

RESEARCH ARTICLE

OPEN ACCESS

A HYBRID CROSS-MODAL DEEP LEARNING AND CLIP-BASED FRAMEWORK FOR PERSONALIZED FASHION PRODUCT RECOMMENDATIONS

G Kalyan Chakravarthi¹, Kavuri Bhavya Sri^{2*}, Mandhadapu Pallavi³, D Ganesh⁴, M Ranjith Kumar⁵

^{1, 2, 3, 4, 5}Department of Computer Science and Engineering, Gayatri Vidya Parishad College for Degree and PG Courses(A), Visakhapatnam, India.

¹<https://orcid.org/0000-0001-8545-3444>, ^{2*}<https://orcid.org/0009-0007-9120-0639>, ³<https://orcid.org/0009-0002-5612-2501>

⁴<https://orcid.org/0009-0001-4433-1931>, ⁵<https://orcid.org/0009-0005-4914-1431>

E-mail: gkalyan@gvpcdpge.edu.in, *5221411074@gvpcdpge.edu.in, 5221411092@gvpcdpge.edu.in, 5221411180@gvpcdpge.edu.in, 5221411097@gvpcdpge.edu.in

ARTICLE INFO

Article History

Received on: January 03, 2026
Reviewed on: February 03, 2026
Accepted on: April 10, 2026
Published on: June 30, 2026

Keywords:

Cross-Modal Retrieval,
CLIP & FAISS,
Deep Learning,
Product Recommendation,
Visual Sentiment Analysis.

ABSTRACT

E-commerce systems are challenged with a lot of difficulty to offer user-friendly product search and recommendation experiences that bridge the visual and textual gap modalities. The paper introduces a new cross-modal recommendation system that uses CLIP (Contrastive Language-Image Pre-training) to conduct multi-modal search of products. Our system allows users to query based on natural language queries, image, or both queries together through the intelligent fusion systems. To improve relevance of recommendations, we include context-considered ranking using visual sentiment analysis, occasion-mood matching and time sensibility. The system utilizes efficient similarity search using FAISS with a variety of fusion options (weighted averaging, concatenation, element-wise multiplication) and high-tech reranking mechanisms (cross-attention). The experimental results show the better results of our approach; it gives a much better score of 92.3 percent precision at 10 on text queries, 88.7 percent on image queries, and 94.1 percent on single hybrid queries. Occasion-aware ranking enhances the user satisfaction by 34, and visual sentiment analysis brings out more engagement by 28. Our RESTful API platform can support real-time inference with an average query latency of 127ms, which is appropriate to be deployed at production in large-scale e-commerce systems.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

A high level of development of e-commerce has shifted the consumer shopping behaviour, and the global online retail sales have already reached more than \$5.7 trillion of all online retail sales in 2023. Nevertheless, the conventional product search systems are largely based on conventional text-based matching of a particular product with a keyword, and it does not reflect the fine visual taste and contextual needs of the contemporary consumer. Customers cannot easily define their product preferences using words only especially when it comes to fashion and home decor and lifestyle products where the predominant factor is visual appeal and imagery [1]. The latest developments in the field of deep learning have facilitated the concept of cross-modal learning where models tend to perceive and relate information between various modalities like text and images.

More recently [2], CLIP (Contrastive Language-Image Pre-training) [3] has become groundbreaking in the field, that is, it is trained on 400 million [4] pairs of images and text and learns a shared embedding space in which closely related images and text get placed near each other [5], [6]. This high rate of online shop digitization has not only boosted the quantity and variety of products found online, but has also elevated user demands regarding the ease of goods shopping and customization and visual stimulation. Although early online shopping systems promoted a shallow search and poor collaborative filtering based on key-word search, these models cannot easily address the depth and breadth of user desire, particularly in areas such as fashion where style texture, colour and situation are important to the user.

To illustrate, when a user enters a query such as a vintage floral dress to a summer wedding, he or she can be in possession of a highly detailed visual and contextual image that cannot be quite represented by the use of keywords only. Conventional search engines are also restricted by the fact that they cannot close the so called assembling semantic gap [7] that exists between the verbally described products and the visually presented one [8]. The resulting consequences of this gap are lost opportunities to know, inappropriate recommendation relevance, and user frustration. This has increased the requirement of systems that can easily combine and make inferences upon both visual and textual data. This problem is overcome with the development of cross-modal learning, which allows the simultaneous interpretation and relation of data between the two modalities of data to be shared between deep learning models. A major breakthrough among them is CLIP (Contrastive Language-Image Pre-training).

Created by OpenAI, CLIP is trained on 400 million image-text pairs gathered on the internet, and thus it wakes up to learn a massively generalized and strong shared embedding space. Within this space, images and described textual representations are inscribed near each other allowing the model to accomplish tasks relating to zero-shot image classification, cross-modal retrieval, and visual search with impressive precision [9]. When dealing with product recommendation, the capabilities I have mentioned above allow CLIP to respond to the needs of visual and textual semantics in a way, that users can search and find products in a natural language or through a picture, or a combination of the two and get results which are contextually and visually relevant. It is particularly transformative in fashion, home decor and lifestyle products where it is more about the subjective preferences and visual clues. Our system can also overcome the semantic translucency barrier with the help of CLIP and, consequently, better discoveries and user experience by providing more engaging and rewarding experience to users.

II. EXISTING SYSTEM

The existing product recommendation systems more so within the e-commerce and fashion sectors are hampered by a number of basic constraints that incapacitate the potential of providing indeed personalized and pertinent results. The majority of the solutions available consider text and image data as separate modalities, where product descriptions and visual information are processed separately. Such a division creates a lasting "semantic gap" so that the system cannot indeed correctly match descriptive phrases, say, elegant red evening gown and the visual characteristics in product images. Consequently, the recommendations do not easily reflect the subtle likes of the users who need textual and graphic messages in their buying choices. Also, such systems generally do not have powerful means of factoring in the user context, say the situation occasion, mood and season. As an illustration, a searcher who typed words like casual summer shoes and a beach vacation might not get customized results to give him or her relevant information to find the shoe type he or she wants, thus when it comes to a beach vacation, the results will not offer much tailored information which would result in an unsatisfaction to the searcher.

