Disentangled Latent Action Tokens vs Standard Models in Few-Shot Generalization on CALVIN Benchmark

Assignee Research

June 12, 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: How does the performance of DiLA's disentangled latent action tokens compare to standard Latent Action Models on the CALVIN benchmark when evaluated for few-shot generalization using CLAM's policy transfer metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

9 papers retrieved. 14 claims extracted; 9 independently verified. Quality review score: 7.4/10.

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 a	✓	0.26
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
BC-AL using action-labeled data does not perform well due to imitating suboptimal demonstrations.	✓	0.22
CLAM scales with |Dunlabeled| while supervised IDMs only scale with |Dlabeled|.	×	0.13
CLAM can leverage vast, unstructured observation data to learn latent actions in an unsupervised manner.	✓	0.17
CLAM enables scalable learning from easy-to-collect, cheap play data avoiding the need for expensive task-specific data	✓	0.28
The Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, a feedforward dimension of 2048, 8 attention heads,	×	0.03
The CALVIN Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, a feedforward dimension of 2048, 8 attention	×	0.03
The MetaWorld environment has a maximum of 100 episode steps, a state dimension of 39, an action dimension of 4, an imag	×	0.03
The CALVIN environment has a maximum of 200 episode steps, a state dimension of 39, an action dimension of 7, an image s	×	0.03
The evaluation environments in simulation include locomotion tasks from the DMControl benchmark (Hopper and HalfCheetah)	✓	0.21

References

- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/2605.15725v1
- http://arxiv.org/abs/2508.11117v1