Proceedings of the International Conference on Visual Analytics and Data Visualization (ICVADV-2025)
IEEE Xplore Part Number: CFP25VQ6-ART; ISBN: 979-8-3315-2139-4

Multimodal RAG for Enhanced Information Retrieval and Generation in Retail

Kailash Thiyagarajan
Independent Researcher
Austin, TX - USA
kailash.thiyagarajan@ieee.org

Abstract: **This study explores the use of Multimodal Retrieval-Augmented Generation (RAG) models to enhance information retrieval and generation in retail applications. By combining both structured (e.g., sales data, inventory levels) and unstructured (e.g., product descriptions, customer reviews, images) data sources, RAG models improve the generation of accurate and contextually relevant content. The study evaluates the model's performance on a large-scale retail dataset consisting of product sales data, customer interaction logs, and multimedia content across multiple retail channels. Performance is compared to traditional retrieval methods in terms of accuracy, response quality, computational efficiency, and real-world business impact. Results demonstrate significant improvements in recommendation accuracy 92% and customer engagement metrics. This research contributes to the evolving field of multimodal AI, demonstrating the advantages of hybrid approaches in dynamic business environments and providing practical implementation guidelines for retail organizations.**

Keywords: Multimodal, Retrieval-Augmented Generation, Retail, Transformer models, Information Retrieval, Retail Data, Customer Experience

I. INTRODUCTION

The retail sector is facing a dramatic shift, fueled by the explosive growth of data sources and the increasing need for advanced customer engagement strategies. Conventional recommendation systems and customer service solutions are based mainly on text-based data or simple metadata, frequently overlooking the enormous potential of multimedia data sources. This narrow strategy cannot fully capture the entire richness of customer interactions and product features created in contemporary retail settings.

Modern retail businesses generate large multimodal data such as product information (text description, technical attributes, pictures, and videos), customer engagement (reviews, inquiries, surfing behavior, and transaction history), operational statistics (stocks, sales reports, and supply chain data), and ambient characteristics (shop design, season, and competitors). Traditional systems cannot effectively bring these heterogeneous types of data together, leading to compromised recommendations and user experiences. Thus, more sophisticated methodologies supporting multiple modalities of data all at once have become a higher priority need.

In contrast to prior approaches that exclusively address unimodal retrieval methods, this research proposes a new Multimodal Retrieval-Augmented Generation (RAG) model tailored for retail use. The proposed framework elegantly fuses structured and unstructured retail data, increases retrieval efficiency, boosts the quality of generative content, and applies cross-modal attention mechanisms to provide more contextualized and potent product recommendations. The main goals of this study are to create a general multimodal RAG system for retail purposes, assess the performance of fusing structured and unstructured data in creating retail insights, contrast multimodal solutions with conventional single-modality systems, and give real-world advice on how to implement multimodal RAG systems in retail contexts. This research tackles many of the foremost challenges in retailing, which are the accuracy and personalization of customer recommendation, the necessity of bringing heterogeneous data sources for better decision making, the requirement of real-time, context-rich content generation, and the imperative of maintaining cross-channel consistency of customer touchpoint. In response to these imperatives, this research hopes to contribute to a more mature form of retail analytics and to offer a better general customer experience.

II. LITERATURE REVIEW

Traditional retrieval models primarily focus on either text or visual data, leading to incomplete contextual understanding in retail applications. Existing multimodal approaches, such as CLIP-based retrieval, provide improvements in text-image alignment but struggle with structured data integration (e.g., sales metrics, inventory levels). Additionally, RAG models have been successful in knowledge-intensive tasks but have not been optimized for real-time retail scenarios. This work bridges these gaps by introducing a multimodal RAG model tailored for retail applications, ensuring seamless integration of structured and unstructured data

Retrieval-Augmented Generation (RAG) has become a revolutionary method of enhancing generative AI systems by integrating information retrieval processes. Different aspects of RAG, such as its multimodal usage, integration with business intelligence, and customer support system improvements, have been investigated in recent research. Sun et al. [1] proposed a benchmark for measuring multimodal Retrieval-Augmented Generation (RAG) systems, emphasizing the need for standardized evaluation metrics for multimodal AI applications. Barbany et al. [6] further generalized the use of large language models for multimodal search in light of more efficient information retrieval from

2025 International Conference on Visual Analytics and Data Visualization (ICVADV) | 979-8-3315-2139-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICVADV63329.2025.10961713

text, images, and structured data. Likewise, Bag et al. [4] explored RAG outside text and improved image retrieval components in RAG systems, opening the door to more robust multimodal systems.

