optimality trade-off** in highly-dynamic environments and multi-context settings.

Showing diversity and acting differently in the world **is fundamental** to create **engaging experiences for users**

Context-conditioned diversity learning

Diversity(Π) is a metric of dissimilarity among policies in a set Π with a common goal. Formally, if ψ_{π_i} *and* ψ_{π_j} *are a function of state-occupancy of relevant features of any two policies in Π, then their dissimilarity is given by* $||\psi_{\pi_i} - \psi_{\pi_j}||$*. A non-zero value of this norm indicates dissimilarity, with larger values indicating greater divergence between the policies. Mathematically, diversity is defined as the sum of the minimum L2 dissimilarity norms in*

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

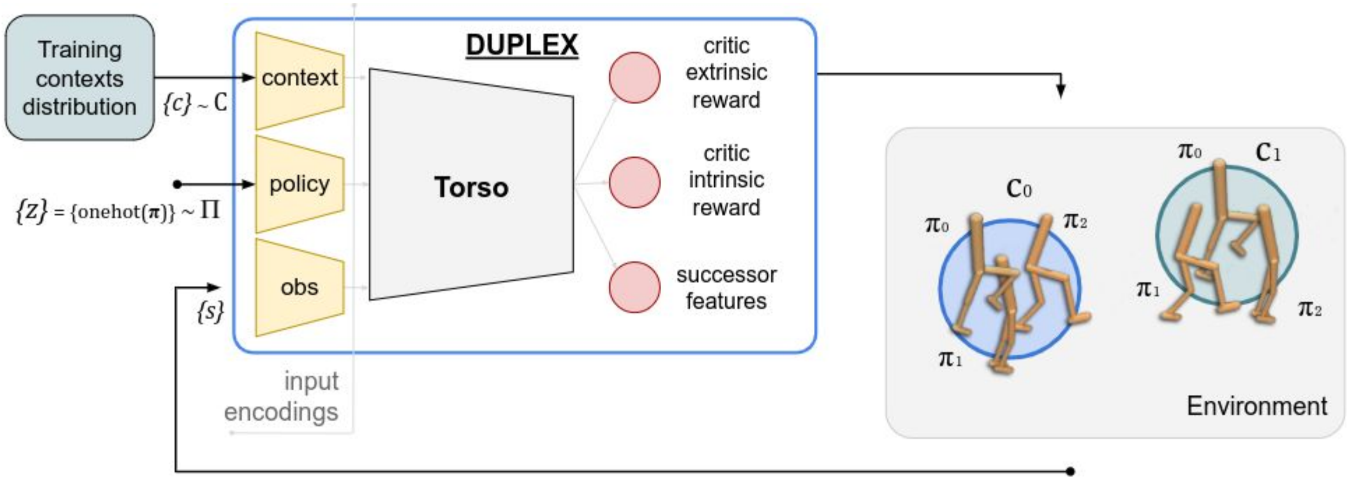
Context-conditioned diversity learning

Accordingly, we measure ψ distances to enforce **context-conditioned diversity within Π**. We aim at training an RL agent that, given a context *c*, discovers a set of *n* near-optimal policies by optimizing our objective function

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

that forces the policies in Π to only **explore for diversity within the near-optimal region of the target value**

Data flow



DUPLEX receives three inputs: (i) a context vector describing **task requirements** and environment dynamics; (ii) an **encoding of the policy**; (iii) and the current **state of the environment**. The critic network returns estimates for the **intrinsic and extrinsic rewards and successor features** to drive diverse behavior discovery. Finally, the algorithm samples policies in Π uniformly and rolls them out to collect more experience.

DUPLEX **stabilizes training and achieves diverse competitive policies** by introducing novel components to modulate the contribution of the intrinsics reward

Dynamic intrinsic reward factor

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

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

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

$$r_I = \lambda \cdot \chi \cdot r_d$$

Soft-lower bound

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

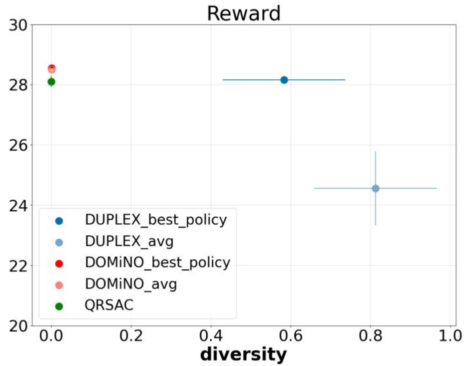
λ to bound the near-optimal subspace for each policy

