Self-Attention Layer Depth and Zero-Shot Cross-Modal Retrieval Performance

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the depth of self-attention layers in a multimodal model influence zero-shot cross-modal retrieval performance on out-of-domain datasets like Conceptual Captions. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners. Research question: How does the depth of self-attention layers in a multimodal model influence zero-shot cross-modal retrieval performance on out-of-domain datasets like Conceptual Captions?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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

- http://arxiv.org/abs/2212.04979v3
- http://arxiv.org/abs/2502.06338v1
- http://arxiv.org/abs/2605.22903v1