Human Attention Benchmark vs. Synthetic Metrics in Model Performance Correlation

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the human attention benchmark compare to existing synthetic attention evaluation metrics in terms of correlation with model performance on downstream tasks. Many computational models of visual attention have been created from a wide variety of different approaches to predict where people look in images. Each model is usually introduced by demonstrating performances on new images, and it is hard to make immediate comparisons between. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Benchmark of Computational Models of Saliency to Predict Human Fixations. Research question: How does the human attention benchmark compare to existing synthetic attention evaluation metrics in terms of correlation with model performance on downstream tasks?.

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 8.8/10.

3 Results

12 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The benchmark dataset contains 300 natural images with eye tracking data from 39 observers	✓	0.21
10 models were evaluated for performance at predicting ground truth fixations using three different metrics	✓	0.24
The Judd et al. and Graph-based visual saliency models perform best	✓	0.29
Models with blurrier maps and models that include a center bias perform well	✓	0.28
Human performance increases with the number of humans to a limit	✓	0.22
The benchmark provides a way for people to submit new models for evaluation online	✓	0.21

References

- https://openalex.org/W1517086206
- https://doi.org/10.1109/5.726791
- https://doi.org/10.18653/v1/2023.newsum-1.1