Scaling Unlabeled Demonstration Data for Latent Action Alignment in Robot Control

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of scaling the size of unlabeled demonstration datasets on the alignment of learned latent actions with ground-truth actions, as measured by metrics like action prediction accuracy. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ViSA-Flow: Accelerating Robot Skill Learning via Large-Scale Video Semantic Action Flow. Research question: What is the impact of scaling the size of unlabeled demonstration datasets on the alignment of learned latent actions with ground-truth actions, as measured by metrics like action prediction accuracy or policy success rates in downstream robot control tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
ViSA-Flow achieves a 31.4% success rate in completing five consecutive tasks, which is almost twice the rate of the next	×	0.06
ViSA-Flow maintains an average sequence length of 2.96 in handling long-horizon manipulation tasks.	×	0.06
Performance degradation from single to sequential tasks is notably less severe for ViSA-Flow (64.7% reduction) compared	×	0.03
Removing the human-video pre-training stage (w/o pre.) leads to a near collapse in performance, with a success rate drop	×	0.05
Removing semantic entity grounding (w/o Seg.) reduces the five-task sequence success rate from 31.4% to 9.6%, and the av	×	0.02
Omitting temporal tracking (w/o Trace.) decreases the average successful length from 2.96 to 2.78.	×	0.01
Excluding manipulator grounding (w/o Hand) yields a modest drop in average successful length from 2.96 to 2.83.	×	0.01
ViSA-Flow with 5% data achieves an 18.4% success rate in completing five consecutive tasks, with an average sequence len	×	0.06
ViSA-Flow with 10% data achieves a 31.4% success rate in completing five consecutive tasks, with an average sequence len	×	0.06
ViSA-Flow with 50% data achieves a 58.8% success rate in completing five consecutive tasks, with an average sequence len	×	0.06

References

- http://arxiv.org/abs/2402.04177v3
- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/2505.01288v3