Multimodal Transformer Scaling and Human Attention Alignment in Fine-Grained Spatial Tasks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 20 peer-reviewed papers addressing the following research question: To what extent does model size scaling in multimodal transformers (e.g., ViT, CLIP vs. small-scale CNN-based models) affect the alignment of synthetic metrics with human attention benchmarks in tasks. Tactile sensing provides local essential information that is complementary to visual perception, such as texture, compliance, and force. Despite recent advances in visuotactile representation learning, challenges remain in fusing these modalities and generalizing across tasks. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ViTaPEs: Visuotactile Position Encodings for Cross-Modal Alignment in Multimodal Transformers. Research question: To what extent does model size scaling in multimodal transformers (e.g., ViT, CLIP vs. small-scale CNN-based models) affect the alignment of synthetic metrics with human attention benchmarks in tasks requiring fine-grained spatial understanding, such as tumor delineation?.

2 Methodology

Systematic literature search across multiple databases yielded 20 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

20 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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.

References

- https://www.semanticscholar.org/paper/216a3f701251ae381e634c2e64ae1b845092e738
- https://arxiv.org/abs/2604.22498
- https://arxiv.org/abs/2505.20032