Multimodal Language Model Vision Reasoning Benchmark Performance and Analysis

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Multimodal language model vision reasoning benchmark evaluation analysis. 12 claims were extracted from source literature; 3 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: Second Place Solution of WSDM2023 Toloka Visual Question Answering Challenge. Research question: Multimodal language model vision reasoning benchmark evaluation analysis.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 12 claims extracted; 3 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 Toloka dataset comprises images paired with textual questions, where each entry includes a question-image pair annot	×	0.08
The dataset consists of 45,199 instances distributed across three subsets: 38,990 instances in the training set, 1,705 i	×	0.01
The dataset is structured with several key columns: 'image' contains URLs linking to images hosted on a public content d	×	0.03
Additional metadata includes 'width' and 'height' integers representing the dimensions of each image.	×	0.01
For bounding box annotation, the dataset includes 'left,' 'top,' 'right,' and 'bottom' integers detailing the coordinate	×	0.07
The baseline is defined as the OFA model which directly inputs the competition dataset and reasoning.	×	0.08
The performance of coarse tuning and pseudo answer increased the most because the pseudo answer showed the category of o	×	0.06
On the test public set, our method obtained a seventy-six point five score, and on the test private set, our method obta	×	0.02
Our team achieved a score of 76.342 on the final leaderboard, ranking second.	✓	0.24
The solution of this competition is based on the OFA visual language pre-training model.	✓	0.15
OFA is a large multi-modal pre-training model that can perform various tasks using one decoder module.	×	0.07
The visual grounding task is very similar to the competition, which are both output bounding box coordinates.	✓	0.22

References

- http://arxiv.org/abs/1803.07724v1
- http://arxiv.org/abs/2407.04255v1
- http://arxiv.org/abs/1912.02145v1