EfficientViT Backbone Integration in PaLI: Accuracy Trade-offs on Image-Text and Text-Only Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of swapping the ViT backbone in PaLI with a more efficient architecture (e.g., EfficientViT) on image-text accuracy while maintaining text-only accuracy on LAVIS benchmarks. 9 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: ViT-5: Vision Transformers for The Mid-2020s. Research question: What is the impact of swapping the ViT backbone in PaLI with a more efficient architecture (e.g., EfficientViT) on image-text accuracy while maintaining text-only accuracy on LAVIS benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 9 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
ViT-5 improves the representational capacity of ViT backbones and narrows the architectural gap between vision models an	×	0.06
ViT-5 aims to facilitate more efficient construction of multimodal systems and inspire the development of unified Transf	×	0.05
ViT-5-Small achieves 82.18% top-1 accuracy on ImageNet.	×	0.04
ViT-5-Base achieves 84.15% top-1 accuracy on ImageNet.	×	0.11
ViT-5-Large achieves 84.82% top-1 accuracy on ImageNet.	×	0.07
Replacing LayerScale by post-norm yields highly similar ImageNet top-1 accuracy across different model sizes.	×	0.03
ViT-5 with LayerScale achieves 82.16% top-1 accuracy on ImageNet for the small model size.	×	0.04
ViT-5 with LayerScale achieves 84.16% top-1 accuracy on ImageNet for the base model size.	×	0.09
ViT-5 with LayerScale achieves 84.86% top-1 accuracy on ImageNet for the large model size.	×	0.06

References

- http://arxiv.org/abs/2602.08071v1
- http://arxiv.org/abs/2209.09019v1
- http://arxiv.org/abs/2310.19909v2