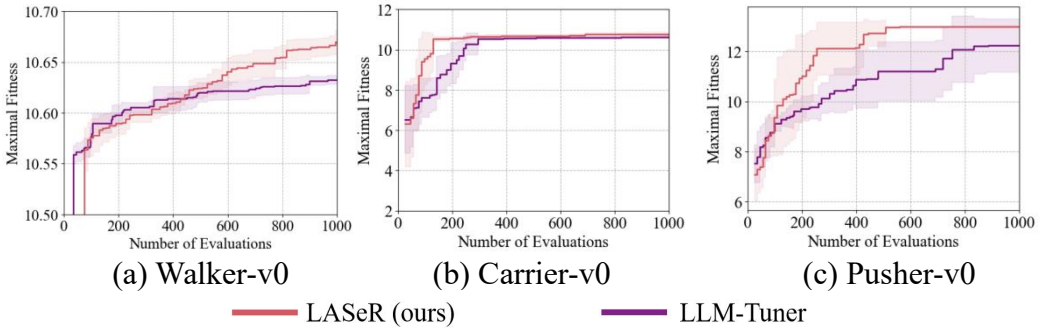
Supplementary material 5: Additional repeated experiments

In response to reviews' suggestions, we conducted two more sets of repeated experiments on LASeR and LLM-Tuner (the most competitive baseline), and now present the averaged results from a total of 5 repeated experiments. As demonstrated in **Supplementary figure 5**, the advantageous optimization efficiency of LASeR remains obvious, and the generally non-overlapping confidence intervals indicate the **statistical significance** of such superiority. The morphological diversity achieved by LASeR remains higher than LLM-Tuner on average (**Supplementary table 2**). However, we note that both LASeR and LLM-Tuner exhibit relatively high variability in their diversity outcomes, suggesting that even more repetitions would be needed to establish statistical significance. To this end, we will continue with additional repeated experiments.



Supplementary figure 5. Comparison between LASeR (red) and LLM-Tuner (purple) based on five repeated experiments, in terms of optimization efficiency. We will continue with the remaining baselines and include the complete results in our paper once they are available.

Supplementary table 2. Comparison between LASeR and LLM-Tuner in terms of morphological diversity, based on five repeated experiments. The results are reported as mean ± standard deviation.

	Walker-v0	Carrier-v0	Pusher-v0
LASeR (ours)	**23.09 (5.96)**	**20.87 (4.77)**	**20.90 (9.89)**
LLM-Tuner	11.60 (5.63)	18.26 (6.70)	14.17 (7.88)