Deep-View X-Modalities Visio-Linguistics (DV-XML) Features Engineering Image Retrieval Framework

Adel Alkhalil
Department of Software Engineering, College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia
a.alkalel@uoh.edu.sa (corresponding author)

Received: 9 January 2025 | Revised: 5 February 2025, and 16 February 2025, and 20 February 2025 | Accepted: 21 February 2025

 DOI: https://doi.org/10.48084/etasr.10175

ABSTRACT

This research proposes an advanced framework for efficient image retrieval by integrating visual and linguistic modalities into a unified system. The Deep-View X-Modalities Visio-Linguistics (DV-XML) framework is designed to handle user queries that include both text and image inputs while allowing modifications to align with user preferences. By employing a multimodal Content-Based Image Retrieval (CBIR) system, the framework combines features extracted by a ResNet-50 model for images and a Bidirectional Encoder Representations from Transformers (BERT) model for textual data. These features are harmonized using an inductive learning-based fusion methodology within Multi-Layer Perceptrons (MLPs). A novel Reverse Re-ranking (RR) algorithm enhances retrieval accuracy by optimally aligning the combined representations with the target images during inference. Extensive evaluations on the Fashion-200K and MIT-States datasets demonstrate the model's superior performance compared to baseline CBIR methods. This work advances the field by efficiently merging dual modalities and streamlining the retrieval process with innovative RR strategies, setting a benchmark for future research in multimodal image retrieval systems.

Keywords-image retrieval framework; image and textual embedding extraction; ResNet; BERT-base/large; multimodal embedding fusion; reverse re-ranking

I. INTRODUCTION

The rapid growth of digital information in the modern global community presents people with huge amounts of visual and textual data, making it increasingly challenging to efficiently retrieve desired content. Content-Based Image Retrieval (CBIR) systems address this issue by retrieving images that align with user preferences based on input queries, which may be visual, textual, or a combination of both [1-4]. Traditional CBIR systems often rely on shallow machine learning and symbolic AI approaches, which limit their ability to handle complex user requirements, particularly those involving dual-modality queries. This limitation has motivated researchers to explore more robust techniques for bridging the gap between visual and linguistic modalities to enhance retrieval accuracy. CBIR systems can be classified into three main categories: those that process image-only queries, those that rely on textual queries, and those that integrate visual and textual inputs for enhanced retrieval. The third category is particularly significant as it allows users to modify their visual queries with additional textual descriptions, such as specifying changes in color, shape, or context [1, 2]. This capability is consistent with real-world scenarios where users often describe nuanced requirements using both images and text. An illustration of a real-world CBIR system in operation is depicted in Figure 1, where a user instructs an agent to locate a shoe using an image query, while also specifying changes to the color and shape of the shoe in text.

Despite advancements, several challenges remain in dual-modality image retrieval systems. These include effectively extracting meaningful features from both visual and textual data, and seamlessly fusing these modalities into a compact, unified representation [1-3]. To address these challenges, we propose a novel Deep-View X-Modalities Visio-Linguistics (DV-XML) framework. This framework employs a ResNet-50 architecture for visual feature extraction and leverages the Bidirectional Encoder Representations from Transformers (BERT) model to generate semantic and contextual textual embeddings. By integrating these features through an inductive learning-based approach within Multi-Layer Perceptrons (MLPs), the framework ensures a robust, unified representation of dual-modality inputs. Furthermore, the proposed DV-XML framework introduces an innovative Reverse Re-ranking (RR) methodology during the inference stage. This method optimally aligns combined query representations with target images, thereby enhancing retrieval precision and reducing computational complexity [5, 6]. The framework's performance is evaluated on two diverse datasets, Fashion-200K and MIT-States, which are widely recognized benchmarks for assessing

multimodal CBIR systems. The results demonstrate significant improvements over existing baseline methods, validating the effectiveness of the proposed approach.

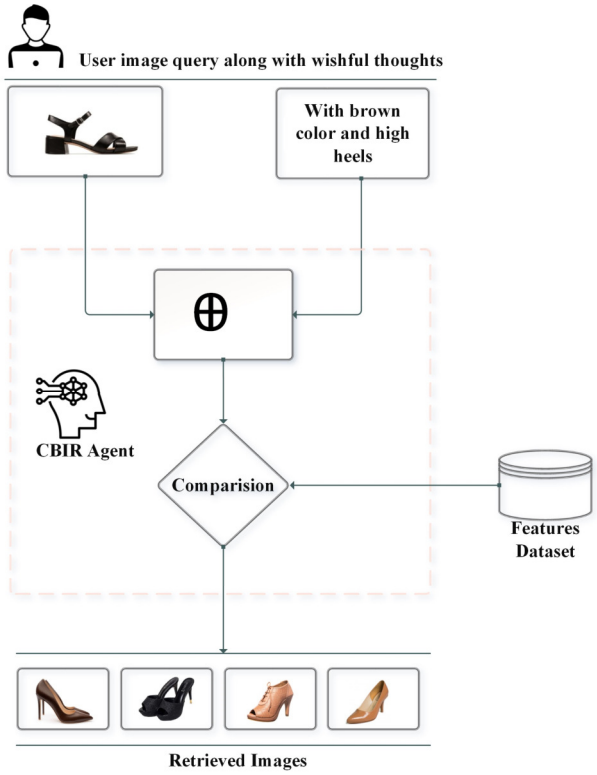


Fig. 1. CBIR system example in a daily activity.

The contributions of this study are summarized as follows:

- Development of a robust CBIR framework that integrates visual and linguistic modalities through ResNet-50 and BERT models.
- Introduction of an inductive learning-based feature fusion approach for generating unified representations.
- Implementation of an RR algorithm to enhance retrieval efficiency and accuracy during inference.

II. RESEARCH PROBLEM

Information retrieval systems are challenged by the vast availability of multimodal data containing both visual and textual information. CBIR systems must navigate complex requirements involving the fusion of visual and linguistic data to meet user expectations. The primary challenge is to create an efficient and accurate dual-modality framework capable of processing and understanding heterogeneous data sources to deliver relevant results. This section explores the research problem in detail, identifying key gaps in existing approaches and outlining the motivation for the proposed framework.

