# Multimodal Search Exploration for E-Commerce

Gajeendran Ratnalingam
Department of Computing
*Informatics Institute of Technology*
Colombo, Sri Lanka
gajanhcc007@gmail.com

Mithushan Jalangan
Department of Computing
*Informatics Institute of Technology*
Colombo, Sri Lanka
mithushan.j@iit.ac.lk

***Keywords— product search(PS), vision-language model(VLM), contextual embeddings(CE), information retrieval(IR)***

## I. INTRODUCTION

The landscape of electronic commerce search has been changing drastically in a constant manner with an exponentially growing number of products, user-generated content and complex consumer behaviour patterns [2]. These aspects have made e-commerce search a challenging problem in order to provide accurate, relevant, and personalized search results.

The challenges in e-commerce search are complex where the misalignment between visual and textual modalities in multimodal search systems may lead to poor search experiences, especially when users submit detailed, ambiguous or more natural queries [2]. This research addresses these issues by introducing an integrated approach that fuses text and image data within a unified space utilizing the ColPali mechanismwith late interaction for seamless multimodal alignment.

The existing search systems are moving towards the conversational search or chatbots which uses the natural queries like “Black Jacket” or “Black Jacket with fleets” and through this multimodal system proposed, the specific and compelling use case of using VLM comes into the effect where a user can upload an image of a specific piece of clothing while asking a query as “Find me some similar jackets in black with fleets and a waterproof lining”. Hence the traditional systems will still struggle to interpret such queries holistically.

The Vision Language Model has the advantage of leveraging their ability to jointly analyze the uploaded image and textual query where they can identify critical visual features like the jacket’s style, texture, colour etc. while integrating this understanding the nuanced context and intent behind the user’s requirements like “black”, “waterproof” etc. When they are combined with the RAG, the system attains the capability to retrieve the relevant product data from external sources to ensure a more comprehensive and accurate search experience. In this paper, we initially review the existing e-commerce search systems and then discuss the proposed novel architecture which resolves the stated gap in the domain of e-commerce search with its search and information retrieval components

## II. LITERATURE REVIEW

E-commerce search systems have evolved rapidly starting from the basic search with boolean keyword matching [6] followed by the TF-IDF-based ranking approach to the faceted filtering to refine search results [1]. With that the introduction to fuzzy logic improved the handling of vague or imprecise queries and then the Learning-to-Rank algorithms infused machine learning to optimize the search rankings based on the user behaviour [5]. Later, the hybrid search systems combined the lexical and semantic techniques to achieve more precise information retrieval and gradually the shift towards the multimodal search systems began to appear which promised with good performance in information retrieval with models like CLIP, BLIP etc.

Leading e-commerce platforms like Amazon, Walmart leverages advanced retrieval techniques including deep learning, semantic search, and vision-language models in order to enhance search functionalities and personalization [4]. Recent advancements in the domain of vision-language models have not been utilized into product retrieval but they enabled open-world retrieval and contextual understanding of unseen data by making e-commerce search more dynamic [4] and user-centric. .

## III. THE PROPOSED APPROACH

The proposed solution leverages vision-language models to improve the image-text alignment which utilizes enriched product images and its descriptions capturing the fine-grained semantic details into a unified representation space ensuring seamless integration of different modalities while maintaining the semantic consistency using the ColPali model which is prolific in producing high-quality COLBERT-style multi-vector embeddings from images of document pages outperforming modern document retrieval pipelines by indexing purely on visual features allowing for subsequent fast query matching with the late interaction mechanism integrated within it [3].

The implementation pipeline of the proposed approach is exhibited in Figure 1 which begins with the dataset collection where the high-quality product catalogs containing detailed textual descriptions and images are served as a primary dataset. And the images and text undergo pre-processing is done to maintain the consistency while conserving the context.

https://doi.org/10.31705/ADScAI.2025.38

Then the key part of the system is feature engineering and embedding generation processes where the image embeddings are created using the encoder of the ColPali model extracting higher-dimensional visual features. Similarly textual descriptions are processed using the encoder to generate embeddings that should align with the extracted visual features and GPT model is used to enhance the image descriptions by summarizing key attributes.

The contextual and multimodal embeddings are projected into a lower dimensional space like 128 dimensions to reduce the storage constraints while preserving rich information. And when a user submits a query, it is processed through the same embedding generation pipeline while ensuring alignment with the stored product embeddings. Then the similarity search system will be employed for efficient similarity searches while ranking top-k results based on the relevance. A late interaction mechanism is employed here for the multi-modal fusion at this shared and unified embedding spaces which refines this ranking process while ensuring contextually relevant recommendations by providing a common embedding space for the multi-vector representation.

