Grounded Synthetic Data Generation in Vision-Language Models for Cross-Domain Transfer

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does grounding synthetic data generation in vision-language models improve cross-domain transfer performance compared to latent feature similarity metrics. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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: To what extent does grounding synthetic data generation in vision-language models improve cross-domain transfer performance compared to latent feature similarity metrics?.

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 4.5/10.

3 Results

15 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
The ARAS400k dataset is available at zenodo.org/records/18890661 and the code base at github.com/caglarmert/ARAS400k.	×	0.09
Models trained on a combination of real and synthetic data consistently outperform those trained on real data alone, par	×	0.15
SynthCLIP and SynGround show that models trained exclusively on synthetic image-caption pairs can achieve performance co	×	0.13
Combining detail attention sampling with a teacher-student network effectively integrates local and global features, yie	×	0.06
Denoising Diffusion Probabilistic Models often require longer training and inference times compared to GAN architectures	×	0.05
GAN models tend to have problems such as mode collapse, vanishing gradients, non-converging or unstable training, and se	×	0.03
The CLIP-Score metric aligns more with human assessment, enabling reference-free caption evaluation.	×	0.02
The generative models were trained exclusively on a fixed training partition containing 80,182 real samples.	×	0.08
The ARAS400k dataset consists of 100,240 real images and 300,000 synthetic images, each paired with semantic segmentatio	✓	0.17
The automated pipeline for context-aware caption generation and evaluation utilizes composition statistics available fro	×	0.07
The data was acquired from ESA Sentinel-2 RGBNIR true-color images and WorldCover 2021.	×	0.02

References

- http://arxiv.org/abs/2603.09625v2
- http://arxiv.org/abs/2411.15497v3
- http://arxiv.org/abs/2406.04295v2