Inference Latency of Fine-Tuned Video Encoders on Synthetic vs. Human Gesture Data

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the inference latency of large pre-trained video encoders change when fine-tuned on synthetic gesture data versus human-annotated datasets across varying batch sizes. In this work, we explore the possibility of using synthetically generated data for video-based gesture recognition with large pre-trained models. We consider whether these models have sufficiently robust and expressive representation spaces to enable "training-free". 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: An Evaluation of Large Pre-Trained Models for Gesture Recognition using Synthetic Videos. Research question: How does the inference latency of large pre-trained video encoders change when fine-tuned on synthetic gesture data versus human-annotated datasets across varying batch sizes?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

12 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The RoCoG-v2 dataset consists of 7 gesture categories.	×	0.11
The synthetic training data consists of 44K videos.	×	0.12
The small real dataset consists of 203 videos.	×	0.06
K=3 is used for all KNN classification experiments.	×	0.05
The UMT model is pre-trained on K710 videos.	×	0.08
Eight frames are sampled from each video using the TSN frame-sampling strategy.	×	0.04
The ViCLIP model is pre-trained on a filtered version of the InternVid dataset with 10M video-text pairs.	×	0.07
The VideoMAE models are pre-trained on a larger dataset of 1.3B videos.	×	0.08
The KNN accuracy for ViT-B/16 UMT K710 is 18.2% for synthetic train and 31.2% for real train.	×	0.02
The KNN accuracy for ViT-B/16 ViCLIP InternVid FLT-10M is 19.2% for synthetic train and 40.4% for real train.	×	0.02
The KNN accuracy for ViT-B/16 UMT K710 + K400 is 38.4% for synthetic train and 45.5% for real train.	×	0.02
The KNN accuracy for ViT-B/16 VideoMAE UnlabeledHybrid SSv2 is 43.4% for synthetic train and 68.7% for real train.	×	0.02
The KNN accuracy for ViT-L/16 VideoMAE UnlabeledHybrid SSv2 is 64.6% for synthetic train and 71.7% for real train.	×	0.02
The zero-shot classification accuracy for ViCLIP-B InternVid FLT-10M with original text descriptions is 25.3%.	×	0.13
The zero-shot classification accuracy for ViCLIP-B InternVid FLT-10M with transformed text descriptions is 26.3%.	×	0.13

References

- http://arxiv.org/abs/2410.02152v1

- http://arxiv.org/abs/2410.21676v4
- http://arxiv.org/abs/2602.09439v1