

Peer Review

Review of: "What's the Move? Hybrid Imitation Learning via Salient Points"

Junpei Zhong¹

1. University of Wollongong, Australia

The paper proposes a hybrid architecture combining point clouds (for spatial awareness) and wrist images (for precision) that leverages the strengths of both modalities. This overcomes the limitations of both methods and shows promising results.

The results also demonstrate scientific rigor as they include four real-world (drawer-opening, cup-stacking, coffee-making, train assembly) and two simulated tasks, covering diverse manipulation challenges. Additionally, generalization tests were included to show its robustness.

One weakness I can see is that it requires explicit human annotation of salient points during data collection and relies on a rule-based mode switch between waypoint and dense policies. This introduces human bias, limits scalability, and complicates deployment in dynamic environments. The manual annotation and guidance also bring bias to the test.

The second problem for me is that the waypoint policy uses point clouds for spatial reasoning, while the dense policy uses wrist images for precision. This separation forces the system to handle two distinct representations rather than learning a unified scene understanding. If the method could use a unified representation to handle both in a single architecture, this would be a more concise and elegant method.

Declarations

Potential competing interests: No potential competing interests to declare.