Large Multimodal Models Enhance Causal Graph Alignment in Visual Scenes

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Do larger multimodal models demonstrate improved alignment with ground-truth causal graphs in complex visual scenes compared to smaller variants on standardized benchmarks. 12 claims were extracted from source literature; 1 was 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: CausalVLBench: Benchmarking Visual Causal Reasoning in Large Vision-Language Models. Research question: Do larger multimodal models demonstrate improved alignment with ground-truth causal graphs in complex visual scenes compared to smaller variants on standardized benchmarks?.

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. 12 claims extracted; 1 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.

5 Extracted Claims

Claim	Verified	Confidence
LVLMs have shown tremendous potential in tasks such as recognition, grounding, and VQA.	×	0.11
Zong et al. (2025) benchmarked LVLMs on diverse tasks with multi-modal in-context learning (ICL).	×	0.12
It is significantly more challenging for AI systems to learn causal relationships from high-dimensional data such as ima	×	0.06
Chen et al. (2024) evaluated the performance of LVLMs on causally-motivated VQA tasks where the causal relationships wer	×	0.06
The evaluation by Chen et al. (2024) focused on scene-specific relations originating from human-object interactions.	×	0.03
The authors focus on the ability of LVLMs to perform formal visual causal reasoning with systems described by determinis	×	0.13
The visual causal reasoning task is formulated as the ability of LVLMs to disentangle causal variables and reason about	×	0.14
The authors construct a benchmark, CausalVL-Bench, encompassing three representative tasks: causal structure inference, i	✓	0.25
The authors study the effect of prompting without the causal graph, demonstration.	×	0.04
The Structural Hamming Distance (SHD) is computed with respect to the ground-truth causal graph.	×	0.09
The average exact match accuracy of model predictions is reported.	×	0.02
In the intervention target prediction task, the number of targets is evaluated.	×	0.08

References

- http://arxiv.org/abs/2506.11034v2
- http://arxiv.org/abs/2301.05169v2
- http://arxiv.org/abs/1911.07420v1