Latent Action Granularity Effects on Imitation Learning Success Rates from Unlabeled Videos

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of latent action granularity on the final task success rate of imitation learning policies trained on unlabeled demonstration videos. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CLAM: Continuous Latent Action Models for Robot Learning from Unlabeled Demonstrations. Research question: What is the impact of latent action granularity on the final task success rate of imitation learning policies trained on unlabeled demonstration videos?.

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 4.8/10.

3 Results

13 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 4.8/10.

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
CLAM outperforms all baselines and nearly matches the performance of BC with expert data in both state- and image-based	×	0.05
CLAM improves upon the best baseline VPT by more than 2$\times$ average normalized return on the DMControl (locomotion) tasks.	×	0.08
CLAM improves upon the best baseline VPT by around 2-3$\times$ success rate on the Meta-World (manipulation) tasks.	×	0.12
Transformer-CLAM achieves performance close to or even better than that of BC-Expert which uses the same amount of privi	×	0.08
All variants of CLAM outperform the best baseline VPT.	×	0.05
CLAM outperforms state-of-the-art methods in the problem setting where only play data is available as action-labeled dat	✓	0.17
CLAM scales with |Dunlabeled| while supervised IDMs only scale with |Dlabeled|.	×	0.02
VPT learns a suboptimal IDM, underscoring the benefit of latent action models which can leverage vast, unstructured obse	×	0.14
CLAM enables scalable learning from easy-to-collect, cheap play data avoiding the need for expensive task-specific data	×	0.05
Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, 512 feedforward dimension, 2048 num attention heads, 0.1	×	0.02
CALVIN Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, 512 feedforward dimension, 2048 num attention hea	×	0.02
MetaWorld environment has max episode steps of 100, state dim of 39, action dim of 4, image shape of [84, 84, 3], num fr	×	0.03
CALVIN environment has max episode steps of 200, state dim of 39, action dim of 7, image shape of [84, 84, 3], num frame	×	0.02
CLAM is evaluated on DMControl, Meta-World, and CALVIN environments.	×	0.03

References

- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/2309.17203v3
- http://arxiv.org/abs/2407.15840v3