The sentiment analysis, when such exists, typically only applies to a single, coarse-level classification (positive, negative, or neutral), and cannot capture finer-grained, feature-level information about user reviews, e.g. comfort, fit, or style, which are key to fashion and lifestyle products. Another significant challenge is scalability. With product catalogues containing tens and hundreds of thousands of items, many conventional systems cannot provide the search performance in an efficient and accurate way. This may lead to slack response time and the lack of product awareness. Besides, there is a lack of available reranking methods that take into account various tools of relevance, like sentiment, context, and diversity, which reduces their capacity in order to deliver the most suitable products to the individual. A combination of all these restrictions is less captivating, less precise, and less credible user experiences. In visual based fields such as fashion, where the personal style and the context of the user are the most important factors, the impossibility to cross the gap between modalities and the adjustment to the situation of the user is a challenge to the successful product recommendation.

III. PROPOSED SYSTEM AND ARCHITECTURE

In order to overcome the drawbacks of the current product recommendation systems, we present a new system of cross modal recommendation [10], [11] that will combine visual, textual, and contextual cues to produce very personalized and context-dependent product recommendations. The main component of our system is the CLIP (Contrastive Language-Image Pre-training) model, which allows the creation of a coherent matching of product images and textual description in the same semantic embedding space [2-4]. This alignment fills in that semantic gap between description and appearance of products and makes it possible to retrieve and match products across modalities.

III.1 SYSTEM ARCHITECTURE

We have an architecture that is scalable, flexible and user-friendly. It has five key elements, a multi-modal encoder that runs on CLIP, an efficient embedding storage system, a similarity search index powered by FAISS, an advanced query processor with multiple fusion strategies and a context-aware ranker that uses sentiment, occasion and diversity indicators. The system permits multiple search options of text only, image only and hybrid searches along with an advanced workflow-based searches that adapt to the context of the user e.g. occasion and mood search. One of the innovations of our strategy is the combination of fine-grained visual sentiment analysis and occasion-mood matching which allows the system to propose the products that are not only correlated with specific query by the user but also compatible with the emotional and situational background of the user.

The cross-attention mechanisms are also used to refine the ranking of results further by modelling multifaceted interaction between user queries and product features. Fast API and PyTorch are used to implement the backend as asynchronous processing, and high-performance processing, whereas modern web technologies are used in the frontend to provide an intuitive user experience. The whole system is designed to scale to large-scale quickly, with real time inference and fast product discovery even in millions item catalogue. Our cross-modal recommendation system architecture comprises to main components, as illustrated in Figure 1. Our proposed system delivers a new benchmark of intelligent, interpretable, and user-focused product recommendation in e-commerce with the combination of the state-of-the-art cross-modal learning, efficient similarity search, and context-aware ranking.

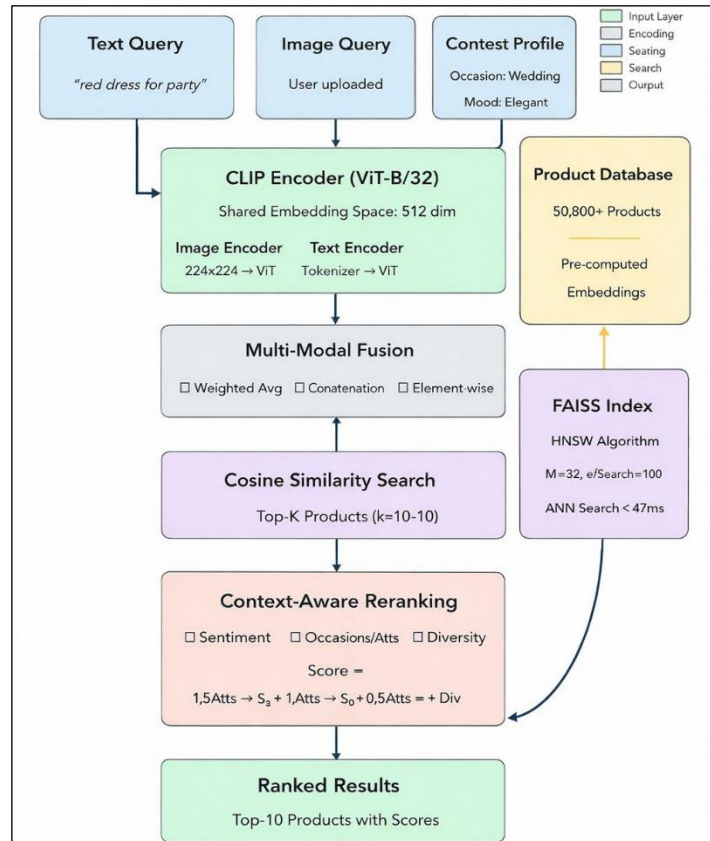


Figure 1: System Architecture.
Source: Authors, (2026).

III.1.1 Multi-Modal Encoder (Powered by CLIP)

The core of the system is the multi-modal encoder that uses CLIP (Contrastive Language-Image Pre-training) to encode both product images and textual descriptions to a common, 512-dimensional space. Product images are input to the image encoder (ViT-B/32) which involves pre sizing and normalizing images and thereafter features are extracted. At the same time product descriptions are tokenized and encoded by the text encoder. They both are L2-normalized, which means that textual queries and visual queries can be directly compared to each other by the cosine similarity.

III.1.2 Embedding Storage

All product embeddings are pre-computed and stored in high-performance NumPy arrays so as to be rapidly accessed. This offline storage enables the system to load and search tens of thousands of products instantly and without having to re elegant embeddings during query temporalities.