A. Challenges in Dual-Modality Retrieval

The main challenges in dual modality retrieval are:

- Feature extraction from distinct modalities: A fundamental challenge in CBIR is the extraction of meaningful and representative features from visual and textual modalities. While deep Convolutional Neural Networks (CNNs) such as ResNet have shown success in extracting detailed visual features, textual data require models such as transformers (e.g., BERT) to capture contextual and semantic meanings. Integrating these two inherently different data types into a unified representation remains a non-trivial task.
- Effective fusion of visual and textual features: Existing CBIR systems struggle with effectively fusing visual and textual features into a compact embedding space. Traditional techniques such as concatenation, average pooling, or gating mechanisms often fall to capture the nuanced interplay between modalities. This results in suboptimal representations that do not generalize well to diverse queries.
- Handling complex user queries: Users often combine visual inputs (e.g., an image) with textual descriptions that specify modifications or preferences (e.g., "change the color to red" or "make it rounder"). This is missing in current methods.
- Scalability and computational efficiency: Real-world applications of CBIR require systems that can efficiently scale to large datasets without compromising retrieval accuracy. High computational costs associated with Deep Learning (DL) architectures and feature fusion processes often hinder the deployment of existing systems in practical scenarios.
- Retrieval optimization and ranking: Another critical challenge is to optimize the ranking of retrieved images to ensure that the most relevant results are prioritized. Existing CBIR frameworks often use simple ranking techniques that can lead to suboptimal retrieval performance, particularly for dual-modality queries.

B. Limitations of Existing Approaches

The limitations of existing approaches are:

- Early CBIR systems relied on shallow learning methods and symbolic AI, which were limited in their ability to integrate multimodal data effectively.
- Recent advancements, such as DenseBert4Ret [1] and DvLIL [2], have introduced feature fusion techniques and transformer-based architectures, but these methods still face limitations in handling complex user-defined queries and achieving computational efficiency.
- The high computational overhead of existing models, such as those using LSTMs or complex gating mechanisms [7, 8], limits their applicability in real-world, time-sensitive environments.

C. Motivation for the Proposed Framework

To address these challenges, this research introduces the DV-XML framework, which focuses on:

- Developing a robust feature extraction pipeline by leveraging ResNet-50 for visual features and BERT for textual embeddings.
- Employing an inductive learning-based approach to ensure seamless and efficient fusion of visual and textual features into a unified representation.
- Incorporating a novel RR mechanism to optimize the alignment between user queries and target images, thereby enhancing retrieval precision.
- Ensuring scalability and computational efficiency for handling large datasets through streamlined processing pipelines and post-processing techniques.

By addressing the above research problem, the proposed DV-XML framework aims to set new benchmarks in multimodal CBIR systems, enabling accurate, efficient, and user-centric image retrieval in diverse real-world scenarios. Figure 2 visually demonstrates the workflow of the DV-XML paradigm, incorporating mathematical notations for clarity. The symbols f_i and f_t denote the processes of extracting features from images and text, respectively. The operations represented by Ω and Φ signify the role of MLPs, which facilitate the generation of the joint representation J. Additionally, the similarity coefficient, denoted as S_{RR}, is a critical component applied within the RR mechanism to improve retrieval performance. This figure aims to provide a clear representation of the framework's feature integration and ranking process, emphasizing the interaction between visual and textual modalities.

III. LITERATURE REVIEW

The field of CBIR has seen significant advancements through the integration of dual modalities, namely visual and textual data. Researchers have explored various methodologies to enhance the retrieval accuracy and efficiency of such systems, which are summarized in Table I. Key approaches include the use of CNNs, Recurrent Neural Networks (RNNs), and attention mechanisms for feature extraction, as well as different strategies for feature fusion and query representation. The AutoRet framework described in [6] employs a Deep Convolutional Neural Network (DCNN) to extract features from various segments of an image, which are then grouped into unique clusters. Training is optimized within a fixed number of clusters, enhancing the systems retrieval accuracy and efficiency. The method described in [9] defines the visual attributes of objects, with model training and embedding learning based solely on interpretations of the features of related entities. To create a composite representation, the model modifies the visual features of the images according to the text embedding. This approach uses a limited set of attributes for the query image.

Significant research has been conducted on dual-modality frameworks, leading to the development of the Text Image Residual Gating (TIRG) framework [7], which generates a unified representation of text and visual features. In this approach, a CNN architecture extracts visual features, while an LSTM network is used to obtain textual features. A gating mechanism then integrates both sets of features into a common representation. However, the method suffers from high computational cost due to convolutional processes and effectiveness limitations, particularly due to the LSTM's unidirectional text feature extraction. An enhanced version of the TIRG framework is presented in [10], employing ResNet for combined image and text feature fusion with a simplified gating mechanism. The Visio-linguistic Attention Learning (VAL) model [8] introduces a composite architecture that effectively combines visual and linguistic features by aligning visual features with language semantics. This semantic adaptation is achieved by incorporating multiple textual features at various layers, resulting in a more enriched and integrated representation. In this model, an RNN is used to process the textual features within the framework.

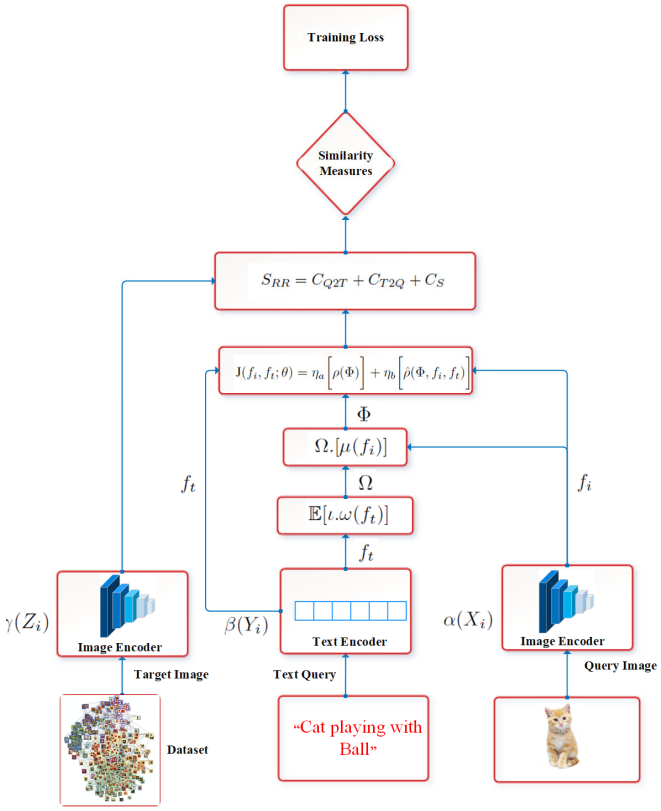


Fig. 2. Flowchart of the proposed DV-XML inference-based image retrieval framework.

