Curriculum-Based Multi-Task Learning Enhances Cross-Domain Generalization in Large Multimodal Models

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does curriculum-based multi-task learning affect the cross-domain generalization accuracy of large multimodal models on the RadNet benchmark compared to standard joint training. 12 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: MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. Research question: How does curriculum-based multi-task learning affect the cross-domain generalization accuracy of large multimodal models on the RadNet benchmark compared to standard joint training?.

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 3.8/10.

3 Results

10 papers retrieved. 12 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.

5 Extracted Claims

Claim	Verified	Confidence
The random accuracies of the two metrics (accuracy and accuracy+) are equal to 50% and 25%, respectively.	×	0.03
The full scores of perception and cognition are 2000 and 800, respectively.	×	0.05
The images for coarse-grained recognition are sampled from COCO, but the instruction-answer pairs are manually construct	×	0.07
In each perception subtask of existence, count, color, and position, there are 30 images with 60 instruction-answer pair	×	0.07
The fine-grained recognition subtasks consist of recognizing movie posters, celebrities, scenes, landmarks, and artworks	×	0.01
The images for fine-grained recognition are from publicly available datasets and all of the instructions are manually de	×	0.04
The images for OCR are sampled from [33] and all of the instruction-answer pairs are manually designed.	×	0.11
WeMM has the highest score of 1621.66 in the first benchmark table.	×	0.01
GPT-4V has the highest score of 517.14 in the second benchmark table.	×	0.02
Otter, Lynx, WeMM, Muffin, and SPHINX all have the highest score of 195.00 in the third benchmark table.	×	0.01
Muffin has the highest score of 163.33 in the fourth benchmark table.	×	0.01
Lion and SPHINX both have the highest score of 153.33 in the fifth benchmark table.	×	0.01

References

- http://arxiv.org/abs/2507.20499v2
- http://arxiv.org/abs/2306.13394v5
- http://arxiv.org/abs/2106.09017v1