Do multimodal video-language models like ViPRA demonstrate robust zero-shot transfer to unseen robotic manipul

Assignee Research

June 10, 2026

Abstract

Can we turn a video prediction model into a robot policy? Videos, including those of humans or teleoperated robots, capture rich physical interactions. However, most of them lack labeled actions, which limits their use in robot learning. We present Video Prediction for Robot Actions (ViPRA), a simple pretraining-finetuning framework that learns continuous robot control from these actionless videos. Instead of directly predicting actions, we train a video-language model to predict both future visual observations and motion-centric latent actions, which serve as intermediate representations of s

1 Introduction

This paper examines: ViPRA: Video Prediction for Robot Actions. Research question: Do multimodal video-language models like ViPRA demonstrate robust zero-shot transfer to unseen robotic manipulation tasks when evaluated on standard imitation learning benchmarks like ViZDoom or RoboTHOR compared to specialized imitation learning models?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.4/10.

3 Results

4 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 5.4/10.

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
ViPRA extracts motion-centric latent action sequences from large-scale actionless videos.	×	0.13
ViPRA pretrains a video-language model to jointly predict future visual observations and latent action chunks.	✓	0.16
ViPRA finetunes a flow matching decoder to map latent actions to smooth, continuous action chunks with minimal labeled d	×	0.14
ViPRA predicts state transitions through video prediction and outputs a sequence of fine-grained motion-centric latent a	×	0.13
ViPRA incorporates optical flow consistency as an additional supervision signal, promoting physically plausible and moti	×	0.08
ViPRA's pretraining leverages both unlabeled human and robot videos, enabling generalization across embodiments.	×	0.09
ViPRA uses a flow matching decoder for finetuning on teleoperated robot demonstrations.	×	0.13
ViPRA's decoder aligns latent transitions with embodiment-specific motor behaviors.	×	0.04
ViPRA's policy can support control rates up to 22 Hz.	×	0.08
ViPRA demonstrates empirical gains of 16% on the SIMPLER benchmark.	×	0.07
ViPRA demonstrates empirical gains of 13% on real-world tasks over the strongest prior continuous control baselines.	×	0.11
ViPRA uses human videos without action labels for pretraining.	×	0.06
ViPRA uses robot videos without action labels for pretraining.	×	0.09
ViPRA predicts future visual states and motion-centric latent actions within a unified video-language model.	✓	0.17
ViPRA integrates flow matching and action chunking to enable smooth, high-frequency continuous control.	×	0.14

References

- http://arxiv.org/abs/2001.07798v4
- http://arxiv.org/abs/2511.07732v2
- http://arxiv.org/abs/2402.16349v2