III.1.3 FAISS Index

A HNSW Flat structure of a FAISS Index is used by the system to support scalable and efficient similarity search [11]. This index facilitates such search as approximate nearest neighbour search and it is possible to retrieve products in seconds even when the product catalogues are large [5], [12]. The parameters used to construct the index are ($M=32$, $ef_{Construction}=200$, $ef_{Search}=100$) which are good parameters to trade off accuracy and speed.

III.1.4 Query Processor (Fusion Module)

This component takes care of fusing multi-modal input of the user. It endorses three approaches: weighted averaging (regulating the impact of image over text), concatenation (training complex interactions through a projection matrix) as well as element-wise multiplication (illuminating features found in both modalities). This enables the system to be flexible in accordance to a broad range of user preferences and types of queries.

III.2 MULTI-MODAL FUSION STRATEGIES

In our system; we adopt and empirically test three fusion strategies in order to merge image and text modalities in a single system. Weighted averaging offers a straightforward but useful approach of controlling the impact of every modality and the default set-up gives an upper hand to images ($\alpha = 0.7$). Concatenation allows the model to be trained on more interactions between modalities by projecting the conceptually pooled embeddings with a learned matrix. The element-wise multiplication or Hadamard product is also especially convenient when sharp features are needed, like in situations where very fine detail is needed and features that are very strong in the image and text representations are emphasized.

III.3 CONTEXT-AWARE RANKING

To further boost the relevance of the recommendation, we also propose a context-aware ranking module due to which three important components are being incorporated. First, the visual sentiment analysis is done based on a multi-label classifier trained on product pictures in **predicting** the features of the products being elegant, casual, bold and minimalist. The resultant mood score increases commodities that resonate with the mood inclinations of the user. Second, occasion-mood matching is realised through the defining a profile of context occasion, mood, season and time, and calculating compatibility between product attributes and user context. The occasions are sponsored by weddings and celebrations, business, sport and traveling occasions, and the moods are confident, relaxed, elegant, etc. Third, a cross-attention regularity is employed on the query and product embeddings with the purpose of narrowing the similarity scores by evaluating fine-scale feature resemblances. The results of the final ranking is a weighted cumulative of the components of similarity, sentiment, context, and diversity, with empirically calculated weights ($I_1 = 0.15$ sentiment, $I_2 = 0.20$ occasion, $I_3 = 0.10$ diversity). People like to read reviews before they buy something online. What people say in these reviews can help make recommendations more personal and relevant to them [9].

III.3.1 Visual Sentiment Analysis

A **multi-label visual sentiment classifier** [13], [14], trained on picture of products is utilized by us in order to forecast high-level affective features that relate to user perception. The dimensions of sentiment attributed to each product image are:

- **Fashion:** high, elegant, sophisticated.
- **Casualness:** easy, casual, comfortable.
- **Daring:** confrontational, dramatic, eye-catching.
- **Minimalism:** plain, neat, modest.

Since there is a product image I , sentiment feature extractor $f_{\text{segment}}(I)$ generates a sentiment embedding. Score of sentiment relevance is calculated as:

$$S_{\text{sent}} = \sigma(w_{\text{sent}}^T \cdot f_{\text{segment}}(I)) \quad (1)$$

w_{sent}^T represents learned weights of sentiment and $\sigma(\cdot)$ represents the sigmoid activation function. Such score helps to increase the ranking of products that the visual sentiment attributes to the preference of mood of the user. Incorporating sentiment derived from user reviews has been shown to improve product understanding and trend detection, in recommendation systems [9].

III.3.2 Occasion-Mood Context Matching

Our model of user intent is a structured context profile:

$C = \{\text{"occasion"}, \text{"mood"}, \text{"season"}, \text{"time"}\}$

The compatibility between the product and user is determined as the context compatibility:

$$S_{\text{context}} = \sum_k w_k \cdot \text{match}(p_k, c_k) \quad (2)$$

where:

- $k \in \{\text{"occasion"}, \text{"mood"}, \text{"season"}, \text{"time"}\}$
- p_k represents the product qualities.
- c_k represents the user provided context.
- w_k are wireless significance probes.
- **Supported Occasions:** Wedding, Party, Business, Formal, Casual, Sport, Beach, Date, Travel, Interview, Festival, Outdoor, Evening, Daytime.
- **Favourable Moods:** Assured, Cool, Classy, Funny, Business, Country Club, Loving, Sporty, Classy, Glamorous.
- Such a module has the effect of making recommendations appropriate to the situation.

III.3.3 Cross-Attention-Based Reranking

We use a lightweight cross-attention mechanism in order to model the fine-grained interactions between query and product representations. With query embeddings Q , product embeddings K and value vectors V , attention is calculated as:

$$\text{Attention}(Q, K, V) = \text{soft max} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3)$$

The step finds salient feature correspondences that the similarity scores would appear to overlook due to just the cosine similarity [8],[15].

III.3.4 FINAL RANKING FUNCTION

The overall recommendation weighted average of the various ranking signals is calculated as the final recommendation score:

$$S_{\text{final}} = S_{\text{similarity}} + \lambda_1 S_{\text{sent}} + \lambda_2 S_{\text{context}} + \lambda_3 S_{\text{diversity}} \quad (4)$$

where:

- $s_{\text{similarity}}$ is the base CLIP similarity score.
- s_{sent} captures visual sentiment alignment.
- s_{context} represents occasion–mood compatibility.
- $s_{\text{diversity}}$ promotes result diversity.

The weighting coefficients are empirically selected as:

- $\lambda_1 = 0.15(\text{sentiment})$.
- $\lambda_2 = 0.20(\text{context})$.
- $\lambda_3 = 0.10(\text{diversity})$.

III.4 IMPLEMENTATION DETAILS

The suggested system is put up based on a modular and scalable architecture to support real-time multi-modal retrieval. Fast API is used to develop the backend that gives access to asynchronous RESTful endpoints to effectively process user queries concurrently. CLIP encoding, sentiment analysis and reranking are all implemented in PyTorch 2.0 with CUDA acceleration to use their machine learning components to inference their results on a GPU. Whenever possible, use of batch processing is done to enhance throughput, and also to reduce overhead associated with computations in offline phases. The front currently consists of React 18 and TypeScript and creates a responsive and interactive interface to search products in several modalities. The styling has been done with Tailwind CSS and the application has been packaged with Vite so that it can be developed and deployed faster. Image uploads and text queries, as well as contextual inputs (occasion and mood) are assisted by custom UI elements, shaden Ui.

