Point-Embedded Transformers vs. Traditional Attention in Multi-View Hand Mesh Reconstruction

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How does the integration of a point-embedded transformer in multi-view hand mesh reconstruction compare to traditional attention mechanisms in terms of inference efficiency on benchmark datasets like. This work introduces a novel and generalizable multi-view Hand Mesh Reconstruction (HMR) model, named POEM, designed for practical use in real-world hand motion capture scenarios. The advances of the POEM model consist of two main aspects. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multi-view Hand Reconstruction with a Point-Embedded Transformer. Research question: How does the integration of a point-embedded transformer in multi-view hand mesh reconstruction compare to traditional attention mechanisms in terms of inference efficiency on benchmark datasets like MPI INTERACT?.

2 Methodology

Systematic literature search across multiple databases yielded 6 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

6 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
POEMv1 is a model trained under the configuration of Model Architecture Analysis.	×	0.03
POEMv2 is a model trained under the configuration of Scalability Assessment.	×	0.06
POEMv2 is designed to scale in terms of both the parameters count and data used for training.	×	0.04
POEMv2 is trained on five multi-view datasets plus one monocular dataset in a randomly mixing patterns.	×	0.08
POEMv2 is evaluated on each single test set to assess its performance.	×	0.06
POEMv2 exhibits the most significant improvement on HO3D-Mv.	×	0.03
POEMv2 leverages the generalization capability of a larger model and larger dataset to better adapt to the camera config	×	0.02
POEMv2 does not perform as well on OakInk-Mv.	×	0.00
POEMv2-param is a variant of POEMv2 that replaces the output of the Transformer decoder with the MANO model parameters θ	×	0.02
POEMv2-param provides rotation angles for each joint, making it more suitable for VR/AR applications that require hand a	×	0.03
The results of POEMv2-param are presented in Tab. 4 row 7,12,17.	×	0.01
The methodology involves aligning observations from different camera spaces into a common representation space and fusin	×	0.03
The methodology involves designing network architectures in the common representation space to make predictions.	×	0.01
The alignment process uses a set of static 3D points in world space as the basis for representing the variable hand surf	×	0.08
The 3D basis points lie within the common view space of all cameras and encompass the hand for reconstruction.	✓	0.16

References

- http://arxiv.org/abs/2408.10581v2
- http://arxiv.org/abs/2406.16137v1
- http://arxiv.org/abs/2304.04038v2