Appendix

0.1 Weighted Linear Recombination

In BEL, we directly adopt the weighted linear recombination from CMA-ES as the population selection mechanism, which has proven to be robust and scale-agnostic. Denoting the episodic reward of policy π as $R(\pi)$, the current generation as g, the recombination weight as ω, and the population size as λ, a constant set of weights is calculated according to performance ranks as follows:

$$w_i = \frac{\ln(\lambda+1) - \ln i}{\lambda \ln(\lambda+1) - \sum_{j=1}^{\lambda} \ln j}, \quad i = 1, \ldots, \lambda \tag{1}$$
$$\text{s.t.} \quad R(\tilde{\pi}_1) \geq \ldots \geq R(\tilde{\pi}_\lambda)$$

Then the center policy for the next generation is obtained by:

$$\pi_\theta^{g+1} = \sum_{i=1}^{\lambda} \omega_i \tilde{\pi}_i^g \tag{2}$$

0.2 Explorative Studies Figures

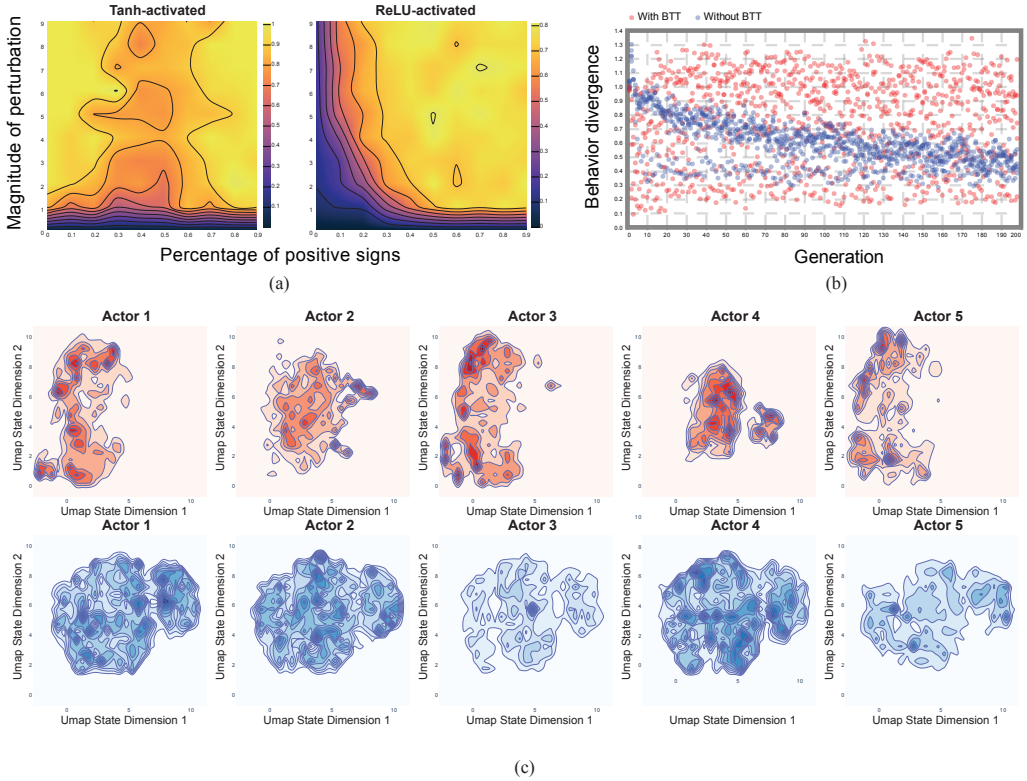

Figure 1: (a) BRP for the Tanh-activated network (left) and the ReLU-activated network (right). The heatmaps revealed different responses to unbiased perturbations. (b) An illustration of the trained offsprings' behavior divergence to the center policy, the BTT-trained population maintained high diversity throughout training. (c) The density maps of state visitation (after UMAP dimension reduction) of five actors trained with BTT (top row) versus without BTT (bottom row) after one generation. More frequent visitation corresponds to darker color. The difference between visitation patterns within each row indicates population diversity level.

0.3 Mujoco Experiments Results in Table

TABLE 1: NUMERICAL RESULTS FOR FINAL BEST MEAN REWARD OF DIFFERENT ALGORITHMS ON SELECTED TASKS

TASKS	STATISTICS	BEL (OURS)	SAC	TD3	CEM-RL	PDERL
HALFCHEETAH-V3	MEAN	**12725.39**	10482.39$^{\circledast}$	10408.62$^{\circledast}$	10636.94$^{\circledast}$	6917.24$^{\circledast}$
	STD	**202.89**	1253.81	1093.64	2131.36	444.95
	MEDIAN	**12751.99**	11058.84	10810.48	11323.23	7026.15
	WALLCLOCK	4.50H	5.34H	**2.31H**	5.52H	3.20H
ANT-V3	MEAN	**6082.41**	5208.65$^{\circledast}$	5090.81$^{\circledast}$	3455.95$^{\circledast}$	1609.40$^{\circledast}$
	STD	**166.28**	282.64	651.13	1359.87	542.42
	MEDIAN	**6147.19**	5259.90	5385.89	3487.73	1582.24
	WALLCLOCK	5.81H	8.03H	**3.09H**	6.47H	3.61H
WALKER2D-V3	MEAN	**5723.30**	4637.03$^{\circledast}$	3855.60$^{\circledast}$	4173.30$^{\circledast}$	1588.51$^{\circledast}$
	STD	**498.38**	414.19	760.91	1153.97	641.26
	MEDIAN	**6087.36**	4682.27	4138.82	4358.34	1253.77
	WALLCLOCK	4.14H	7.54H	**2.54H**	5.91H	3.42H
HOPPER-V3	MEAN	**3717.14**	3543.35$^{\circledast}$	3426.26$^{\circledast}$	3597.87$^{\circledcirc}$	1293.66$^{\circledast}$
	STD	**101.08**	103.29	192.31	495.28	356.54
	MEDIAN	**3740.41**	3580.16	3333.04	3749.87	1160.93
	WALLCLOCK	4.29H	7.94H	**2.30H**	5.82H	3.35H
HUMANOID-V3	MEAN	5337.20	**5617.94**$^{\ominus}$	5319.09$^{\circledcirc}$	215.79$^{\circledast}$	815.96$^{\circledast}$
	STD	113.53	133.93	114.38	**0.44**	90.86
	MEDIAN	5364.52	**5588.50**	5333.41	215.76	821.11
	WALLCLOCK	7.74H	8.48H	**4.51H**	9.25H	4.85H
DELAYED-HALFCHEETAH-V3	MEAN	**6777.87**	4763.14$^{\circledast}$	4730.74$^{\circledast}$	6276.42$^{\circledcirc}$	2865.77$^{\circledast}$
	STD	**596.49**	758.29	806.42	857.72	658.37
	MEDIAN	**6857.65**	4734.52	4469.37	6372.91	3095.51
	WALLCLOCK	4.69H	5.73H	**2.45H**	6.82H	3.15H

[1] The Wilcoxon rank sum test was conducted between the performance of BEL and comparing algorithms at a 0.05 significance level. $\circledast$ denotes that BEL significantly outperforms the competing algorithm, and $\ominus$ denotes the opposite. $\circledcirc$ denotes no significant difference detected.

[2] The bold text denotes the best metric within each row.