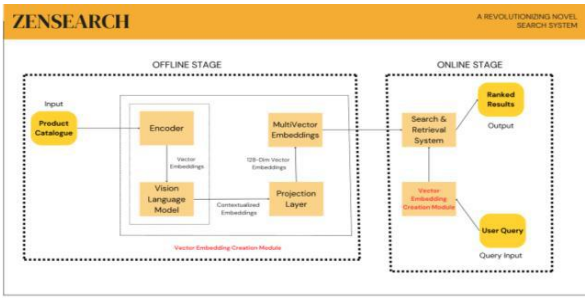


Figure 1. Proposed Model Pipeline

The architecture is flexible for real-time applications and to be integrated with other e-commerce functionalities where it can be extended to personalized advertising with the user's improved search and purchase history to refine the recommendations dynamically. And analysing search patterns and preferences for tailored shopping experiences are helpful with future integrations like reinforcement learning by improving performance on user feedback. And fine-tuning of the ColPali model and its usage in the approach enables the approach to specialize in visual-textual alignment for e-commerce applications, improving its ability to interpret and retrieve products effectively.

## IV. TESTING AND BENCHMARKING

The below Table 1 is the results metrics that was retrieved from testing the base ColPali model with the finetuned ColPali model that is proposed to be used in the approach. And the below Table 2 is the ViDoRe benchmarking metrics results of multiple systems obtained from the ColPali paper [3]. The datasets used below to benchmark the system are part of the standard ViDoRe benchmarking.

TABLE I. RESULTS OF BASE AND FINETUNED MODELS

| Metric | vidore/colpali-v1.3 | Base: vidore/colpali-v1.2 | Finetuned: gajanhcc/finetune_colpa li_v1_2- |
|---|---|---|---|
| Accuracy | 78.5% | 45.5% | **89%** |
| NDCG | 0.79 | 0.46 | **0.89** |
| MAP | 0.79 | 0.46 | **0.89** |
| Precision | 0.70 | 0.36 | **0.84** |
| Recall | 0.79 | 0.46 | **0.89** |
| F1-Score | 0.72 | 0.38 | **0.86** |
| MRR | 0.8728 | 0.5907 | **0.9316** |

TABLE II. VIDORE BENCHMARKING-COLPALI VS FINETUNED

| Recall@1 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Arxi vQ | Doc Q | Info Q | Tab F | TAT q | Shift | Health |
| ColPali | **72.4** | 45.6 | 74.6 | 75.4 | **53.1** | 55.0 | **88.0** |
| Finetuned | 70.4 | **47.0** | **75.5** | **82.1** | 52.6 | **63.0** | **88.0** |
| nDCG@5 | | | | | | | |
| | Arxi vQ | Doc | Info | TabF | TAT q | Shift | Health |
| ColPali | **79.1** | 54.4 | 81.8 | 83.9 | 65.8 | 73.0 | **94.4** |
| Finetuned | 77.2 | **54.6** | **82.0** | **88.3** | **66.4** | **79.0** | 94.3 |


## ACKNOWLEDGMENT

The authors would like to thank all the wellwishers for their overall support with the valuable insights and also extend the gratitude to the open-source communities whose contributions greatly enriched this work and especially to the ColPali and the HuggingFace teams.


## REFERENCES


[1] J. Koren, Y. Zhang, and X. Liu, "Personalized interactive faceted search," in Proc. 17th Int. Conf. World Wide Web, Beijing, China, 2008, pp. 477-486.

[2] M. A. Ghossein, C.-W. Chen, and J. Tang, "Shopping queries image dataset (SQID): An image-enriched ESCI dataset for exploring multimodal learning in product search," arXiv:2405.15190, 2024.

[3] M. Faysse et al. ColPali: Efficient Document Retrieval with Vision Language Models. 2024, doi: 10.48550/ARXIV.2407.01449

[4] N. Maio and B. Re, "How Amazon's e-commerce works," Zenodo, 2020. doi: 10.5281/ZENODO.3894408.

[5] P. G. Anick and S. Vaithyanathan, "Exploiting clustering and phrases for context-based information retrieval," in Proc. 20th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '97), Philadelphia, PA, USA, 1997, pp. 314-323.

[6] S. O. Kimbrough and S. A. Moore, "On obligation, time, and defeasibility in systems for electronic commerce," in Proc. 26th Hawaii Int. Conf. System Sciences, Wailea, HI, USA, 1993, pp. 493-502