# Efficient Multimodal Retrieval-Augmented Generation via Adaptive Context Pruning: Towards Scalable Organizational AI Systems

Harshvardhan Soni, Dr. Parikshit Saikia, Dr. Badal Soni
Department of Computer Science and Engineering, National Institute of Technology Silchar, Assam, India
Email: harshverdhan24_rs@cse.nitc.ac.in
(Supervisor: Dr. Parikshit Saikia; Co-Supervisor: Dr. Badal Soni)

***Abstract*—Retrieval-Augmented Generation (RAG) has emerged as a paradigm to ground large language models (LLMs) with external knowledge, reducing hallucinations and enhancing reliability. While traditional RAG has focused primarily on text, modern organizational data increasingly spans multimodal formats such as documents, images, charts, and tables. This creates challenges in efficiency, latency, and accuracy when fusing heterogeneous evidence into LLM prompts. Moreover, resource constraints in local LLM deployments impose strict *token budgets*, making naïve concatenation of all retrieved content infeasible. This paper proposes Adaptive Multimodal Context Pruning (AMCP): a budget-aware retrieval and selection framework that dynamically allocates context tokens across modalities depending on query intent. We argue that AMCP will enable organizations to deploy scalable multimodal RAG systems with improved efficiency, reduced inference latency, and stronger grounding compared to text-only RAG.**

***Index Terms*—Multimodal Retrieval-Augmented Generation, Adaptive Context Pruning, Token Budgeting, Efficiency Optimization, Organizational AI Systems.**

## I. Introduction

The integration of Retrieval-Augmented Generation (RAG) into enterprise workflows has significantly enhanced generative AI capabilities, particularly by grounding responses with domain-specific knowledge. However, existing RAG systems are predominantly text-centric, ignoring the reality that organizational data is inherently multimodal: annual reports include tables, research papers contain diagrams, and corporate knowledge bases store images, charts, and mixed-format documents. Relying solely on text retrieval risks incomplete grounding and reduced utility.

At the same time, organizations face strict computational budgets when deploying local LLMs. Since inference cost grows non-linearly with context length, unrestricted context fusion results in latency, memory exhaustion, and noise-diluted accuracy. These challenges motivate the need for **multimodal yet budget-aware RAG systems**.

## II. Background and Related Work

Foundational RAG work by Lewis et al. (2020) introduced retrieval-augmented pipelines for open-domain question answering. Extensions such as FiD (Izacard et al., 2021) and Retro (Borgeaud et al., 2022) demonstrated improvements in text fusion efficiency but remain limited to textual modalities. Recent advances in multimodal RAG highlight the potential of integrating text and image retrieval, yet these works often overlook token budgets and latency trade-offs.

Parallel literature on *context pruning* (e.g., Provence, Attention-guided pruning) shows promise in token-efficient RAG but remains underexplored in multimodal settings. Our work builds on these directions by unifying multimodal retrieval, token budgeting, and adaptive pruning.

## III. Budget Constraints in Local LLM Deployment

The concept of a *budget* in RAG refers to the maximum number of tokens allocated to the input context of an LLM. This constraint arises from four major factors:

1) **Latency:** Increased tokens slow inference, particularly on CPU-bound deployments.
2) **Memory Usage:** Transformer attention complexity scales with $O(n^2)$, limiting scalability.
3) **Accuracy Degradation:** Irrelevant tokens dilute the attention signal, reducing grounding effectiveness.
4) **Energy and Compute Cost:** Even local deployments incur electricity and hardware costs for large contexts.

Hence, **budgeted pruning**—the adaptive selection of context under token constraints—is critical for sustainable deployment.

## IV. Adaptive Multimodal Context Pruning (AMCP)

We propose AMCP as a three-stage framework:

1) **Cross-Modal Scoring:** Rank retrieved evidence across text, images, tables, and structured data by relevance, diversity, and complementarity.
2) **Adaptive Budget Allocation:** Dynamically allocate token budget to modalities based on query intent. For example, image tokens dominate in "show architecture" queries, while text dominates in "define policy" queries.
3) **Structured Fusion:** Concatenate pruned multimodal evidence in a structured prompt format with modality markers to guide the LLM.

## V. Need in Organizational Setup

Modern enterprises require AI systems capable of handling knowledge bases that span multimodal content. Text-only RAG pipelines fail to extract insights from diagrams, dashboards, and tables, leading to incomplete decision support. Further, efficiency bottlenecks limit scalability across departments. AMCP offers:

- **Reduced Latency:** Ensuring timely responses in enterprise chatbots and decision systems.
- **Improved Reliability:** Stronger grounding by integrating multimodal evidence.
- **Cost Efficiency:** Lower computational overhead on local hardware.

We argue that multimodal, budget-aware RAG will soon replace text-only RAG in organizational contexts.

## VI. Research Direction and Expected Contributions

Our ongoing research will focus on:

1) Designing adaptive pruning algorithms that optimize accuracy-latency trade-offs under token budgets.
2) Extending context pruning to multimodal retrieval pipelines.
3) Evaluating performance on organizationally relevant multimodal corpora.
4) Demonstrating Pareto-optimal trade-offs between efficiency and reliability.

## VII. Conclusion

This paper outlines the motivation and framework for Adaptive Multimodal Context Pruning in RAG systems. By explicitly addressing token budget constraints and extending retrieval beyond text, AMCP provides a pathway for scalable and efficient AI deployments in organizations. We anticipate this paradigm to supersede core text-based RAG in enterprise environments within the near future.

## References


[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

[2] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain QA," arXiv, 2021.

[3] S. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," arXiv, 2022.

[4] Y. Wang et al., "Provence: Efficient and Robust Context Pruning for RAG," OpenReview, 2025.

[5] Z. Zhang et al., "Multi-RAG: A Multimodal Retrieval-Augmented Generation System," arXiv, 2025.