Synthetic Data Generation Techniques and Multimodal LLM Alignment in CALVIN Benchmark

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do different synthetic data generation techniques (e.g., diffusion models vs. GANs) influence the alignment of multimodal LLMs in the CALVIN benchmark, measured by both task success rate and. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Everything to the Synthetic: Diffusion-driven Test-time Adaptation via Synthetic-Domain Alignment. Research question: How do different synthetic data generation techniques (e.g., diffusion models vs. GANs) influence the alignment of multimodal LLMs in the CALVIN benchmark, measured by both task success rate and human preference metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

13 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
Fine-tuning the source model on the synthetic data generated by the Mix of Diffusion process improves performance by 5.5	✓	0.17
The diffusion synthetic data xsyn_0,u and source data xsrc_0 exhibit no noticeable visual differences across different t	×	0.09
SDA consistently outperforms all baseline methods across different model architectures and sizes on ImageNet-C.	×	0.04
SDA improves accuracy by 2.5%-2.9% compared to DDA.	×	0.01
SDA achieves an improvement of 2.2% with ConvNeXt-T compared to the recent SOTA GDA.	×	0.02
Diffusion-driven methods (SDA, DDA, and GDA) demonstrate superior performance compared to the model adaptation method, M	×	0.10
DiffPure presents worse results since it is primarily designed for adversarial attacks.	×	0.04
SDA surpasses DiffPure in all 15 corruption types on ImageNet-C.	×	0.01
SDA achieves an average accuracy of 51.9% on ImageNet-C, which is 2.5% higher than DDA.	×	0.01
SDA improves performance by 6.0% for Swin-B and 5.5% for ConvNeXt-B when fine-tuned on synthetic data generated by the M	×	0.11

References

- http://arxiv.org/abs/2308.12898v2
- http://arxiv.org/abs/2406.04295v2

- http://arxiv.org/abs/2503.14504v2