Modular Zero-Shot VQA Performance Degradation Under Distributional Shift

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of modularized zero-shot VQA systems degrade under distributional shift when evaluated on out-of-domain visual question answering datasets. 12 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: Modularized Zero-shot VQA with Pre-trained Models. Research question: How does the performance of modularized zero-shot VQA systems degrade under distributional shift when evaluated on out-of-domain visual question answering datasets?.

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 4.5/10.

3 Results

13 papers retrieved. 12 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 GQA dataset consists of questions requiring multi-step reasoning and various reasoning skills, with around 94% of th	×	0.13
The VQAv2 dataset requires fewer reasoning steps and is of diverse semantics compared to GQA.	×	0.07
Standard accuracy is reported for the GQA dataset while soft accuracy is reported for the VQAv2 dataset.	×	0.02
Experiments were conducted on NVIDIA Tesla V100 GPU.	×	0.02
The proposed Mod-Zero-VQA method is more effective on the GQA dataset, which contains many multi-step reasoning question	×	0.11
Mod-Zero-VQA surpasses CLIP and several methods that utilize large language models, which often require caption generati	×	0.07
PNP-VQA generates 100 captions per question, which is laborious and may have redundancy.	×	0.03
Supervised VQA models give fluctuated performance in different scenes, demonstrating the robustness of the proposed Mod-	×	0.06
The proposed Mod-Zero-VQA method gives correct predictions for questions requiring multiple reasoning steps, while QIP a	×	0.09
The proposed method decomposes questions into sub-tasks and assigns appropriate sub-tasks to PTMs without any adaptation	✓	0.16
The proposed method uses OWL as the object detector, MDETR for reference expression localization, and CLIP as the answer	×	0.02
The proposed method defines simple heuristics to address the limited capabilities of current pretrained vision-language	×	0.05

References

- http://arxiv.org/abs/2403.14783v1
- http://arxiv.org/abs/2305.17369v2

- http://arxiv.org/abs/2511.00504v2