EfficientViT Backbone Integration in PaLI for Image-Text Matching on COCO

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does replacing the ViT backbone with EfficientViT in PaLI affect the inference latency and memory efficiency during Image-Text Matching tasks on the COCO dataset, as measured by FPS and GPU. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. Research question: How does replacing the ViT backbone with EfficientViT in PaLI affect the inference latency and memory efficiency during Image-Text Matching tasks on the COCO dataset, as measured by FPS and GPU memory usage?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

13 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 5.8/10.

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
EfficientViT-M5 surpasses MobileNetV3-Large by 1.9% in accuracy, while getting 40.4% and 45.2% higher throughput on Nvid	✓	0.31
EfficientViT-M2 achieves 1.8% superior accuracy compared to MobileViT-XXS, while running 5.8$\times$/3.7$\times$ faster on the GPU/CPU	✓	0.16
EfficientViT-M5 gets 77.1% top-1 accuracy on ImageNet with throughput of 10,621 images/s on an Nvidia V100 GPU and 56.8	×	0.09
EfficientViT-M2 gets 70.8% accuracy, surpassing MobileViT-XXS by 1.8%, while running 5.8$\times$/3.7$\times$ faster on the GPU/CPU, an	×	0.10
EfficientViT-M2 model runs 2.3$\times$ faster than MobileViT-XXS on Apple A13 Bionic chip in iPhone 11.	×	0.06
EfficientViT-M3 achieves 73.4% accuracy with 16644 parameters, 96.4 FLOPs, and 120.8 inference time on GPU.	×	0.03
EfficientViT-M4 achieves 74.3% accuracy with 15914 parameters, 88.5 FLOPs, and 108.6 inference time on GPU.	×	0.03
EfficientViT-M5 achieves 77.1% accuracy with 10621 parameters, 56.8 FLOPs, and 62.5 inference time on GPU.	×	0.04
EfficientViT-M4$\uparrow$384 achieves 79.8% accuracy with 3986 parameters, 15.8 FLOPs, and 22.6 inference time on GPU.	×	0.03
EfficientViT-M5$\uparrow$512 achieves 80.8% accuracy with 2313 parameters, 8.3 FLOPs, and 10.5 inference time on GPU.	×	0.04

References

- http://arxiv.org/abs/2305.07027v1
- http://arxiv.org/abs/2205.14756v6
- http://arxiv.org/abs/2406.15306v1