

Towards Universality: A Performance Benchmark of Pretrained Vision Models in Visuomotor Control

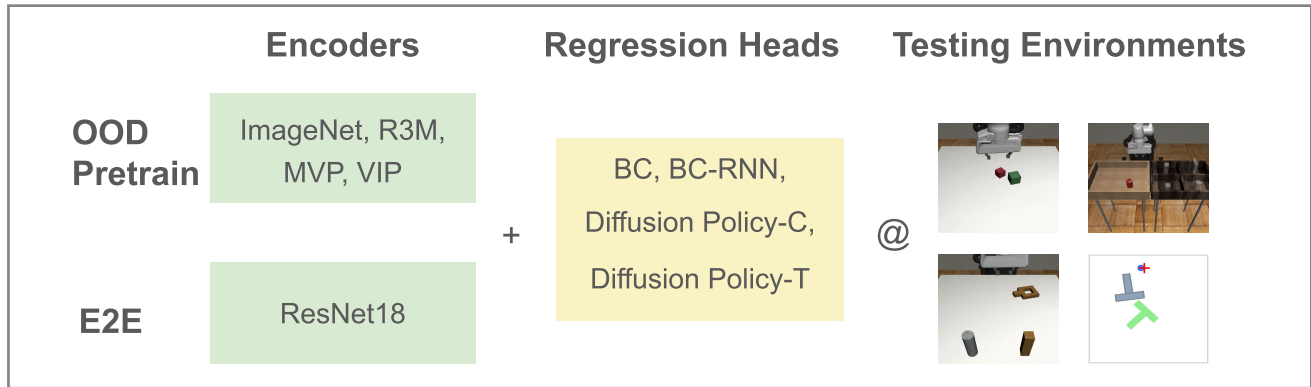
Ruiyu Wang, Zheyu Zhuang, Shutong Jin and Florian T. Pokorny

{ruiyuw, zheyuzh, shutong, fpokorny}@kth.se

Motivation:

The growing trend of developing universal visual encoders for robotic visuomotor control underscores the need for a performance benchmark to evaluate their effectiveness and ensure safe deployment.

We present a performance benchmark between frozen out-of-domain pretrained visual encoders and the state-of-the-art performance achieved by end-to-end policies.



Benchmark Setups: We evaluate the performance of models (BC, BC-RNN, DP-C and DP-T) with frozen pretrained encoders (Imagenet, R3M, MVP and VIP) and their E2E-trained counterparts on four manipulation tasks in simulation, *PushT*, *Can*, *Stack* and *Square*.

Benchmark Results: The results indicate a consistent performance gap, 42% in average, between models trained E2E with those adopting frozen pretrained visual encoders. The gap is narrowed in simple tasks and when using high-capacity models as the regression head.

The universality of visual pretraining for visuomotor control has not been realized to date.