DL has seen significant progress since the development of AlexNet [11]. The authors utilized the well-known ImageNet dataset, which contains a million images divided into a thousand classes, to train and evaluate the model. In addition to increasing the model's depth, essential architectural features were introduced, including max-pooling layers, padding, ReLU for nonlinear activation, and a softmax layer for classification. Another notable DL benchmark, ZFNet [12], offers an enhanced version of AlexNet with improved accuracy. A key difference between ZFNet and AlexNet is the choice of filter size: ZFNet uses 7×7 filters, whereas AlexNet uses larger 11×11 filters. This difference is important in evaluating the two models, as the use of larger filters in AlexNet can result in the

loss of considerable pixel information. The VGGNet architecture, introduced in [13], is widely recognized as a top-performing model in computer vision applications. In particular, VGG16 is notable for its simple and consistent design, which relies on 3×3 convolutional layers with a stride of 1 and consistent padding. The model also incorporates max-pooling layers with 2×2 filters and a stride of 2. Towards the end of the network, two fully connected layers are used, culminating in a softmax layer for classification output. The DenseBert4Ret [1] and DvLIL [2] frameworks introduce dual-modality architectures that handle both image and text embeddings. DenseBert4Ret utilizes DenseNet to extract deep features from visual data and includes a transformer to process text-based query features. DvLIL, on the other hand, employs ResNet-50 to analyze reference images and BERT-base to generate textual embeddings, thereby enhancing computational efficiency. Both DenseBert4Ret and DvLIL use learning strategies to achieve a unified feature representation across modalities. The autoencoder-based framework presented in [14] uses unsupervised learning approaches for image similarity that effectively learn image representations and assess similarity through cosine distances. The DQU-CIR framework [15] processes multimodal data by integrating textual and visual queries into a unified representation. It employs BLIP-2 for image-text alignment and a Large Language Model (LLM) for textual understanding. The framework encodes text using a Vision-Language Pre-trained Model (VLPM) textual encoder, whereas the corresponding visual features are extracted via the VLPM vision encoder. An MLP is then used to fuse these extracted features into a single representation. The effectiveness of the model was evaluated on four datasets, three from the fashion domain and one from a real-life scenario.

The TG-CIR framework [16] follows a three-stage retrieval process. In the first stage, both high-level and fine-grained features are extracted from source images, target images, and text queries using VLPM-based visual and textual encoders. To minimize interdependencies among queries and improve generalization, it incorporates orthogonal regularization. The second stage leverages transfer learning, enabling the model to refine multimodal queries autonomously. This helps in resolving conflicts between the reference image and textual modifications, leading to better query composition and retrieval accuracy. In the final stage, TG-CIR applies two types of loss functions: (1) a batch-wise classification loss, and (2) a batch-wise similarity-guided matching loss, both of which contribute to improving ranking performance. The framework has been evaluated using fashion-oriented and real-world datasets. The CLIP4CIR framework [17] builds upon CLIP, a Vision-Language Pre-trained (VLP) model, to extract and align textual and visual features. After feature extraction, it employs a combiner network to integrate these representations into a cohesive multimodal embedding. During training, batch-wise contrastive loss is applied between the fused features and target images to fine-tune the parameters. The model has been tested on several datasets, including Fashion-IQ [18], Fashion-200K [19], and CIRR [20], demonstrating exceptional performance despite its relatively simple architectural design.

Recent advances in DL have significantly improved multimodal image retrieval frameworks, particularly in terms of feature extraction, fusion techniques, and retrieval optimization. Several studies have explored methods that integrate neural networks to improve classification, retrieval accuracy, and scalability. For instance, authors in [21] introduced DNNBoT, a deep neural network-based model for botnet detection and classification, demonstrating the efficacy of DL in high-dimensional data processing and classification tasks. While their approach focuses on cybersecurity, the underlying principles of deep feature extraction and classification align with the objectives of our DV-XML framework in refining retrieval accuracy. Moreover, authors in [22] proposed the intelligent and smart navigation system for visually impaired friends leveraging DL techniques to process multimodal inputs for decision-making and navigation. Their study highlights the importance of integrating multiple data modalities to enhance AI-driven decision systems, a concept central to our proposed inductive learning-based feature fusion methodology. Additionally, authors in [23] introduced a spatial-spectral image classification with edge preserving method to improve feature preservation in image-based classification tasks. Their method enhances the spatial-spectral image representation, which is relevant to our ResNet-50-based image feature extraction, by ensuring that essential image attributes remain intact during processing.

The comparative analysis presented in Table I highlights the evolution of CBIR frameworks, focusing on their strategies for feature extraction and fusion from dual modalities. Early approaches such as AutoRet relied solely on visual features, limiting their scope to image-only queries. Later models such as TIRG and VAL introduced mechanisms to integrate textual features, albeit with limitations in computational efficiency and handling of complex textual queries. DenseBert4Ret and DvLIL made significant advancements by leveraging transformer-based architectures for textual feature extraction and using advanced fusion strategies to enhance dual-modality representation. The literature review on CBIR reveals that further work is essential to advance the bimodal CBIR framework. In bimodal CBIR systems, a textual query serves to alter object attributes, such as color, shape, contour, and environment, according to human intent. In contrast, the proposed DV-XML framework integrates ResNet-50 for visual features and BERT for textual embedding using an inductive learning-based fusion methodology. This approach addresses the limitations of previous models by ensuring robust representation and retrieval accuracy, while incorporating an innovative RR mechanism to optimize the retrieval process during inference. It utilizes an attention-augmented, transformer-based approach to capture the nuanced semantic and contextual elements of text queries. Concurrently, a deep convolutional framework variant is employed for embedding visual queries with an emphasis on computational efficiency. To support dual-modal query integration, the model incorporates an inductive learning approach that stacks multiple layers of sequential MLPs to develop a unified feature representation by leveraging extracted features from both modalities. Finally, an RR method is applied to optimally sort

combined queries and target images, thereby improving retrieval efficiency.

