Scaling Unlabeled Video Data Improves CLAM Sample Efficiency Over Latent Action Methods

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the impact of scaling unlabeled video demonstration data on the sample efficiency of CLAM compared to existing unsupervised latent action learning methods. 17 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 scaling unlabeled video demonstration data on the sample efficiency of CLAM compared to existing unsupervised latent action learning methods?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

7 papers retrieved. 17 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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.	×	0.07
CLAM improves around 2-3$\times$ success rate on the MetaWorld (manipulation) tasks compared to the best baseline VPT.	×	0.13
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 [11].	×	0.04
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
CLAM can leverage vast, unstructured observation data to learn latent actions in an unsupervised manner.	×	0.10
CLAM enables scalable learning from easy-to-collect, cheap play data [21] avoiding the need for expensive task-specific	×	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
DMControl tasks include Hopper and HalfCheetah.	×	0.02
MetaWorld tasks include Assembly, Bin Picking, Peg Insert Side, and Shelf Place.	×	0.02
CALVIN tasks include Close Drawer and Slider Left.	×	0.01

References

- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/1906.03248v1
- http://arxiv.org/abs/2211.10412v3