Synthetic Image-Text Alignment and Cross-Domain Generalization in Vision-Language Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the alignment of synthetic image-text pairs generated by multimodal models affect the cross-domain generalization performance on vision-language benchmarks like VQAv2 or COCO-Caption. 17 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Grounding Synthetic Data Generation With Vision and Language Models. Research question: How does the alignment of synthetic image-text pairs generated by multimodal models affect the cross-domain generalization performance on vision-language benchmarks like VQAv2 or COCO-Caption?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.9/10.

3 Results

15 papers retrieved. 17 claims extracted; 2 independently verified. Quality review score: 4.9/10.

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
The ARAS400k dataset is available at zenodo.org/records/18890661.	×	0.12
The code base for the project is available at github.com/caglarmert/ARAS400k.	×	0.06
Models trained on the experimental dataset (real + synthetic) consistently outperform those trained on real data alone.	✓	0.17
The combined real and synthetic dataset is particularly effective in addressing class-imbalance for under-represented ca	×	0.05
SynthCLIP and SynGround show that models trained exclusively on synthetic image-caption pairs can achieve performance co	×	0.13
Denoising Diffusion Probabilistic Models often require longer training and inference times compared to GAN architectures	×	0.04
GAN models tend to suffer from mode collapse, vanishing gradients, non-converging or unstable training, and sensitivity	×	0.01
The CLIP-Score metric aligns more with human assessment than standard metrics for reference-free caption evaluation.	×	0.05
The generative models were trained exclusively on a fixed training partition containing 80,182 real samples.	×	0.09
The study utilized an implementation of StyleGAN3, a U-Net based discriminator, and SPADE architectures.	×	0.02
GAN models were trained until their training FID score reached a plateau.	×	0.03
The ARAS400k dataset consists of 100,240 real images.	×	0.07
The ARAS400k dataset consists of 300,000 synthetic images.	×	0.09
Each image in the ARAS400k dataset is paired with semantic segmentation maps.	✓	0.18
The ARAS400k dataset contains over 2 million descriptive captions.	×	0.02
Data was acquired from ESA Sentinel-2 RGB-NIR true-color images.	×	0.03
Data was acquired from WorldCover 2021 land cover data.	×	0.02

References

- http://arxiv.org/abs/2603.09625v2
- http://arxiv.org/abs/2201.05729v3
- http://arxiv.org/abs/2605.00721v1