Synthetic Caption Diversity Effects on Vision-Language Model Performance in Remote Sensing

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does varying the diversity of synthetic captions in vision-language models impact performance on remote sensing benchmarks like EuroSAT or RSICD. 15 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Modeling Caption Diversity in Contrastive Vision-Language Pretraining. Research question: How does varying the diversity of synthetic captions in vision-language models impact performance on remote sensing benchmarks like EuroSAT or RSICD?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

16 papers retrieved. 15 claims extracted; 4 independently verified. Quality review score: 4.8/10.

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
Llip with a ViT-G/14 backbone achieves a zero-shot top-1 accuracy of 83.5% on ImageNet.	✓	0.17
Llip outperforms SigLIP by +0.3% in zero-shot top-1 accuracy on ImageNet.	✓	0.16
Llip processes 4 times fewer samples during pre-training compared to SigLIP.	×	0.06
Llip outperforms EVA-CLIP by 2.5% in zero-shot top-1 accuracy on ImageNet.	✓	0.16
EVA-CLIP is pre-trained with a ViT-E/14 backbone which has 2.5 times more parameters than the ViT-G/14 used by Llip.	×	0.07
DFN achieves a higher zero-shot top-1 accuracy (84.4%) on ImageNet than Llip (83.5%).	×	0.10
DFN is trained on a dataset of 5 billion curated samples.	×	0.03
DFN uses an input image resolution of 378 pixels, whereas Llip uses 224 pixels.	×	0.02
Llip outperforms MetaCLIP (ViT-G/14) by +1.4% in zero-shot top-1 accuracy on ImageNet.	✓	0.16
Llip achieves the best average performance across 22 standard zero-shot classification benchmarks compared to OpenCLIP,	×	0.12
Llip reaches the best performance in 19 out of 22 classification tasks.	×	0.04
On the Food-101 benchmark, Llip64 (ViT-B/32) achieves an accuracy of 70.4%.	×	0.01
On the CIFAR10 benchmark, Llip64 (ViT-B/32) achieves an accuracy of 84.1%.	×	0.02
On the SUN397 benchmark, Llip64 (ViT-B/32) achieves an accuracy of 80.8%.	×	0.01
The singular value spectrum of the covariance matrix of visual features indicates that Llip's representation is more exp	×	0.10

References

- http://arxiv.org/abs/2505.14361v1
- http://arxiv.org/abs/2405.00740v4
- http://arxiv.org/abs/2508.11919v3