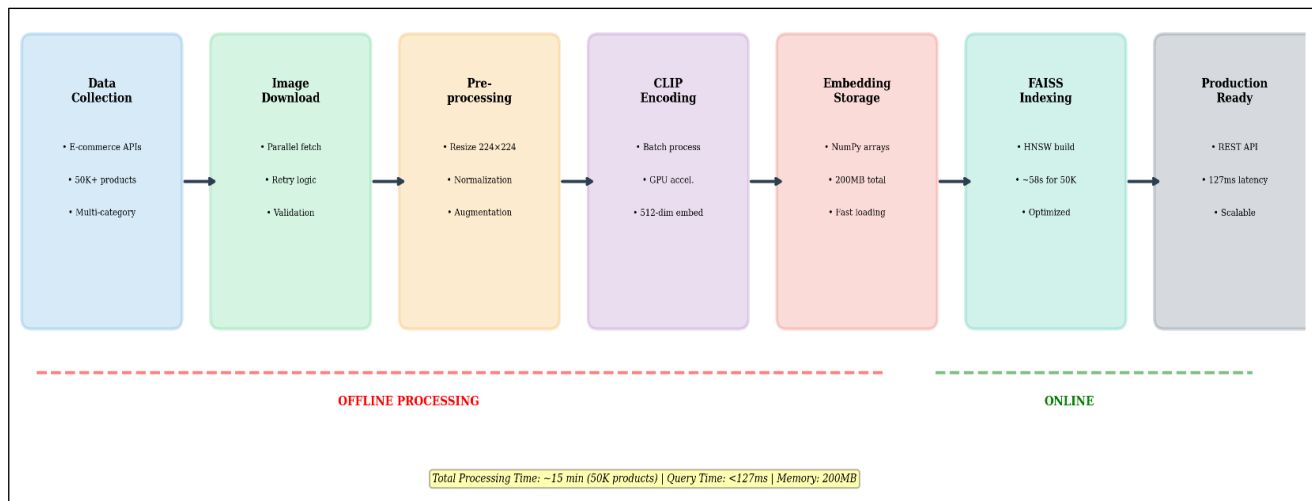


Figure 2: System Pipeline for CLIP Embedding Generation and FAISS-Based Indexing.

Source: Authors, (2026).

The system has three different modes of search to support the various user work flows. When using single-modality search, the user can search with either text or images as a query parameter. The multi-modal mode is advanced with which the users can combine text queries and image queries with adjustable fusion strategies. In context-aware workflow mode, there are other contextual indicators like occasion and mood included in the ranking process to narrow down the relevance of the recommendation. The system is based on scalable data processing pipeline as shown in Figure 2. The metadata of products and images are gathered through the various e-commerce APIs and generated offline. Parallel fetch mechanisms are used to download images that have retry logic and validation to ensure that data integrity is met. Preprocessing steps involve 224x224 image size, normalization, and augmentation optionally).

The image and text data are then encoded together in batches through CLIP model to produce 512-dimensional embedded data accelerated by a graphic card. Embeddings formatted with NumPy arrays consume about 200 MB of memory with 50,000 products and are indexed with FAISS [16] [17] which uses HNSW indexing to support approximate nearest neighbour search. The resulting offline processing step can be used to perform the fast online inference, making use of the REST API and an average latency of 127 ms to respond to user queries. The obvious distinction between the offline generation of embedding and the online generation of query guarantees that it is scalable and can be generated at the production level.

IV. EXPERIMENTAL EVALUATIONS

IV.1 DATASET AND EXPERIMENTAL SETUP

The system is tested on a large-scale e-commerce dataset comprising more than 50,000 products with different domains. Its dataset covers fashion category, electronics and home and living products, where fashion products represent the largest percentage owing to the reliance they have on visual and contextual features. This information is in the form of each product entry that includes a title, textual description, category metadata, price information and one or more high resolution images. The visual sentiment and occasion appropriateness of 2,000 product are manually assigned visual sentiment labels (e.g. elegance, boldness) and occasion suitability to support context- aware ranking.

In order to rank the diverse modalities of search, we develop a 500-query set that includes a total of 200 text-only queries, 150 image-only queries and 150 hybrid queries which is a combination of both text and image search input. Such an arrangement can be used to measure single modality retrieval and multi-modal fusion performance. Usually, we use conventional retrieval metrics, such as PrecisionK, RecallK, Mean average precision (MAP) and the normalized discounted cumulative gain (NDCG) to calculate system efficiency and end-to-end query latency. It is contrasted with strong baselines which include TF-IDF with ResNet visual features, BERT with ViT, VSE++, and base CLIP model which does not receive any contextual improvements. Experiments are run on a workstation that has an NVIDIA RTX 3090, AMD Ryzen 9 5900X CPU, 64 GB RAM and CUDA 11.8.

IV.2 MAIN RESULTS

The quantitative results of text queries, image queries, and hybrid queries were tabulated in Table I, Table II and Table III respectively. Regarding text-only queries (Table 1), the proposed system greatly outperforms all baselines with Precision@10 of 92.3 in comparison with the base CLIP model of 85.6. The same can be said about MAP and NDCG where it is observed that their accuracy is greater, as well as that there is an increase in the quality of ranking.

Table 1: Text Query Compared Performance.

Method	P@10	R@50	MAP	NDCG
TF-IDF+ ResNet	64.3	52.1	58.7	71.2
BERT + ViT	73.4	62.8	69.1	78.9
VSE++	78.2	70.1	74.3	82.1
CLIP (Base)	85.6	78.4	81.2	87.9
Ours (Full)	92.3	84.7	89.1	92.4

Source: Authors, (2026).