TABLE I. COMPARATIVE ANALYSIS OF CBIR FRAMEWORKS

Paradigm	Visual	Textual	Fusion	Strength	Weakness
AutoRet [6]	CNN	N/A	Clustering	Efficient retrieval with fixed cluster ranking	Limited selection of attributes for query image
TIRG [7]	CNN	LSTM	Gating	Unified representation	Computation cost
VAL [8]	CNN	RNN	Attention mechanism	Aligns visual-language	Needs multiple text features
DenseBert4Ret [1]	DenseNet	Transformer	Feature fusion	Strong dual-modality representation	Extensive training
DvLIL [2]	ResNet-50	BERT	Inductive learning	Efficient computation	Large-scale retrieval cost
DQU-CIR [15]	BLIP-2	VLPM	MLP	Multimodal integration	Limited real-world evaluation
TG-CIR [16]	VLPM	Text encoder	Transfer learning	Improves retrieval accuracy	Complex multi-stage training
CLIP4CIR [17]	CLIP	CLIP	Combiner network	Strong image-text embedding	Needs large-scale data
DV-XML (proposed)	ResNet-50	BERT	Inductive learning	RR	High fusion demand

IV. METHODOLOGY

In the proposed framework, ResNet-50 serves as a tool for extracting detailed image features, and advanced NLP techniques, specifically BERT, are used to generate text embeddings. The proposed framework consists of four main subsections: (1) image feature extraction, (2) text embedding generation, (3) fusion of textual and visual features, and (4) RR of features and target images for streamlined retrieval.

A. Visual Features Engineering

The proposed model utilizes ResNet-50 as the backbone framework for extracting query image features. In addition to its core components, ResNet incorporates a distinctive feature known as residual blocks or skip connections. These residual block connections serve the purpose of circumventing certain layers, reducing computational complexity, and mitigating information loss.

B. Textual Features Engineering

To capture textual features from the input query, we utilize a transformer-based approach, specifically the pre-trained BERT model [24]. BERT is built on the transformer's encoder architecture, making it highly capable of capturing linguistic meaning, after extensive training on large English corpora, including Wikipedia [25] and the Book Corpus [26], which contain approximately 2,500 million and 800 million words, respectively. BERT is available in two main versions: BERT-base, consisting of 768 feed-forward network modules with 12 encoder layers, and BERT-large, which has 1,024 network modules with 24 encoder layers. Consequently, the output feature vectors of the base and large versions of the BERT model have dimensions of 768 and 1,024, respectively. BERT is known for its computational efficiency and its ability to provide dynamically informed embedding representations.

C. Visual and Textual Embedding Conjugation

Despite the use of conventional techniques to achieve dual-modality composition, we have adopted an inductive learning-based feature integration strategy. Both visual and textual features are processed by MLPs whose weights are fine-tuned based on the characteristics of each individual modality. To achieve this, we sequentially train the MLPs using the features extracted from both the query image and the text. This approach is referred to as inductive learning-based joint feature integration because the behavior of the subsequent MLP layers is influenced by the individual modality features. Inductive learning is a machine learning paradigm where a model learns generalized patterns from observed data and applies them to unseen instances. In our framework, the MLP layers leverage inductive learning to adaptively integrate visual and textual features. During training, the MLP layers sequentially learn feature dependencies by incorporating prior information from each modality before generating a unified representation. This process ensures that the model generalizes well to novel image-text queries, thereby improving retrieval performance. To create a unified representation of these features, we utilize two sequential MLP layers, where image features are fed into the first layer, followed by linguistic features. In our framework, feature fusion is achieved through an inductive learning based MLP network. First, image embeddings (f_i) extracted via ResNet-50 undergo a transformation through a fully connected layer to align their dimensions with the textual embeddings (f_t) extracted via BERT. The transformed features are then passed through sequential MLP layers, where activation functions ensure non-linearity, capturing complex interdependencies between the modalities. This fusion approach outperforms conventional concatenation-based fusion in that it dynamically learns feature relationships rather than simply combining raw embeddings. The role of MLPs (Ω and Φ) is to progressively refine these features into a common representation (J), ensuring robust alignment between visual and textual modalities. To ensure compatibility in feature dimensions across modalities, a fully connected projection layer is added to the output of ResNet-50. This layer is set up to match the output dimensions of the BERT encoder, which are 768 for BERT-base and 1024 for BERT-large, aligning the feature dimensions for effective integration.

D. Reverse Re-Ranking

Following the feature fusion, we implemented an RR mechanism [5, 27] to systematically arrange the common feature embeddings and target images to improve retrieval accuracy. This approach works on the principle that if a combined query (denoted by Q_c) retrieves a target image (denoted by T_i), then T_i should reciprocally prioritize $\mathrm{Q_c}$. The RR process functions as a post-processing technique that integrates elements from the Multivariate Re-rank (MR)

algorithm and the cross-modal re-ranking scheme. Specifically, the method exploits bidirectional retrieval consistency by employing k-reciprocal nearest neighbors to refine rankings. The process involves two stages: 1) arranging the top-K T_i for each Q_c based on similarity measures, and 2) reordering the top-K Q_c for each T_i using reverse retrieval. By incorporating the MR methodology, additional ranking factors such as initial position and similarity confidence are fused to enhance the ranking's robustness. Similarly, the cross-modal re-ranking scheme introduces the concept of combining Image-to-Text (I2T) and Text-to-Image (T2I) retrievals to ensure that retrieval consistency is maintained across modalities. For instance, initial retrieval lists for I2T and T2I are refined by mutual confirmation: if an image ranks a text highly in I2T, the same text should rank the image highly in T2I. This bidirectional re-ranking mechanism results in optimized retrievals, reducing redundancy and improving efficiency. As a result, target image retrieval based on combined queries becomes more accurate and time-efficient without requiring additional training, thus addressing sparsity and scalability challenges. The innovation of the proposed RR algorithm lies in its ability to effectively bridge the gap between the training and inference stages.

