# Multimodal Retrieval Latency-Accuracy Trade-offs in Gemini 1.5 Flash for Video and Image Data

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the trade-off between inference latency and multimodal retrieval accuracy in Gemini 1.5 Flash when processing hours of video data versus static image sets. 13 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: REIS: A High-Performance and Energy-Efficient Retrieval System with In-Storage Processing. Research question: What is the trade-off between inference latency and multimodal retrieval accuracy in Gemini 1.5 Flash when processing hours of video data versus static image sets?.

## 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 7.9/10.

## 3 Results

10 papers retrieved. 13 claims extracted; 12 independently verified. Quality review score: 7.9/10.

## 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 |
|---|---|---|
| Large Language Models (LLMs) have knowledge confined to the data they were trained on. | ✓ | 0.19 |
| Retraining Large Language Models incurs significant cost. | × | 0.14 |
| Retrieval-Augmented Generation (RAG) complements static LLM knowledge with an external knowledge repository. | ✓ | 0.20 |
| RAG consists of three stages: indexing, retrieval, and generation. | ✓ | 0.15 |
| The indexing stage of RAG creates a database facilitating similarity search on text embeddings. | ✓ | 0.19 |
| The retrieval stage of RAG searches and retrieves relevant data from the database given a user query. | ✓ | 0.31 |
| The generation stage of RAG uses the user query and retrieved data to generate a response. | ✓ | 0.25 |
| The retrieval stage of RAG is a significant performance bottleneck in inference pipelines. | ✓ | 0.21 |
| In the retrieval stage, a user query is mapped to an embedding vector. | ✓ | 0.23 |
| In the retrieval stage, an Approximate Nearest Neighbor Search (ANNS) algorithm searches for semantically similar embedd | ✓ | 0.30 |
| Large database sizes cause ANNS to incur significant data movement overheads between the host and the storage system. | ✓ | 0.24 |
| Prior works propose In-Storage Processing (ISP) techniques to accelerate ANNS workloads by performing computations insid | ✓ | 0.34 |
| Existing works leveraging ISP for ANNS employ algorithms that are not tailored to ISP systems. | ✓ | 0.22 |

## References


- https://doi.org/10.3390/bdcc9120320
- https://doi.org/10.1145/3695053.3731116
- https://doi.org/10.32604/cmc.2024.052618