Curriculum-Based Multi-Task Learning Effects on Large Multimodal Model Latency and Throughput in RadNet Benchmarks

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of curriculum-based multi-task learning on the inference latency and throughput of large multimodal models evaluated on the RadNet medical image-text dataset. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Exploring Transfer Learning in Medical Image Segmentation using Vision-Language Models. Research question: What is the impact of curriculum-based multi-task learning on the inference latency and throughput of large multimodal models evaluated on the RadNet medical image-text dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

14 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
VLSMs adapt better to non-radiology images in Zero-Shot Setting (ZSS).	×	0.03
CRIS and CLIPSeg barely work in ZSS for radiology images except for CRIS in the BUSI dataset but get a Dice score in the	×	0.03
Adding more attributes to the prompt generally improved performance, but the gain is inconsistent across prompts and dat	×	0.03
CRIS performs better on endoscopy datasets when prompts contain image-specific attributes (size, number, and location; P	×	0.04
CRIS degrades with non-image-specific attributes added (P7, P8, P9) in the ZSS.	×	0.02
Prompts with general descriptions (P8 and P9) achieve the highest performance on the DFU 2022 dataset.	×	0.02
The DSC variation across prompt type is minimal in the finetuned setting for all the models.	×	0.03
Prompt with only class name (P1) improves segmentation performance in radiology datasets for all four VLSMs.	×	0.10
CRIS' performance almost saturates after adding the class name and mask shape (P2).	×	0.02
BiomedCLIPSeg and BiomedCLIPSeg-D, despite being based on a VLM pretrained on medical data, consistently perform poorly	×	0.04
Four medical VLSMs were created using CLIP and BiomedCLIP: CLIPSeg, CRIS, BiomedCLIPSeg-D, and BiomedCLIPSeg.	×	0.06
CLIPSeg accommodates both CNN and ViT backbones, whereas CRIS only supports a CNN-based CLIP backbone.	×	0.02
BiomedCLIPSeg-based models include transformer-based backbones for both the encoders.	×	0.02

References

- `http://arxiv.org/abs/2410.21696v4`

- http://arxiv.org/abs/2308.07706v3
- http://arxiv.org/abs/2108.12828v4