E. Mathematical Illustration

Consider a set of p input query images, which can be expressed as $X = [X_1, X_2, \dots, X_p]$, and a set of q input text queries, which can be expressed as $Y = [Y_1, Y_2, \dots, Y_q]$. In line with common practice in supervised learning scenarios, we have a set of target images that the model must retrieve in response to these input queries. These target images are represented by the set $Z = [Z_1, Z_2, \dots, Z_n]$. The operation of the DV-XML paradigm can be represented as follows:

$$E_s = \max_\theta \Delta\,[Q_c, \gamma(Z)] \quad (1)$$

$$Q_c = J(\alpha(X), \beta(Y);\, \theta) \quad (2)$$

In this context, E_s represents the degree of similarity between the merged embeddings from the combined visual and textual user input Q_c and the inherent properties of the target image, denoted as $\gamma(Z)$. The combined modality query is formed by combining the visual and textual query features, as indicated in (2). The function $J(\alpha(\cdot), \beta(\cdot);\, \theta)$ merges both modalities of the user input into a unified representation, while $\gamma(Z)$ represents the visual feature extraction from the visuals of the target dataset. The notation α and β corresponds to the methods employed to extract the respective visual and textual query features, and θ represents the associated parameters for model learning. To generate visual features, the proposed model utilizes the ResNet−50 architecture as its foundation. The architecture considers a user input query of size 224×224, which is subjected to various preprocessing steps. The ResNet−50 model employs a bottleneck design with a combination of max-pooling and average-pooling layers, along with 48 convolutional layers, resulting in a 50-layer deep network. In the proposed framework, a mapping layer is added at the model output to align the dimensions with those of the parallel BERT model.

The generation of visual features for the i-th image is mathematically represented as follows:

$$I_F = \alpha(X_i) \quad (3)$$

The visual features extracted from the input image query, denoted as I_F, have the same dimensions as those of the BERT model. In addition to the extraction of visual features, it is equally important to capture the complex features present in the query text. This responsibility is delegated to the linguistic model, which in our case is the dual model of BERT. BERT is an advanced language model grounded in DL methods, specifically engineered to capture the context and semantics of textual input. A notable advantage of using BERT is its availability as a client-server setup provided by Google, which eliminates the need for local implementation and conserves onboard computational resources. In our implementation, we ran the client-side component on the local development system, and retrieved the textual input query from the server. The model utilized four workers to process both BERT-base/large versions.

Mathematically, the formulation of the text query features can be expressed as follows:

$$T_F = \beta(Y_i) \quad (4)$$

The text query features, represented as T_F, have a dimensionality determined by the BERT model used. In the case of the base model, this dimension is 768, whereas for the large model it goes up to 1024.

After extracting features from both modalities, the next step is to integrate them into a common representation. While the literature offers various techniques to achieve this fusion, such as average pooling, max pooling, concatenation, and gating mechanisms, our approach leverages an influential learning mechanism. This procedure involves the step-by-step training of three layers in an MLP using the non-linear nature of activation functions. The fully connected MLP layers act as projection functions, allowing the transformation of features from distinct modalities into a shared embedding space. The process begins with the training of a textual feature mapping, as shown in (5):

$$\Omega = E[\iota.\, \omega(T_F)] \quad (5)$$

In (5), the symbol E denotes the matrix exponential function, while ι represents the imaginary unit. This equation illustrates the transformation of the vectorized feature map of the textual query into a matrix format, preparing it for integration with the visual features. To perform this transformation, we utilize a mapping function, denoted ω, which takes the textual features T_F as its input. Following this transformation, we procced to another mapping step that communicates the impact of Ω on the visual features, as detailed in (6):

$$\Phi = \Omega.\,[\mu(I_F)] = E[\iota.\, \omega(T_F)].\,[\mu(I_F)] \quad (6)$$

In this context, Φ acts as a unified representation that combines the embeddings of both modalities into a single shared space. Equation (6) illustrates the process of inductive training, where features from one modality influence the other. It first transforms the textual embeddings T_F, resulting in Ω. This Ω then exerts a significant influence on the visual features I_F, ultimately leading to the creation of Φ, the joint

representation that encapsulates both modalities. Both (5) and (6) are essential for achieving a unified representation of both visual and textual features. To evaluate the proposed framework's effectiveness, our objective is for (1) to reach its maximum value. This indicates the highest degree of similarity between the retrieved target photos $\gamma(Z)$ and the composite alignment of the user's input dual modality query Q_c.

In (1), $\gamma(Z)$ symbolizes the features extracted from the retrieved target images, obtained using the ResNet-50 visual model. Since the features of the target images are derived solely from their visual content, it is imperative that the joint representation of the input query be effectively aligned with these features. To facilitate this alignment, we have introduced a mapping function, denoted as ρ, which facilitates the comparison between the target and retrieved image and provides a means to assess the performance of the proposed model. The function ρ consists of three fully connected perceptrons and the joint representation will be input to the function. The expected result is the function ρ will produce an output that matches the desired visual features. While ρ operates solely on the joint representation Φ, the function $\hat{\rho}$ incorporates not only Φ but also T_F and I_F, allowing it to integrate the combined effects of the similarity kernel.

$$J(T_F, I_F; \theta) = \eta_a[\rho(\Phi)] + \eta_b[\hat{\rho}(\Phi, I_F, T_F)] \quad (7)$$

The parameters η_a and η_b have been optimized using heuristic methods to enhance the model's performance. Both η_a and η_b have an impact on the weighting of the joint representation of the retrieved visual features in the shared space.

After successfully combining the textual and visual features, we propose an RR approach. The primary function of RR is to organize the combined queries and corresponding target images in a manner that optimizes the efficiency of the retrieval process. Consider the set of combined features query $Q = [Q_1, Q_2, \dots, Q_n]$ and the set of target image features $T = [T_1, T_2, \dots, T_n]$. The fundamental assumption is that a combined query and a target image can be retrieved from each by Q2T and T2Q retrieval, both forward and backward. In simple terms, the target image that closely matches the combined query should be ranked at the top of the candidate list, and conversely, the combined query should also be ranked high for the target image. We define two re-ranking strategies, R_{Q2T} re-ranking and R_{T2Q} re-ranking based on this assumption.

$$R_{Q2T(Q,k)} = [T_1, T_2, \dots, T_m, \dots, T_k] \quad (8)$$

In (8), $R_{Q2T(Q,k)}$ indicates the ranking of k target images against the dual-modality query Q. Let T_k be the top k nearest neighbors among which T_m is the most similar to the combined query. As of now, we have the ranking position information of the most similar image $P_{T_m} \in (0,1, \dots, M)$. To efficiently utilize this information of the most similar image, we defined a coefficient as shown in (9):

$$C_{Q2T} = e^{\left\{-\epsilon\left\{\left(P_{T_m}+1\right)\right\}\right\}} \quad (9)$$

where ϵ represents the ranking coefficient. The primary goal of (9) is to normalize the ranking data of $Q2T$, where the most similar target image (in this case, the T_m image) would have a higher C_{Q2T} component value.

