Cross-Domain Data Augmentation for Multimodal Language Models in VQA and OCR Tasks

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Does cross-domain adaptation of data augmentation strategies from image-based models to multimodal language models enhance performance on tasks like VQA and OCR-VQA, and how does this compare to. The field of computer vision has experienced significant advancements through scalable vision encoders and multimodal pre-training frameworks. However, existing approaches often treat vision encoders and large language models (LLMs) as independent modules, limiting the. 14 claims were extracted from source literature; 1 was 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: Hierarchical Pre-Training of Vision Encoders with Large Language Models. Research question: Does cross-domain adaptation of data augmentation strategies from image-based models to multimodal language models enhance performance on tasks like VQA and OCR-VQA, and how does this compare to text-only approaches?.

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 3.5/10.

3 Results

15 papers retrieved. 14 claims extracted; 1 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
HIVE is evaluated on CIFAR-10, CIFAR-100, ImageNet-1K, Tiny-ImageNet, Food-101, Stanford Cars, Oxford-IIIT Pets, and Cal	×	0.02
HIVE is evaluated on MME, GQA, OK-VQA, and ScienceQA for vision-language model (VLM) evaluation.	✓	0.18
The 'Base' baseline configuration uses CLIP (clip-vit-large-patch14-336) and SigLIP (siglip-large-patch16-384) without a	×	0.05
The 'SA' baseline is a self-attention-based vision encoder trained using a three-stage pre-training method.	×	0.14
MobileLLM-350M is used as the language model backbone for pre-training in both self-attention and hierarchical cross-att	×	0.13
Optimization is performed using decoupled AdamW with a peak learning rate of 1 $\times$ 10-3 and a cosine decay schedule.	×	0.03
For VLM fine-tuning, the study adopts LLaVA using the Llama-3.2-1B-Instruct model.	×	0.02
For classification tasks, the vision encoder is kept frozen while only the appended classifier head is fine-tuned.	×	0.07
All experiments were conducted on a single RTX 3090 GPU with 24GB VRAM.	×	0.05
A maximum batch size of 256 was used for the early stages of training due to hardware constraints.	×	0.03
HIVE consistently outperforms self-attention baselines across classification and vision-language benchmarks.	×	0.13
HIVE significantly reduces computational overhead compared to self-attention baselines.	×	0.05
CLIP and SigLIP align visual and textual representations through contrastive pre-training.	×	0.06
AIMv2 introduced autoregressive pre-training that processes image patches and text tokens.	×	0.05

References

- http://arxiv.org/abs/2604.00086v1
- http://arxiv.org/abs/2312.08865v1
- http://arxiv.org/abs/2403.10075v2