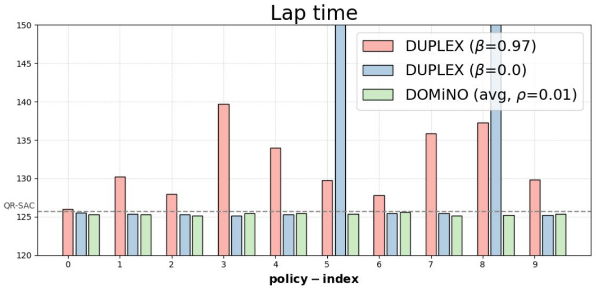
Averaging critic networks

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

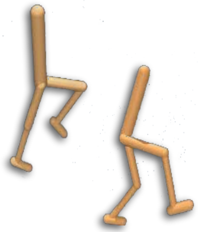
Results: GranTurismo™ 7

DUPLEX trains **diverse competitive policies** in hyper-realistic driving simulators

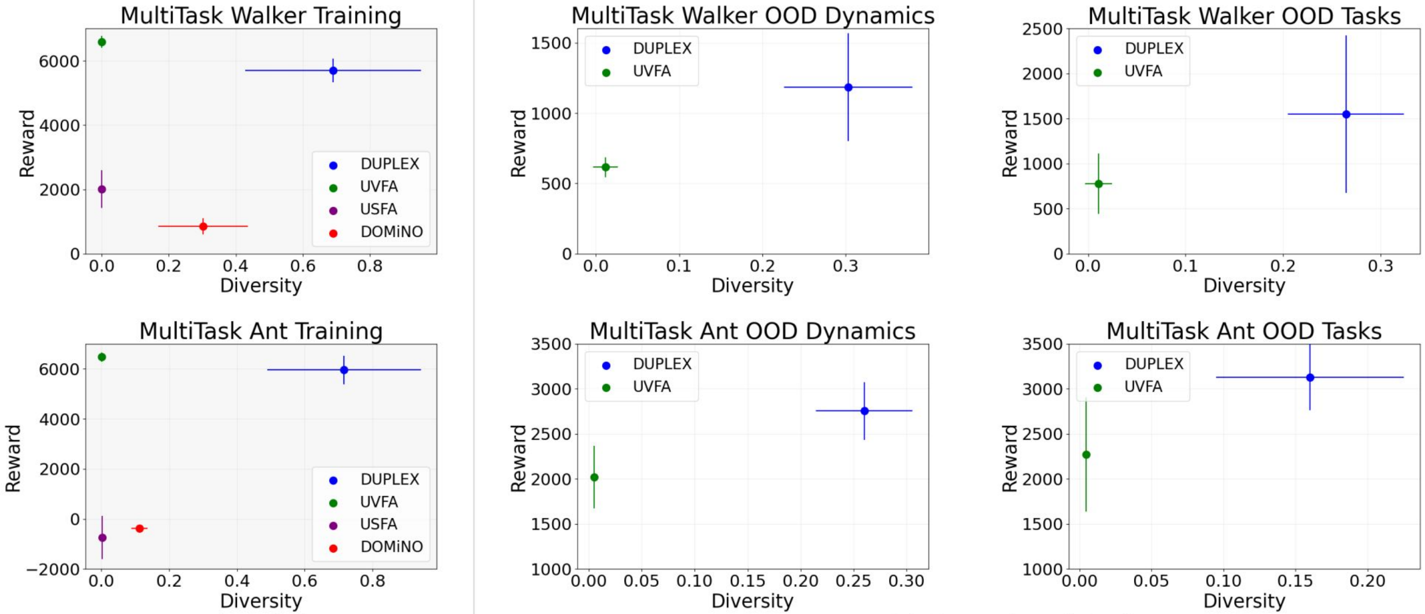




Results: MuJoCo Walker2d and Ant



DUPLEX improves the diversity vs near-optimality trade-off both **within- and out-of- distribution** settings

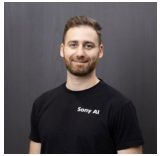


a) Training

b) Zero-shot (Test)

Discovering Creative Behaviors through **DUPLEX**: **D**iverse **U**niversal Features for **P**olicy **Ex**ploration


Borja G. León[†,1]


Francesco Riccio[2]


Kaushik Subramanian[2]


Peter R. Wurman[2]


Peter Stone[2,3]

Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024

[†]internship project while at Sony AI.

1 ICONIC 2 Sony AI 3 TEXAS The University of Texas at Austin

more videos and results