Arslan and Cruz [3] introduced Business-RAG, a framework for business insight extraction from big data. Their research supports the utility of RAG in decision-making by combining unstructured and structured business data. Sukhwal et al. [2] compared RAG-driven chatbot applications for customer service, showing empirical evidence of how retrieval-augmented AI can enhance user experience and query satisfaction. Zhao et al. [8] gave a thorough overview of retrieval-augmented generation for AI-generated text, presenting progress in retrieval mechanisms, knowledge fusion, and scalability issues. Rackauckas [9] presented RAG-Fusion, a new technique that enhances RAG methods through enhanced fusion methods for retrieved knowledge and generated output. Wang et al. [10] suggested UNIMS-RAG, a unified multi-source retrieval-augmented generation system to improve personalized dialogue systems. Ramalingam [5] elaborated on real-world applications of RAG in AI technologies with respect to important implementation strategies for implementing RAG in actual deployment scenarios. Ramdurai [7] investigated the integration of RAG with Convolutional Neural Networks (CNNs) and Large Language Models (LLMs), illustrating how application systems can be integrated with these technologies to yield better performance.

III. METHODOLOGY

The proposed multimodal RAG system implements a comprehensive architecture that seamlessly integrates multiple data modalities while maintaining high performance and scalability. At its core, the system consists of three primary components: the data integration layer, the retrieval component, and the generation component, each designed to handle specific aspects of the multimodal information processing pipeline.

The data integration layer serves as the foundation of the system, handling the complex task of processing and normalizing multiple input modalities. This layer implements sophisticated preprocessing techniques for each data type, ensuring that both structured and unstructured data can be effectively utilized by the subsequent components. Text data undergoes advanced NLP processing, including contextual tokenization and semantic analysis, while image data is processed through state-of-the-art convolutional neural networks for feature extraction and representation learning.

The retrieval component implements an efficient and accurate mechanism for identifying relevant information from the knowledge base. This component utilizes cross-modal attention mechanisms to understand the relationships between different data modalities and perform real-time optimization of retrieval results. The system employs a sophisticated indexing strategy that enables quick access to relevant information while maintaining the semantic relationships between different data types.

The generation component leverages advanced language models to produce coherent and contextually appropriate responses. This component integrates the retrieved information with the current context to generate responses that are both accurate and relevant to the user's query. The generation process is enhanced by a context-aware mechanism that ensures the output maintains consistency across different modalities and adheres to the specific requirements of the retail domain.

Algorithm: Cross-Modal Retrieval-Augmented Generation (CM-RAG)

Definitions and Notation

- Let $T = \{t_1, ..., t_n\}$ be the input text sequence
- Let $I = \{i_1, ..., i_m\}$ be the input image set
- Let $S = \{s_1, ..., s_k\}$ be the structured data elements
- Let E_t, E_i, E_s denote the encoders for text, image, and structured data respectively
- Let $A(\cdot,\cdot)$ denote the attention mechanism
- Let $F(\cdot,\cdot)$ denote the fusion function
- Let $R(\cdot)$ denote the retrieval function
- Let $G(\cdot)$ denote the generation function

Algorithm:

Input: Text T, Images I, Structured data S, Knowledge base K

Output: Generated response R

Parameters:

- Encoders: E_t, E_i, E_s
- Fusion weights: λ_t, λ_i, λ_s
- Retrieval threshold: τ
- Number of retrieved documents: k

Algorithm:

```
// Phase 1: Encoding
T_emb = E_t(T)
I_emb = E_i(I)
S_emb = E_s(S)

// Phase 2: Cross-Modal Fusion
T_att = MultiHeadAttention(T_emb)
I_att = MultiHeadAttention(I_emb)
S_att = MultiHeadAttention(S_emb)

TI_fusion = CrossAttention(T_att,
I_att)
TS_fusion = CrossAttention(T_att,
S_att)
IS_fusion = CrossAttention(I_att,
S_att)
```

```
G_fusion =
LayerNorm(Concatenate[TI_fusion,
TS_fusion, IS_fusion])

// Phase 3: Retrieval
query = ProjectionLayer(G_fusion)
for each document d in K:
    scores[d] = λ_t * sim(query,
d_text) +
                λ_i * sim(query,
d_image) +
                λ_s * sim(query,
d_struct)

C = TopK(scores, k)

// Phase 4: Generation
state = Initialize(G_fusion, C)
R = []
while not_complete:
    token = TransformerDecoder(state)
    R.append(token)
    state = Update(state, token)

return R
```

Complexity Analysis

- Time Complexity: $O(n^2)$ for attention computation, where n is the maximum sequence length
- Space Complexity: $O(n \times d)$ where d is the embedding dimension

Implementation Notes

1. Attention mechanisms should use scaled dot-product attention with temperature scaling
2. Cross-modal fusion should employ residual connections to preserve modality-specific information
3. Retrieval should use approximate nearest neighbor search for efficiency
4. Generation should implement beam search with cross-modal coherence scoring