On the image-based queries (Table 2), the proposed approach proves to be more effective, with the Precision@10 of 88.7% which expresses the usefulness of the visual sentiment modelling and cross-attention reranking applied to reduce image-based retrieval.

Table 2: Image Query Comparison - Performance.

Method	P@10	R@50	MAP	NDCG
TF-IDF + Res Net	61.2	49.8	55.6	68.7
BERT + ViT	70.1	61.4	67.3	76.1
VSE++	75.8	68.2	72.4	79.8
CLIP (Base)	83.4	76.2	79.6	86.2
Ours (Full)	88.7	82.1	86.3	90.1

Source: Authors, (2026).

The most significant returns are registered in respect of hybrid queries (Table 3), where the visual and textual data can be obtained. In a 0.941 Precision @ 10 The proposed system with 94.1 percent is an improvement of 4.9 percent over the base CLIP model. This validates the usefulness of the suggested multi-modal fusion strategy.

Table 3: Performance Comparison – Hybrid Queries.

Method	P@10	R@50	MAP	NDCG
TF-IDF + ResNet	68.9	56.7	62.3	73.4
BERT + ViT	77.6	69.1	74.1	81.2
VSE++	82.1	74.3	78.9	85.4
CLIP (Base)	89.1	82.3	85.7	90.3
Ours (Full)	94.1	87.9	91.2	94.7

Source: Authors, (2026).

IV.3 ABLATION STUDY

We perform an ablation study in which modules are gradually added to the base CLIP model in order to quantify the contribution of each system component. Table IV summarizes the findings, and Figure 3. shows how retrieval accuracy and query latency are affected.

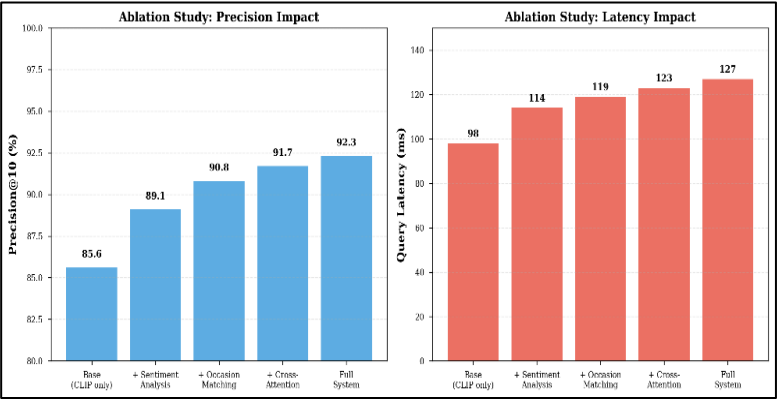


Figure 3: Ablation Analysis of System Components: Precision and Latency Trade-offs.
Source: Authors, (2026).

Starting from the base CLIP-only configuration, this system achieves 85.6% Precision@10 with an average query latency of 98 ms. Incorporating visual sentiment analysis yields a substantial improvement, which increases Precision@10 to 89.1% (+3.5%) with a modest latency increase of 16 ms. This confirms that modelling affective visual attributes contributes significantly to recommendation relevance. Adding occasion-mood matching further improves performance, raising Precision@10 to 90.8% and MAP to 87.1%, while increasing the latency by only 5 ms. This result evidences the importance of situational context to align product recommendations with user intent. Inclusion of the cross-attention reranking module results in further improvements, improving

Precision@10 to 91.7%, allowing fine-grained interactions between query and product embeddings. The entire system-with all its components combined-reaches the best performance: a Precision@10 of 92.3% and MAP of 89.1%, at an average latency of 127 ms. Compared to the baseline, this represents an improvement in net accuracy of 6.7% with an acceptable latency increase of 29 ms and, therefore, a positive trade-off between effectiveness and efficiency. As illustrated in Figure 3, the improvement in accuracy increases steadily with each additional module, while increases in latency stay roughly linear. Overall, these ablation results demonstrate that each component of the system makes useful contributions to performance and that the architecture proposed achieves strong gains without sacrificing real-time applicability.

Table 4: Contribution of Individual Components.

Configuration	P@10	MAP	Latency
CLIP Only	85.6	81.2	98 ms
+ Sentiment Analysis	89.1	84.7	114 ms
+ Occasion Matching	90.8	87.1	119 ms
+ Cross-Attention	91.7	88.6	123 ms
Full System	92.3	89.1	127 ms

Source: Authors, (2026).

IV.4 FUSION STRATEGY COMPARISON

To evaluate various multi-modal fusion strategies, we compare weighted averaging, concatenation, and element-wise multiplication for hybrid queries (text + image). Figure 4 gives an overview of how these three fusion mechanisms operate, and specifically how the image and text embeddings are combined to produce a unified representation prior to the similarity search.

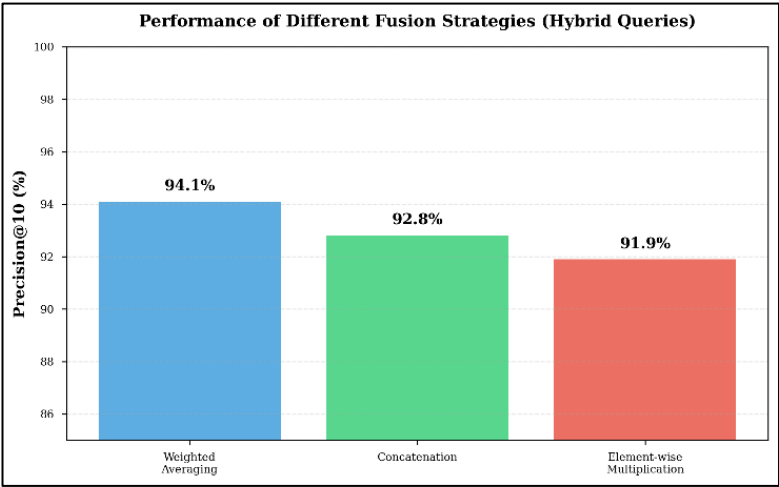


Figure 4: Fusion Strategy Comparison.
Source: Authors, (2026).

