Scaling Unlabeled Video Demonstrations in Continuous Latent Action Models for Multimodal Robot Learning

Assignee Research

June 14, 2026

Abstract

Learning robot policies using imitation learning requires collecting large amounts of costly action-labeled expert demonstrations, which fundamentally limits the scale of training data. A promising approach to address this bottleneck is to harness the abundance of unlabeled observations-e.g., from video demonstrations-to learn latent action labels in an unsupervised way. However, we find that existing methods struggle when applied to complex robot tasks requiring fine-grained motions. We design continuous latent action models (CLAM) which incorporate two key ingredients we find necessary for l

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 convergence rate and final success score of continuous latent action models for multimodal robot learning?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

10 papers retrieved. 14 claims extracted; 9 independently verified. Quality review score: 7.2/10.

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	✓	0.28
CLAM improves upon the best baseline VPT by more than 2$\times$ average normalized return on the DMControl (locomotion) tasks.	✓	0.23
CLAM improves around 2-3$\times$ success rate on the MetaWorld (manipulation) tasks compared to the best baseline VPT.	✓	0.16
BC-AL using action-labeled data does not perform well due to imitating suboptimal demonstrations.	✓	0.22
Transformer-CLAM achieves performance close to or even better than that of BC-Expert which uses the same amount of privi	✓	0.23
All variants of CLAM outperform the best baseline VPT.	✓	0.19
CLAM outperforms state-of-the-art methods in the problem setting where only play data is available as action-labeled dat	✓	0.25
CLAM scales to learn capable robot policies in real-world scenarios.	✓	0.16
CLAM uses a feedforward dimension of 2048, 4 attention heads, and a dropout of 0.1 in the Transformer CLAM model.	×	0.03
CLAM uses a feedforward dimension of 2048, 8 attention heads, and a dropout of 0.1 in the CALVIN Transformer CLAM model.	×	0.03
MetaWorld environment has a max episode steps of 100, state dim of 39, action dim of 4, image shape of [84, 84, 3], num	×	0.03
CALVIN environment has a max episode steps of 200, state dim of 39, action dim of 7, image shape of [84, 84, 3], num fra	×	0.03
CLAM is evaluated on locomotion tasks from the DMControl benchmark (Hopper and HalfCheetah) and manipulation tasks (Asse	✓	0.21
CLAM is also evaluated in CALVIN with the Close Drawer and Slider Left tasks.	×	0.03

References

- http://arxiv.org/abs/2504.11493v1
- http://arxiv.org/abs/2503.00200v3
- http://arxiv.org/abs/2505.04999v1