RankVQA Performance Under Domain-Shift Noise in Multimodal Reasoning Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of varying levels of domain-shift noise on the inference efficiency and accuracy trade-offs of deep learning models evaluated on multimodal reasoning benchmarks. Visual Question Answering (VQA) is a challenging task that requires systems to provide accurate answers to questions based on image content. Current VQA models struggle with complex questions due to limitations in capturing and integrating multimodal information effectively. 18 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: Enhancing Visual Question Answering through Ranking-Based Hybrid Training and Multimodal Fusion. Research question: What is the impact of varying levels of domain-shift noise on the inference efficiency and accuracy trade-offs of deep learning models evaluated on multimodal reasoning benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 18 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 RankVQA model was evaluated using the VQA v2.0 and COCO-QA datasets.	×	0.14
VQA v2.0 contains over 200,000 images and 600,000 questions.	×	0.06
COCO-QA comprises 123,287 images and over 117,000 questions.	×	0.04
The experimental environment used NVIDIA Tesla V100 (32GB) GPUs.	×	0.01
The experimental environment used Intel Xeon E5-2698 v4 CPUs.	×	0.01
The experimental environment had 256GB DDR4 memory.	×	0.01
The experimental environment had 2TB SSD storage.	×	0.01
The experimental environment used Ubuntu 20.04 LTS as the operating system.	×	0.01
The experimental environment used PyTorch 1.10.0 as the deep learning framework.	×	0.09
The experimental environment used CUDA Version 11.2.	×	0.01
The experimental environment used cuDNN Version 8.1.	×	0.01
The experimental environment used Python Version 3.8.10.	×	0.01
Images in the datasets were resized to a uniform size of 224x224 pixels.	×	0.01
Pixel values of the images were normalized to the range of 0 to 1.	×	0.04
The RankVQA model uses Faster R-CNN for visual feature extraction.	×	0.07
The RankVQA model uses a pre-trained BERT model for text feature extraction.	×	0.11
The RankVQA model uses a multi-head self-attention mechanism for multimodal fusion.	×	0.14
The RankVQA model includes a ranking learning module to optimize the relative ranking of answers.	✓	0.18

References

- http://arxiv.org/abs/2212.06370v4
- http://arxiv.org/abs/2105.04026v2
- http://arxiv.org/abs/2408.07303v2