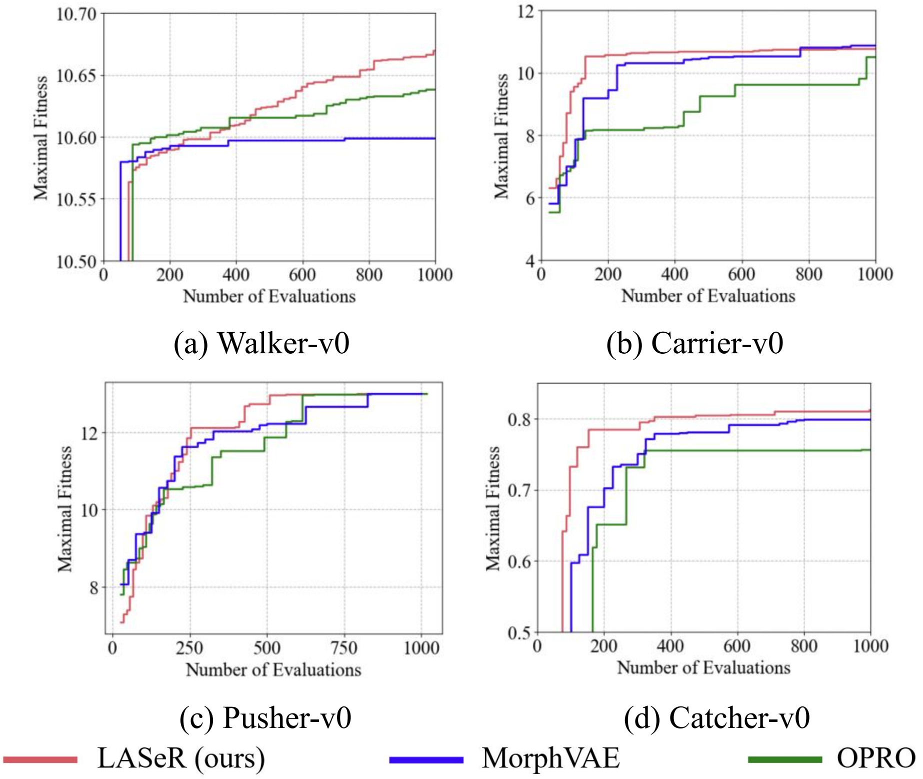
## **Supplementary material 4:** Comparison with two additional baselines

In response to the reviewers' suggestions, we conducted comparative studies between LASeR and two additional baseline algorithms. The first one is OPRO (Yang et al., 2024), another evolutionary strategy that uses LLMs as search operators, which we adapted for voxel-based soft robot (VSR) design. The second one is MorphVAE (Song et al., 2024), a state-of-the-art co-design algorithm that does not employ LLMs but is also developed on the EvoGym platform.

As shown in **Supplementary figure 4**, LASeR consistently outperforms the two baselines in terms of optimization efficiency, reflected by its steeper fitness curves. Here we would like to clarify that we have deliberately chosen **a sufficiently large number of robot evaluations**, i.e. 1000, to hopefully allow all algorithms to converge for fair comparison. This explains why different algorithms end up with rather similar fitness levels. However, in the context of robot design automation, the convergence rate is an important aspect for evaluating design algorithms, as the evaluation of robot designs usually involves computationally expensive control learning, let alone the manufacturing costs of physical robots when deployed in real-world application. In this regard, LASeR exhibits considerable performance gains.



(a) Walker-v0

(b) Carrier-v0

(c) Pusher-v0

(d) Catcher-v0

**Supplementary figure 4.** Comparison between LASeR (red), MorphVAE (blue) and OPRO (green) in terms of optimization efficiency.

As demonstrated in **Supplementary table 1**, LASeR achieves the highest diversity in two out of four tasks, while MorphVAE is dominant in the remaining two tasks. Despite the competitive performance of MorphVAE, we note that LASeR holds distinct advantages: **(a)** with the novel diversity reflection mechanism, LASeR is capable of achieving a **more favorable trade-off between optimization efficiency and diversity**, whereas MorphVAE proposes two variants, each of which focuses on one aspect and compromises the other; **(b)** MorphVAE leverages a variational autoencoder to approximate the high-performing robot distribution and generate offspring solutions, which lacks interpretability. On the contrary, LASeR can instruct an LLM to **explicitly explain its design choices** and thus provide valuable insights of robot design (see **Supplementary material 7**). LASeR is also capable of **more intelligent knowledge transfer** across different tasks, utilizing the reasoning capabilities of LLMs.

**Supplementary table 1.** Comparison between LASeR, MorphVAE and OPRO in terms of diversity. The results are reported as mean ± standard deviation.

| | Walker-v0 | Carrier-v0 | Pusher-v0 | Catcher-v0 |
|---|---|---|---|---|
| LASeR (ours) | **23.09 (5.96)** | 20.87 (4.77) | **20.90 (9.89)** | 6.15 (1.63) |
| MorphVAE | 16.2 (N/A) | **33.16 (16.59)** | 18.18 (12.48) | **11.00 (3.09)** |
| OPRO | 20.77 (7.34) | 5.06 (2.49) | 9.55 (2.92) | N/A |

**Note 1:** In the original paper of MorphVAE, two variants are proposed that focus on either optimization efficiency or diversity. We take an average of them to reflect the overall performance of MorphVAE.

**Note 2:** MorphVAE was implemented 3 times independently for each of the two variants. OPRO was also repeated 3 times except for Catcher-v0, where we were only able to perform a single run of experiment. We plan to continue with the repeated experiments and include the complete results in our paper once they are available.

**References:**

[1] Song, Junru, et al. "MorphVAE: Advancing Morphological Design of Voxel-Based Soft Robots with Variational Autoencoders." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. No. 9. 2024.

[2] Yang, Chengrun, et al. “Large Language Models as Optimizers.” *The Twelfth International Conference on Learning Representations*. 2024.