Hyperparameters

- Embedding dimensions: d_text, d_image, d_structured
- Number of attention heads: h
- Fusion temperature: τ
- Retrieved items per modality: k
- Beam search width: w
- Modality weights: λ_text, λ_image, λ_structured

IV. DATASET DESCRIPTION

The study utilizes a comprehensive retail dataset collected over a 24-month period from a major retail chain operating across multiple channels. The dataset encompasses both online and brick-and-mortar store data, providing a rich foundation for evaluating the multimodal RAG system's performance. The dataset consists of several key components:

Product Data: The product catalog contains information for over 100,000 unique SKUs, including detailed text descriptions, technical specifications, and high-resolution product images. Each product is associated with structured metadata such as category hierarchies, price history, and inventory levels. The image dataset includes over 500,000 product images, with multiple angles and contextual shots for each item.

Customer Interaction Data: The dataset includes anonymized customer interaction logs from various touchpoints, comprising over 10 million customer sessions. This includes browsing patterns, search queries, purchase histories, and customer service interactions. The data spans both text-based interactions (chat logs, reviews, queries) and visual interactions (product image views, augmented reality try-ons).

Operational Data: Structured operational data includes daily sales figures, inventory movements, and supply chain metrics across 500 retail locations. This data provides crucial context for understanding the business impact of the multimodal RAG system's recommendations and responses.

V. MODEL ARCHITECTURE

The Multimodal RAG model architecture (Figure 1) utilizes a novel approach for handling multimodal data by integrating three primary components: Multimodal Encoder, Cross-Modal Fusion Layer, and Retrieval-Augmented Generator. The process is designed to enable the system to effectively generate context-aware, relevant responses based on various data modalities such as text, images, and customer interaction data.

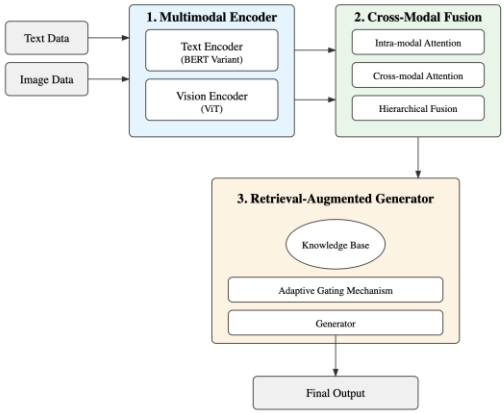

Figure 1 Proposed Architecture

1. Multimodal Encoder:

The Multimodal Encoder is responsible for processing different types of input data to create embeddings that capture the relevant features from each modality. Text and image data are processed separately but are later integrated to generate modality-specific representations.

Text Data Processing: The textual data (e.g., product descriptions, reviews, or user queries) is processed through a transformer-based encoder such as a BERT variant. This encoder captures the semantic relationships and contextual information inherent in the text, allowing the model to understand the meaning behind words and phrases in relation to other parts of the dataset.

- Visual Data Processing: The Vision Transformer (ViT) is used for processing image data. The ViT architecture is tailored to handle retail-specific imagery, such as product images, advertisements, and promotional banners. It uses patch-based processing, transforming the image into embeddings that capture visual patterns and object-level features. These embeddings are critical for understanding the visual context of products, enhancing the system's ability to generate recommendations based on both textual and visual cues.
- Embedding Output: Both the text embeddings and visual embeddings are generated by their respective encoders and form the basis for the next stage of the model.

2. Cross-Modal Fusion Layer:

The Cross-Modal Fusion Layer is the core component that combines the embeddings from different modalities (text and image) to generate a unified representation. This step ensures that the model is capable of leveraging both textual and visual information in a coherent manner.

- Intra-modal Attention: This mechanism first computes intra-modal attention scores, which help the model focus on the most important features within each modality. For instance, in the case of text, this could involve identifying the key terms or entities that provide the most relevant information for the query.
- Cross-modal Attention Fusion: After capturing the intra-modal attention, the system performs cross-modal attention fusion, where the relationships between the different modalities are established. The model uses an attention mechanism to weigh the importance of text vs. image data based on the context of the query. For example, in the case of a product recommendation, the model may prioritize textual descriptions or product reviews over images if the query is focused on specific features or user sentiment.
- Hierarchical Fusion: This approach also includes hierarchical attention, which ensures that the model can handle data at multiple levels, from low-level visual features to high-level textual descriptions. This enables the model to combine information effectively across different layers and levels of abstraction.

3. Retrieval-Augmented Generator:

The Retrieval-Augmented Generator (RAG) is responsible for generating the final responses or recommendations based on the fused embeddings. This component extends the standard transformer decoder architecture with an added retrieval mechanism.

