How does the robustness of structural-causal-augmented multimodal models compare to traditional augmentation m

Assignee Research

June 10, 2026

Abstract

Large Vision-Language Models (LVLMs) are capable of handling diverse data types such as imaging, text, and physiological signals, and can be applied in various fields. In the medical field, LVLMs have a high potential to offer substantial assistance for diagnosis and treatment. Before that, it is crucial to develop benchmarks to evaluate LVLMs' effectiveness in various medical applications. Current benchmarks are often built upon specific academic literature, mainly focusing on a single domain, and lacking varying perceptual granularities. Thus, they face specific challenges, including limited

1 Introduction

This paper examines: GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. Research question: How does the robustness of structural-causal-augmented multimodal models compare to traditional augmentation methods in cross-domain transfer tasks on MMBench subsets, using accuracy as the evaluation metric?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

10 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
GMAI-MMBench consists of 284 diverse clinical-related datasets from worldwide sources, covering 38 modalities.	×	0.11
GMAI-MMBench features 18 clinical VQA tasks and 18 clinical departments, meticulously organized into a lexical tree.	×	0.15
GMAI-MMBench offers interactive methods spanning from image to region level, providing varying degrees of perceptual det	×	0.07
Medical-Diff-VQA consists of 70K samples and covers 7 tasks.	×	0.06
PathVQA consists of 6K samples and covers 7 tasks.	×	0.01
Cholec80-VQA consists of 9K samples and covers 2 tasks.	×	0.02
VQA-RAD consists of 3K samples and covers 11 tasks.	×	0.01
RadBench consists of 137K samples and covers 5 tasks.	×	0.01
MMMU (H & M) consists of 2K samples and covers 5 tasks.	×	0.01
SLAKE consists of 2K samples and covers 10 tasks.	×	0.01
OmniMedVQA consists of 128K samples and covers 5 tasks.	×	0.01
GMAI-MMBench consists of 26K samples and covers 18 tasks.	×	0.07
GMAI-MMBench includes data from 284 datasets from both public and hospital sources.	×	0.08
Med-Flamingo achieved an overall test score of 12.74.	×	0.01
LLaVA-Med achieved an overall test score of 20.54.	×	0.01
Qilin-Med-VL-Chat achieved an overall test score of 22.34.	×	0.01
RadFM achieved an overall test score of 22.95.	×	0.02
MedDr achieved an overall test score of 41.95.	×	0.01

References

- http://arxiv.org/abs/2503.14478v2
- http://arxiv.org/abs/2408.03361v7
- http://arxiv.org/abs/2306.13394v5