Starting with the most similar target image, denoted as T_m, we initiate the reverse search and retrieve the combined queries associated with T_m, as demonstrated below.

$$R_{T2Q(T_m,l)} = [Q_1, Q_2, \dots, Q_n, \dots, Q_l] \quad (11)$$

Following the same nomenclature of R_{Q2T}, we have a combined query Q_n that is most similar to the target images T_m. In (10), l indicates the top- l nearest combined queries of T_m. For every target image from 1 to k, l nearest combined queries can be retained. Like C_{Q2T}, we can also obtain the C_{T2Q} using (11).

$$C_{T2Q} = \begin{cases} e^{-\epsilon^{\left(P_{Q_n}+1\right)}}, & n \in R_{T2Q}(T_m, l) \\ 0, & elsewhere \end{cases} \quad (11)$$

The coefficient C_{T2Q} carries the supporting similarity information that can be used to correct C_{Q2T}. When utilizing both coefficients, we introduced a similarity confidence score to measure the level of confidence in the similarity. For a selected target image T_m, the higher the similarity with Q_n, the greater the confidence level. With this consideration, we calculate the similarity confidence score as follows:

$$C_S = \frac{\cos(T_m, Q_n)}{\sum_{n=0}^{N} \cos(T_m, Q_n)} \quad (12)$$

Finally, the weighted sum of these three ranking coefficients provides the similarity after reverse re-ranking which is denoted by S_{RR}.

$$S_{RR} = C_{Q2T} + C_{T2Q} + C_S \quad (13)$$

S_{RR} considers the outcome of the forward, reverse reranking and accumulates it with the confidence similarity.

F. Training Paradigm

The effective training of the model, which involves weight updates based on a loss value, is crucial. In our methodology, we utilize triplet loss [28] to evaluate the similarity between the combined feature set and the target image from the dataset, ensuring that the retrieval process is aligned with the desired objective. As the name suggests, triplet loss relies on three fundamental components:

1. The anchor, which is the combined representation of the visual and linguistic query, denoted as $J(T_F, I_F)$.
2. The positive, which represents the desired target images in the database, denoted as $\gamma(Z)$.
3. The negative, which corresponds to the retrieved images from the database, denoted as $\tilde{\gamma}(Z)$.

Thus, the triplet loss utilized during the training of the proposed framework can be expressed as follows:

$$\frac{1}{N}\sum_{i}^{N} log\left[1 + exp\{\Delta\big(J(T_F, I_F), \tilde{\gamma}(Z)\big) - \Delta\big(J(T_F, I_F), \gamma(Z)\big)\}\right] \quad (14)$$

where N represents the number of retrieved images, Δ denotes the level of similarity, and the number '1' in the outer brackets denotes the margin employed to control the differentiation between positive and negative samples.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Implementation Details

The proposed framework was implemented on a Ubuntu LTS 64-bit operating system, utilizing a system configuration that included 32GB of RAM and an 8-core AMD Ryzen processor. An NVIDIA 3090Ti GPU was employed for computationally intensive tasks. The implementation was carried out in a Python 3.6 virtual environment, incorporating essential libraries such as Torchvision (v0.4.0) and PyTorch (v1.2.0). To enhance the training process over multiple datasets, basic image augmentation techniques were applied. Given that ResNet-50 requires a specific input size, all images were resized to 224×224 pixels. To optimize system the use of system resources, BERT-Client, a library developed by Google, was utilized for text feature extraction. To mitigate potential memory constraints, an additional 75 GB of swap space was allocated to complement the 16 GB of RAM available on the system. The framework was trained with varying epoch values for different datasets. Notably, the model achieved faster convergence on the Fashion-200K dataset, whereas the MIT-States dataset required a longer training period due to its greater diversity. This system configuration ensured smooth execution and effectively prevented memory-related issues during the experiments.

B. Datasets

Table II provides a quantitative summary of the datasets used for both phases of the DV-XML paradigm. These databases contain images of daily life, fashion domain, and are well established in the research community for their suitability in image retrieval tasks. The MIT-States dataset is recognized for its richness, containing images labeled with descriptive tags, including 245 nouns and 115 adjectives. The Fashion-200K dataset [19] is a large collection of approximately 200,000 fashion-related images. Each image is associated with a descriptive text that specifies its key attributes.

TABLE II. DESCRIPTION OF DATASETS

Dataset	MIT-States	Fashion-200K
Visuals	53,753	201,838
Training queries	43,207	172,049
Testing queries	82,732	33,480
Text query avg. length	2	4.8
Per query visuals	26	3

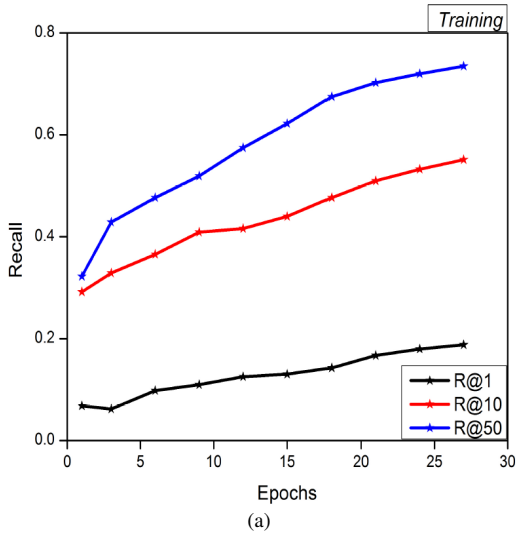
C. Comparative Analysis

As in previous academic studies in the field, the proposed model was evaluated using the metric known as Recall@K. This metric evaluates the performance of the model in retrieving relevant images from a given set of top-K results. In CBIR, Recall@K represents the percentage of visuals that were accurately retrieved in response to a dual-modality query. It's important to note that this retrieval task takes place after the RR phase. The comparison results of the proposed approach with the baseline frameworks, evaluated by Recall@K for both the Fashion-200K and MIT-States datasets, are presented in Tables III and IV, respectively. It is evident that the proposed approach outperforms the baseline frameworks. Figures 3 and 4 illustrate the efficiency of DV-XML on the Fashion-200K and MIT-States datasets, respectively.