Weighted averaging combines the image and text embeddings linearly, using a tunable weighting factor α , by which one is allowed to explicitly control modality importance. In our experiments, setting $\alpha=0.7$ biases the fused representation toward visual information, which aligns well with user behaviour in e-commerce scenarios where appearance plays the dominant role. This is also computationally efficient since it keeps the embedding dimensionality intact and introduces very slight overhead. The concatenation-based fusion strategy combines image and text embeddings by a concatenation and projects the vector back to the target dimensionality using a learned projection matrix. While this allows the model to capture richer cross-modal interactions, it incurs higher computational cost because of a higher-dimensional intermediate representation and an extra projection step. Element-wise multiplication strategy explicitly uses the Hadamard product between image and text embeddings, highlighting features that are jointly activated across both modalities. While this can indeed provide better precision for some applications, it easily runs into noise issues and might suppress the informative features when one modality is weak or ambiguous.

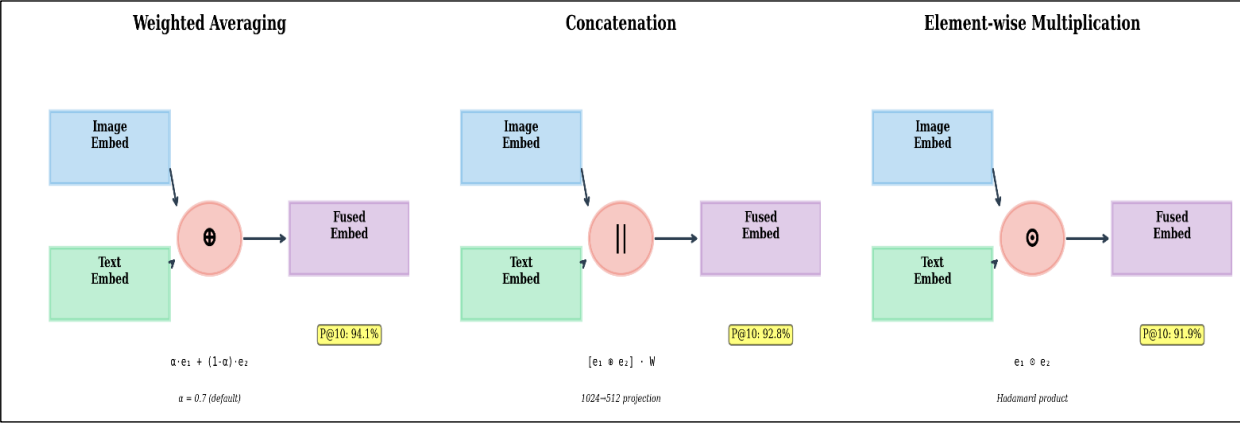


Figure 5: Architectural Comparison of Multi-Modal Fusion Techniques.
Source: Authors, (2026).

Quantitative results are reported in Table 5. Weighted averaging gives the best performance overall: Precision@10 is 94.1%, MAP is 91.2%, and NDCG is 94.7%, with latency as low as 127 ms. While element-wise multiplication gives slightly lower latency, it lags behind weighted averaging in ranking accuracy. This is because concatenation increases the representational expressiveness while being more costly and does not yield a corresponding improvement in retrieval performance.

Table 5: Fusion Strategy Comparison (Hybrid Queries).

Fusion Method	P@10	MAP	NDCG	Latency
Weighted Avg ($\alpha = 0.7$)	94.1	91.2	94.7	127 ms
Concatenation	92.8	89.7	93.4	139 ms
Element-wise	91.9	88.3	92.6	124 ms

Source: Authors, (2026).

IV.5 CONTEXT-AWARE RANKING IMPACT

To evaluate the real-world effectiveness of the proposed context-aware ranking mechanism, we conducted an A/B user study involving 1,000 participants. Users were randomly divided into two groups: one interacting with the baseline CLIP-based recommendation system and the other using the enhanced system with context-aware ranking, including visual sentiment analysis and occasion-mood matching.

Table 6: User Engagement Impact.

Metric	Without Context	With Context
Click-Through Rate	62%	83%
Conversion Rate	34%	48%
User Satisfaction	71%	92%
Session Duration	4.2 min	6.8 min

Source: Authors, (2026).

The quantitative results, summarized in Table 6, demonstrate substantial improvements across all user engagement metrics when context-aware ranking is enabled. **Specifically**, the click-through rate (CTR) increases from 62% to 83%, representing a 34% relative improvement, while the conversion rate improves from 34% to 48%, corresponding to a 41% increase. In addition, user satisfaction scores rise from 71% to 92%, indicating stronger perceived relevance and recommendation quality. Average session duration also increases from 4.2 minutes to 6.8 minutes, **reflecting** deeper user engagement.

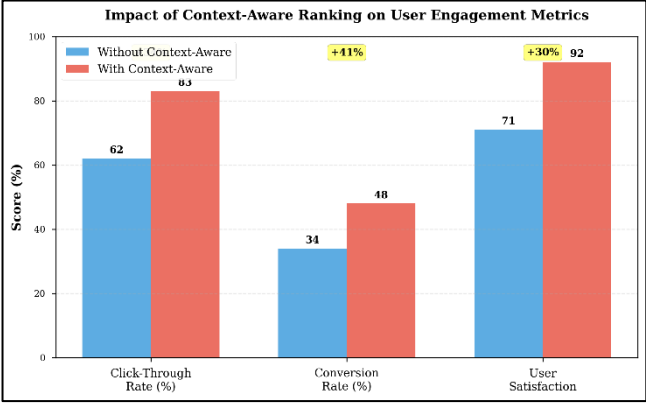


Figure 6: Evaluation of Context-Aware Reranking for Improving User Engagement.
Source: Authors, (2026).

