RankVQA Hybrid Training Strategy and Zero-Shot Accuracy on GQA Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the ranking-inspired hybrid training strategy in Rank VQA impact zero-shot accuracy on the GQA benchmark compared to standard cross-entropy baselines. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: How does the ranking-inspired hybrid training strategy in Rank VQA impact zero-shot accuracy on the GQA benchmark compared to standard cross-entropy baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

10 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 3.8/10.

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 RankVQA model was evaluated using the VQA v2.0 and COCO-QA datasets.	×	0.13
VQA v2.0 contains over 200,000 images and 600,000 questions.	×	0.05
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.	×	0.03
The experimental environment used 2TB SSD storage.	×	0.01
The experimental environment used Ubuntu 20.04 LTS operating system.	×	0.01
The experimental environment used PyTorch 1.10.0 deep learning framework.	×	0.06
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.	×	0.00
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.	✓	0.15
The RankVQA model includes a ranking learning module to optimize the relative ranking of answers.	✓	0.18

References

- http://arxiv.org/abs/2403.14783v1
- http://arxiv.org/abs/2305.17369v2
- http://arxiv.org/abs/2408.07303v2