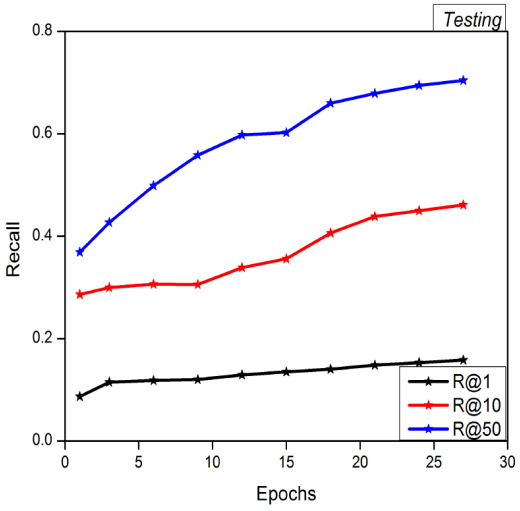
It's important to emphasize that the choice of different K parameters for the two datasets is justified by the differences in the number of images retrieved for a single modality input (query). In the case of the Fashion-200K dataset, where a bi-modal query yields approximately 3 target images, the use of higher K values becomes crucial to ensure a more realistic evaluation of the proposed framework's performance. This larger K value allows to evaluate the performance of the model under more demanding conditions, where it must retrieve a larger number of relevant target images. In contrast, for the MIT-States dataset, where more images are retrieved for a single query, lower K values are sufficient for evaluation purposes.

However, it should be noted that using lower K values could result in retrieving all the target images for a query, even if the model's performance is relatively poor. This scenario wouldn't provide a realistic assessment of the model's effectiveness. Therefore, for the Fashion-200K dataset, we chose to set K to values of 1, 10, and 50. This range of K values allows for a comprehensive evaluation of the model, covering both the retrieval of a single relevant image (K=1) and more extensive retrieval scenarios (K=10 and K=50). Similarly, in the case of the MIT-States dataset, we chose values of K corresponding to 1, 5, and 10. These values are tailored to the characteristics of the dataset and ensure that the DV-XML efficiency is rigorously evaluated across various retrieval scenarios.

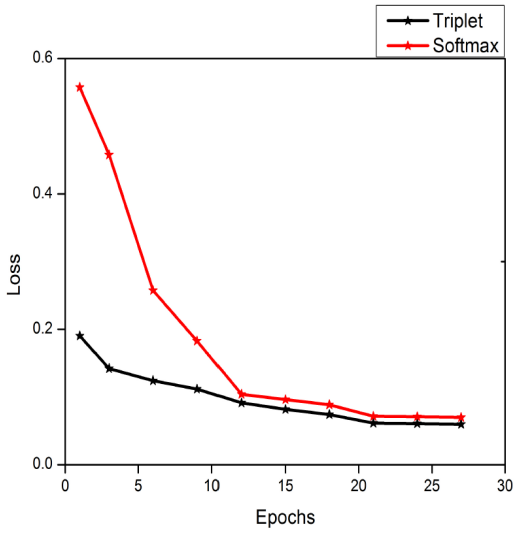
To further analyze the robustness of the proposed framework in practical retrieval scenarios, we categorized queries into three levels based on the degree of modification in the textual descriptions: (1) minor modifications (e.g., slight color changes: 'red dress' → 'dark red dress'), (2) moderate modifications (e.g., adding or removing attributes: 'black jacket' → 'black leather jacket'), and (3) significant modifications (e.g., complex transformations involving multiple attributes: 'white sneakers' → 'blue high-top sneakers with stripes'). Our evaluation reveals that the proposed DV-XML framework performs exceptionally well for minor and moderate modifications, achieving over 85% accuracy at Recall@10. However, for significant modifications, the performance decreases slightly due to the increased semantic complexity. This highlights the need for future improvements in the handling of highly complex composed queries.



(a)

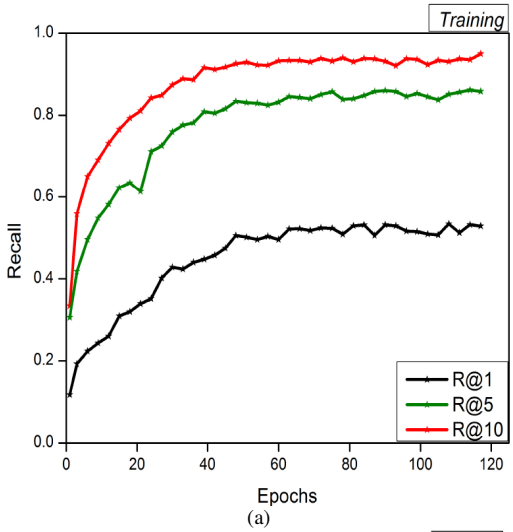


(b)

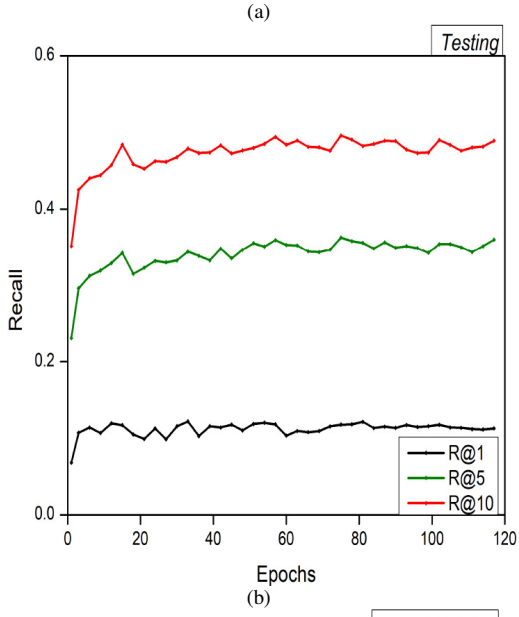


(c)