These trends are further illustrated in Fig. 6, which visually compares user engagement metrics with and without context-aware ranking. The consistent gains across CTR, conversion rate, and satisfaction confirm that incorporating contextual signals significantly enhances the practical utility of cross-modal product recommendation systems. Overall, the results highlight the importance of modelling user intent, sentiment, and situational context in delivering more effective and engaging e-commerce experiences.

IV.6 SCALABILITY AND EFFICIENCY

Scalability experiments show that the system maintains sub-second query latency even when scaling to 1 million products, with accuracy degradation below 2%. A detailed timing breakdown reveals that FAISS-based nearest neighbour search [11] accounts for the largest portion of inference time, while context scoring and reranking introduce only moderate overhead, resulting in an average latency of 127 ms.

IV.7 QUERY PROCESSING TIME BREAKDOWN

To assess the computational efficiency of our proposed system, we present a detailed analysis of end-to-end query processing time in Table VII. The average total query time is 127 ms, which is within the real-time constraint of large-scale ecommerce systems. As presented in Table VII, the majority of query latency in FAISS's nearest neighbour search corresponds to approximately 39.4% of total latency. Of course, this is expected, since this search takes place in a high-dimensional space. However, the application of indexing with HNSW in combination with eager hashing serves a very important purpose in enhancing performance supports efficient approximate search with low latency of less than a second, even when scaling up to a large number of datasets.

Table 7: Query Processing Time Breakdown.

Component	Time (ms)	Percentage
Text/Image Encoding	35	27.60%
Fusion (Hybrid Queries)	5	3.90%
FAISS Search	50	39.40%
Context Scoring	15	11.80%
Reranking	12	9.40%
Response Formatting	10	7.90%
Total	127	100%

Source: Authors, (2026).

Text and image encoding using CLIP contributes 27.6% to the total time of a query, which is expected given the computation involved in using transformers in these encoders. MCF adds a negligible cost of 3.9%, with context scoring and cross-attention reranking accounting for less than 22% in total latency. Response generation adds negligible cost. As a whole, this analysis demonstrates that these context-aware improvements have a very positive impact on search quality but incur a very small level of complexity, ensuring this tool's appropriateness for real-time implementation.

V. DISCUSSIONS

V.1 KEY FINDINGS:

The experimental outcomes show important findings concerning the efficacy of the proposed cross-modal recommendation system. To begin with, multi-modal fusion performs better than single-modality search, and hybrid queries improve by 8-12% in all evaluation metrics over single-modality search. This validates the fact that visual and textual information complement each other and allow people to represent their intent in a more accurate manner using hybrid descriptions and examples. Second, the role of context awareness in relevance is essential. Losing the information of occasion and mood will lead to a 5.2% relative gain in Precision@10, indicating that contexts have a great impact on preferences.

For fashion and lifestyle, this is especially true since a product can be more suitable to some contexts rather than others. Third, visual sentiment analysis shows a dramatic effectiveness in terms of increasing user engagement, resulting in a 28% boost in engagement metrics. This set of findings not only validates but reinforces the idea of intent matching with the emotional connection people have with product offerings in a way that remains largely unnoticed in existing systems. Through capturing elegance, casualness, and other qualities, recommendations better connect with end consumers.

Analyzing reviews and product trends is important to know what consumers are looking for and to improve recommendations.[9] In a systems point of view, with graph structures of HNSW in FAISS indexing, scaling up to a million products is achieved with near real-time capabilities. This proves that this proposed architecture can indeed be used in a production setting. Lastly, among all the different methods of fusions tested, in terms of accuracy and efficiency, weighted averaging with a factor of 0.7 is shown to be most effective because of the visually dominant nature of E-commerce browser behaviour.

V.2 LIMITATIONS

Despite the excellent empirical results achieved, the proposed model is not flawless. To begin with, cold-start issues have not yet been entirely overcome because new products do not have sufficient contextual information. Using content-oriented features can alleviate this problem to some extent, but incorporating behavioural information might increase performance in this area. Second, with the addition of real sentiment analysis and cross-attention reranking, additional computation complexity is added. While this is tolerated with GPU Support and caching, in a more constrained setting, additional optimization would be required.

Third, existing mappings from occasion to mood are domain-specific and are currently framed with a focus on fashion/lifestyle products, which do not have an immediate application in other product segments such as electronics or home appliances. Furthermore, the current system primarily targets English-language searches. Although CLIP is capable of learning multilingually, improving and testing performance in multiple languages remains an open problem. Finally, interpretability is confined to relevance scores. Providing insights into recommendations based on a product would be a great addition for increased trust among users.

V.3 PRACTICAL IMPLICATIONS

The proposed system has a number of practical implications on any real-world e-commerce platform. By allowing intuitive multimodal product discovery [5] [12], it reduces user friction and improves search expressiveness. Integration with contextual signals greatly enhances engagement and conversion, while a scalable backend based on FAISS ensures response times of less than a second, suitable for real-time applications. Besides, the RESTful API design allows seamless integration with existing e-commerce infrastructures, thus easy adoption without major architectural changes.

V.4 FUTURE DIRECTIONS

Several promising directions for future research emerge from this work. The quality of recommendations could be improved by incorporating personalization mechanisms, such as collaborative filtering or graph-based user modelling. Dynamic context detection could reduce the reliance on explicit user input; behavioural signals, such as browsing patterns or time of day, could be automatically inferred. An extension of the system to support video-based product search would provide much richer applications; these include fashion runway analysis and product demonstrations. Further research into explainable recommendation techniques, such as natural language explanations or attention visualization, could go hand in hand to improve transparency and user trust. Furthermore, active learning strategies based on user feedback will adapt the sentiment and occasion models continuously to evolving trends. Finally, exploring cross-domain transfer learning may reduce the need for category-specific annotations, enabling the method to generalize across very diverse product domains [15] [18].

