Multimodal Model Scaling and Accuracy Trade-offs in Image-Text vs. Text-Only Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the scaling of parameter count affect the trade-off between image-text and text-only accuracy in multimodal models like PaLI and IDEFICS when evaluated on LAVIS benchmarks. 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: PaLI: A Jointly-Scaled Multilingual Language-Image Model. Research question: How does the scaling of parameter count affect the trade-off between image-text and text-only accuracy in multimodal models like PaLI and IDEFICS when evaluated 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 3.5/10.

3 Results

12 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
PaLI-3B and PaLI-15B checkpoints are fine-tuned and evaluated at 490$\times$490 resolutions.	×	0.02
PaLI-17B checkpoint is fine-tuned and evaluated at 588$\times$588 resolution unless otherwise stated.	×	0.01
Cross-entropy loss is used for fine-tuning all benchmarks.	×	0.02
PaLI outperforms the latest SOTA trained with cross-entropy loss on COCO Captions, achieving a CIDEr score of 149.1.	×	0.03
PaLI-17B achieves a CIDEr score of 124.4 on the NoCaps test set.	×	0.04
GIT2 achieves CIDEr scores of 124.2, 125.5, and 122.3 on in-domain, near-domain, and out-of-domain splits of the NoCaps	×	0.02
PaLI-17B achieves CIDEr scores of 121.1, 124.4, and 126.7 on in-domain, near-domain, and out-of-domain splits of the NoC	×	0.02
PaLI-17B outperforms all prior models on recognizing and describing long-tail objects outside of COCO's domain.	×	0.04
PaLI-17B is fine-tuned on TextCaps and VizWiz-Cap using OCR strings generated by publicly available automatic service.	×	0.03
PaLI uses a text encoder-decoder Transformer at its core.	×	0.07
PaLI includes vision as input by feeding the text encoder with a sequence of visual tokens from a Vision Transformer.	×	0.08
No pooling is applied to the output of the Vision Transformer before passing the visual tokens to the encoder-decoder mo	×	0.07
PaLI reuses previously trained unimodal checkpoints.	×	0.04

References

- `http://arxiv.org/abs/2406.03496v1`

- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2209.06794v4