Scaling Synthetic Data with Engineered Logical Dependencies for Cross-Modal Reasoning Under Distributional Shift

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does scaling synthetic data with engineered logical dependencies during pretraining enhance multimodal model performance on cross-modal reasoning tasks under distributional shift compared to models. 18 claims were extracted from source literature; 2 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: Modularized Zero-shot VQA with Pre-trained Models. Research question: Does scaling synthetic data with engineered logical dependencies during pretraining enhance multimodal model performance on cross-modal reasoning tasks under distributional shift compared to models trained on natural data?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 18 claims extracted; 2 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
The GQA dataset consists of questions requiring multi-step reasoning and various reasoning skills.	×	0.10
Around 94% of the questions in the GQA dataset require multiple reasoning steps.	×	0.12
Questions on the VQAv2 dataset require fewer reasoning steps compared to GQA.	×	0.10
The study reports standard accuracy for the GQA dataset.	×	0.03
The study reports soft accuracy for the VQAv2 dataset due to multiple ground-truth answers.	×	0.04
Experiments were conducted on an NVIDIA Tesla V100 GPU.	×	0.01
The Mod-Zero-VQA method is more effective on the GQA dataset than on datasets with fewer multi-step reasoning questions.	×	0.10
Mod-Zero-VQA clearly surpasses CLIP in performance.	×	0.05
Several methods utilizing large language models achieve better performance than Mod-Zero-VQA but require caption generat	×	0.08
PNP-VQA generates 100 captions per question.	×	0.05
Supervised VQA models give fluctuated performance in different scenes.	×	0.07
The proposed method assigns sub-reasoning tasks to pre-trained models, specifically using MDETR for reference expression	✓	0.16
In the case study examples shown, the proposed method gave correct predictions while QIP and TAC-P answered wrongly.	×	0.02
The Mod-Zero-VQA method decomposes questions into basic reasoning steps and maps them to PTMs without any adaptation.	✓	0.18
The method uses OWL as the object detector.	×	0.04
The method uses MDETR for reference expression localization, including relational and spatial reasoning.	×	0.03
The method uses CLIP as the answer generator for open-ended questions.	×	0.06
Current pre-trained vision-language models have limited capabilities in spatial relation understanding.	×	0.07

References

- http://arxiv.org/abs/2305.17369v2
- http://arxiv.org/abs/2510.03904v2
- http://arxiv.org/abs/2407.04973v1