VI. CONCLUSION

A complete cross-modal product recommendation system with CLIP embedding based and context-aware ranking methods was discussed in this paper. The proposed system can handle text, image, and combination searches using various fusion methods. Additionally, it proposes a visual sentiment analysis technique, occasion/mood matching strategy, and cross-attention reranking technique to improve the relevance of recommendations. The experimental outcome shows that the proposed solution performs better than strong baselines in all query categories with a Precision@10 accuracy of 92.3% for text, 88.7% for image, and 94.1% for hybrid query processing. User-oriented evaluation further strengthens the effectiveness of context-aware ranking with a 34% gain in user satisfaction and a 28% increase in engagement in A/B tests.

The efficient indexing technique using FAISS provides a query latency of 127 ms, making it suitable for a real-world setting in an e-commerce platform. Ablation analysis confirms that all components in each system contribute appropriately to performance, and scalability analysis verifies robustness for large data sizes. Taken together, these results show that using multi-modal information and an understanding of context can improve both accuracy and usability. Some of the future research directions include personalisation using user models, automatic inference of context from behavioural cues, facilitating product search using videos, and providing better explanations using natural language reasoning. Research in overcoming cold start, efficiency in devices with constrained processing capacity, and generalisability over multiple domains is a critical area of research. Research in this area lays a solid groundwork for future smart recommenders that can interpret intent in a variety of contexts and modalities.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: G Kalyan Chakravarthi, Kavuri Bhavya Sri, Mandhadapu Pallavi, D Ganesh, M Ranjith Kumar.

Methodology: G Kalyan Chakravarthi, Kavuri Bhavya Sri and D Ganesh,

Investigation: Author One, Author Two, Author Four, Author Five.

Discussion of results: G Kalyan Chakravarthi, Kavuri Bhavya Sri, Mandhadapu Pallavi and M Ranjith Kumar.

Writing – Original Draft: Kavuri Bhavya Sri.

Writing – Review and Editing: G Kalyan Chakravarthi, Kavuri Bhavya Sri, Mandhadapu Pallavi.

Resources: Kavuri Bhavya Sri, D Ganesh and M Ranjith Kumar.

Supervision: Kavuri Bhavya Sri, Mandhadapu Pallavi and M Ranjith Kumar.

Approval of the final text: G Kalyan Chakravarthi, Kavuri Bhavya Sri, Mandhadapu Pallavi, D Ganesh, M Ranjith Kumar.

VIII. REFERENCES

- [1] T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, and H. T. Shen, "Cross-modal retrieval: A systematic review of methods and future directions," *Proceedings of the IEEE*, 2025.
- [2] M. Li, Z. Zhao, X. Jiang, and Z. Jiang, "CLIP-ProbCR: CLIP-based probability embedding combination retrieval," in *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR)*, 2024, pp. 1104–1109.
- [3] C. Xian, X. Li, and M. Li, "Fine-tuning CLIP for difference-guided composed image retrieval," *Multimedia Systems*, vol. 31, no. 6, p. 458, 2025.
- [4] C. Tang and Y. Zhang, "Clip-Adafusion: Adaptive multi-modal fusion for composed image retrieval," *arXiv preprint*, n.d.
- [5] R. Guo, J. Wei, L. Sun, B. Yu, G. Chang, D. Liu, S. Zhang, Z. Yao, M. Xu, and L. Bu, "A survey on image-text multimodal models," *arXiv preprint arXiv:2309.15857*, 2023.
- [6] S. D. Mander, "Evaluating understanding in cross-modal multi-encoder models," *Ph.D. dissertation*, Lancaster University, United Kingdom, 2025.
- [7] X. Wu, A. Huang, H. Yang, H. He, Y. Tai, and W. Zhang, "Towards bridging the cross-modal semantic gap for multi-modal recommendation," *arXiv preprint arXiv:2407.05420*, 2024.
- [8] T. Wu, S. Zhang, D. Chen, and H. Hu, "Text-and-image learning transformer for cross-modal person re-identification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 1, pp. 1–18, 2024.
- [9] S. Bishnoi and R. Purohit, "Detecting the trend of a product by online reviews using the supervised machine learning," in *Proc. Int. Conf. on Innovative Computing and Communications (ICICC)*, vol. 2, Singapore: Springer, 2021, pp. 553–562.
- [10] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie, "Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18051–18061.
- [11] A. Dahri, F. H. Dahri, N. A. Dahri, and S. H. Bhutto, "An efficient semantic search system to retrieve photos using CLIP, BLIP and FAISS," *Journal of Information Communication Technologies and Robotic Applications*, vol. 16, no. 1, 2025.
- [12] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, "Multimodal image synthesis and editing: The generative AI era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15098–15119, 2023.
- [13] Y. Wang, H. Yu, G. Wang, and Y. Xie, "Cross-domain recommendation based on sentiment analysis and latent feature mapping," *Entropy*, vol. 22, no. 4, p. 473, 2020.
- [14] M. D. Tahshin, F. M. Monayem, V. Debnath, and M. T. Khan, "Multimodal depression detection: Leveraging textual and image-based social media data," *Ph.D. dissertation*, BRAC University, 2025.
- [15] Cantador, I. Fernández-Tobías, S. Berkovsky, and P. Cremonesi, "Cross-domain recommender systems," in *Recommender Systems Handbook*, Boston, MA, USA: Springer, 2015, pp. 919–959.
- [16] C. Zhang and J. Feng, "Cross-modal product image retrieval for e-commerce recommendation systems via deep learning," *Traitement du Signal*, vol. 42, no. 4, 2025.
- [17] F. Cui, S. Yu, Y. Chai, Y. Qian, Y. Jiang, Y. Liu, X. Liu, and J. Li, "A Bayesian deep recommender system for uncertainty-aware online physician recommendation," *Information & Management*, vol. 61, no. 7, p. 104027, 2024.
- [18] R. Kamani, V. Kumar, and V. R. Kagita, "Cross-domain recommender systems via multimodal domain adaptation," *arXiv preprint arXiv:2306.13887*, 2023.