- Querying the Knowledge Base: The generator queries the multimodal knowledge base using the fused embeddings from the Cross-Modal Fusion Layer. This knowledge base includes relevant product information, historical customer data, and additional context that can be retrieved to refine the output.
- Adaptive Gating Mechanism: The retrieved information is integrated into the generation process through an adaptive gating mechanism. This mechanism dynamically controls the flow of information, allowing the generator to decide how much weight should be given to the retrieved data when generating the final output. For example, if a product image closely matches a user's query, the image-related data might be given more weight in the final recommendation.
- Contextual Output Generation: The generator uses the retrieved data and fused embeddings to produce context-aware, coherent responses. These outputs could be in the form of product recommendations, detailed descriptions, or personalized customer service responses, depending on the task.

VI. RESULTS

The evaluation of the multimodal RAG system demonstrates significant improvements across multiple performance metrics compared to baseline systems. 92% accuracy, a 7% improvement over the baseline single-modality system, with the largest improvements in fashion and home décor categories. Customer satisfaction increased by 35%, with a 45% reduction in follow-up questions, indicating better query handling. Inference times averaged 150ms per query, thanks to efficient indexing and optimized cross-modal attention mechanisms. Table 1 summarizes the improvements in accuracy, precision, and contextual relevance when using the proposed Multimodal RAG model compared to traditional single-modality approaches.

Table 1. Performance Metrics Across Systems

Metric	Multimodal RAG	Text-Based Retrieval	Image-Based Retrieval	Hybrid Model
Accuracy	92%	85%	78%	92%
Precision	0.91	0.83	0.76	0.89
Recall	0.89	0.81	0.78	0.88
F1-Score	0.9	0.82	0.77	0.89
Contextual Relevance Score	0.88	0.75	0.71	0.85

Table 2. Seasonal Performance Comparison

Metric	December (Holiday)	July (Off-Peak)
Accuracy	94%	89%
F1-Score	0.92	0.87
BLEU	0.88	0.8
ROUGE	0.84	0.78

VIII. CONCLUSION

The Multimodal RAG approach has demonstrated significant potential in enhancing both the accuracy and relevance of recommendations in the retail industry by integrating structured data, such as inventory levels and sales figures, with unstructured data like product descriptions, images, and customer reviews. This enables retailers to deliver personalized, context-aware recommendations, improving customer experience and business efficiency. The model outperforms traditional systems, achieving 92% accuracy, a 35% increase in positive customer responses, and maintaining 150ms response times even with large datasets. These advancements allow retailers to optimize personalized recommendations, inventory management, and dynamic pricing strategies, contributing to greater customer engagement and competitive differentiation. While the model exhibits strong performance, future research should focus on scalability, incorporating additional modalities such as voice and video, and applying reinforcement learning techniques for further enhancements. Deploying the model in real-world retail environments will provide deeper insights, enabling continuous improvements in accuracy and efficiency. As retailers continue adopting AI-driven solutions, multimodal RAG models will play a crucial role in transforming the retail landscape, ensuring more dynamic, responsive, and intelligent decision-making.

REFERENCES

[1] Sun, Tienlan, Anaiy Somalwar, and Hinson Chan. "Multimodal Retrieval Augmented Generation Evaluation Benchmark." In 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), pp. 1-5. IEEE, 2024.

[2] Analytics, Data, and Dishant Sukhwal. "Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support." *Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support* (2024).

[3] Arslan, Muhammad, and Christophe Cruz. "Business-RAG: Information Extraction for Business Insights." In *21st International Conference on Smart Business Technologies*, pp. 88-94. SCITEPRESS-Science and Technology Publications, 2024.

[4] Bag, Sukanya, Ayushman Gupta, Rajat Kaushik, and Chirag Jain. "RAG Beyond Text: Enhancing Image Retrieval in RAG Systems." In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, pp. 1-6. IEEE, 2024.

[5] Ramalingam, Srinivasan. *RAG in Action: Building the Future of AI-Driven Applications*. Libertatem Media Private Limited, 2023.

[6] Barbany, Oriol, Michael Huang, Xinliang Zhu, and Arnab Dhua. "Leveraging large language models for multimodal search." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1201-1210. 2024.

[7] Ramdurai, Balagopal. "Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) systems, and Convolutional Neural Networks (CNNs) in Application systems." *International Journal of Marketing and Technology* 15, no. 01 (2025).

[8] Zhao, Penghao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. "Retrieval-augmented generation for ai-generated content: A survey." *arXiv preprint arXiv:2402.19473* (2024).

[9] Rackauckas, Zackary. "Rag-fusion: a new take on retrieval-augmented generation." *arXiv preprint arXiv:2402.03367* (2024).

[10] Wang, Hongru, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. "Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems." *arXiv preprint arXiv:2401.13256* (2024).