Multimodal vs. Text-Only RAG Architectures: Recall and Reasoning on Cross-Domain Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do multimodal RAG architectures (incorporating text and image retrieval) compare to text-only RAG systems in terms of Recall@1000 and reasoning accuracy on cross-domain benchmarks like JURIS-AQA. Retrieval-Augmented Generation (RAG) has been introduced to mitigate hallucinations in Multimodal Large Language Models (MLLMs) by incorporating external knowledge into the generation process, and it has become a widely adopted approach for knowledge-intensive Visual Question. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: QA-Dragon: Query-Aware Dynamic RAG System for Knowledge-Intensive Visual Question Answering. Research question: How do multimodal RAG architectures (incorporating text and image retrieval) compare to text-only RAG systems in terms of Recall@1000 and reasoning accuracy on cross-domain benchmarks like JURIS-AQA and multimodal QA datasets such as VQA or OK-VQA?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

14 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
QA-Dragon achieves an accuracy of 21.31% in single-source tasks.	×	0.13
QA-Dragon achieves an accuracy of 23.22% in multi-source tasks.	×	0.14
Removing the domain router leads to a drop in accuracy from 21.31% to 19.04% in single-source tasks.	×	0.06
Removing the domain router leads to a drop in accuracy from 23.22% to 21.25% in multi-source tasks.	×	0.07
Disabling the tool router results in a performance degradation from 21.31% to 18.32% in single-source tasks.	×	0.04
Eliminating query splitting significantly degrades the performance.	×	0.04
Removing two-stage reranking decreases accuracy from 21.31% to 20.90% in single-source tasks.	×	0.03
Removing two-stage reranking decreases accuracy from 23.22% to 22.14% in multi-source tasks.	×	0.04
Removing two-stage reranking increases latency by nearly 0.3s in the multi-source case.	×	0.03
The framework achieves the best overall balance by dynamically coordinating reasoning and retrieval.	×	0.06
QA-Dragon decomposes the problem into three branches with multiple processes.	×	0.07
The Pre-Answer Module performs domain classification and generates an initial reasoning trace and answer using a domain-	×	0.05
The Search Router inspects the reasoning trace to determine whether additional external evidence is required.	×	0.06
The Tool Router decides whether to invoke an Image Search Agent or a Text Search Agent.	×	0.07

References

- http://arxiv.org/abs/2602.13179v1

- http://arxiv.org/abs/2504.08748v1
- http://arxiv.org/abs/2508.05197v2