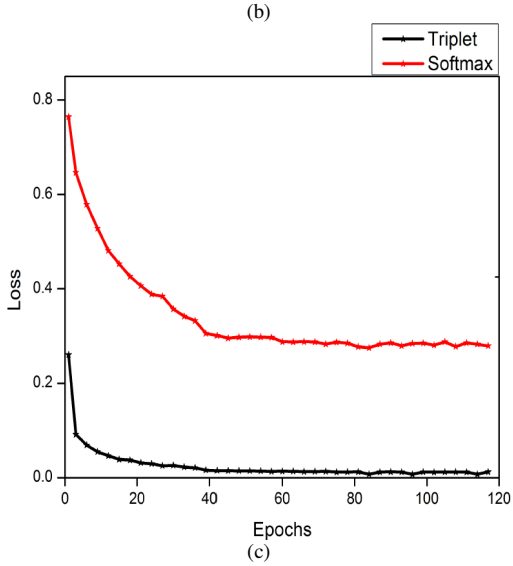
Fig. 3. Efficiency of DV-XML on the Fashion-200K dataset (a): training plots, (b): testing plots, (c): loss.



(a)



(b)



(c)

Fig. 4. Efficiency of DV-XML on the MIT-States dataset (a): training plots, (b): testing plots, (c): loss.

D. BERT-Large / BERT-Base Performance Evaluation

We retrieved linguistic components from the proposed framework by using the base and large versions of the BERT framework to comprehensively evaluate it. We then performed an evaluation of the proposed framework using both variants of the BERT-base/large model, as described above. These two variants differ in the density of the generated text embedding. It's important to note that text queries may vary in terms of word or character count, but the embedding dimensions remain consistent for each variant of BERT. Specifically, the BERT-base generates embeddings with dimensions of 768, whereas the BERT-large results in embeddings with dimensions of 1024. This discrepancy in embedding dimensions leads to an inverse relationship between text query length and BERT embedding density. Consequently, the proposed methodology is less effective with the BERT-large variant compared to the BERT-base variant. The BERT-base variant generates denser text query embeddings, which improves the overall efficiency of DV-XML. We have designated the textual features generated by BERT-base as DV-XML(B) and those generated by BERT-large as DV-XML(L). Tables III and IV clearly illustrate the improved performance of the proposed model when using BERT-base, as indicated by the notably lower recall values for DV-XML(L) compared to DV-XML(B).

TABLE III. QUALITATIVE RESULTS FOR FASHION-200K DATASET

Paradigm	Recall@1	Recall@10	Recall@50
Show & tell [29]	12.3 ± 1.1	40.2 ± 1.7	61.8 ± 0.9
Param. hashing [30]	12.2 ± 1.1	40.0 ± 1.1	61.7 ± 0.8
Relationship [31]	13.0 ± 0.6	40.5 ± 0.7	62.4 ± 0.6
FiLM [32]	12.9 ± 0.7	39.5 ± 2.1	61.9 ± 1.9
TiRG [27]	14.1 ± 0.6	42.5 ± 0.7	63.8 ± 0.8
Aug TiRG [10]	14.2	41.9	63.3
ComposeAE [33]	16.5	44.2	63.1
DCNet [34]	-	46.9	67.6
DenseBert4Ret(B) [1]	14.9 ± 0.2	43.8 ± 0.3	67.3 ± 0.7
DenseBert4Ret [1]	13.7 ± 0.3	42.1 ± 0.4	67.6 ± 0.7
DvLiL [2]	14.7 ± 0.4	44.2 ± 0.6	69.3 ± 0.2
DV-XML(B)	15.8 ± 0.5	46.1 ± 0.2	70.4 ± 0.4
DV-XML(L)	14.9 ± 0.5	44.8 ± 0.2	68.3 ± 0.4

TABLE IV. QUALITATIVE RESULTS FOR MIT-STATES DATASET

Paradigm	Recall@1	Recall@10	Recall@50
Show & tell [29]	11.9 ± 0.1	31.0 ± 0.5	42.0 ± 0.8
Param. hashing [30]	8.8 ± 0.1	27.3 ± 0.3	39.1 ± 0.3
Relationship [31]	12.3 ± 0.5	31.9 ± 0.7	42.9 ± 0.9
FiLM [32]	10.1 ± 0.3	27.7 ± 0.7	38.3 ± 0.7
TiRG [27]	12.2 ± 0.4	31.9 ± 0.3	43.1 ± 0.3
Aug TiRG [10]	12.3 ± 0.6	31.8 ± 0.3	42.6 ± 0.8
DenseBert4Ret(B) [1]	13.9 ± 0.3	36.8 ± 0.1	49.8 ± 0.3
DenseBert4Ret [1]	12.8 ± 0.2	35.7 ± 0.3	49.2 ± 0.3
DvLiL [2]	13.8 ± 0.4	35.7 ± 0.4	48.9 ± 0.3
DV-XML(B)	14.6 ± 0.4	37.4 ± 0.7	50.3 ± 0.5
DV-XML(L)	13.5 ± 0.4	36.4 ± 0.3	49.6 ± 0.2

VI. CONCLUSION AND FUTURE WORK

This study presents a novel framework, Deep-View X-Modalities Visio-Linguistics (DV-XML), for efficient Content-Based Image Retrieval (CBIR) by seamlessly integrating visual and linguistic modalities. By leveraging ResNet-50 for extracting enriched visual features and Bidirectional Encoder Representations from Transformers (BERT) for capturing semantic and contextual textual features, the proposed framework addresses critical challenges in dual-modality image retrieval systems. The integration of inductive learning-based fusion via Multi-Layer Perceptrons (MLPs) ensures robust unified representations of dual-modality inputs. Additionally, the innovative Reverse Re-Ranking (RR) methodology significantly improves retrieval accuracy and efficiency by optimizing the alignment of combined representations with target images during inference. Extensive evaluations on benchmark datasets, including Fashion-200K and MIT-States, demonstrate the superior performance of the proposed framework compared to existing methods, particularly for complex queries require both visual and textual inputs.

Comparative analysis with previous frameworks further validates the effectiveness of DV-XML. The model achieves Recall@10 scores of 46.1% on Fashion-200K and 37.4% on MIT-States, surpassing TIRG (42.5% and 31.9%), DenseBert4Ret (42.1% and 35.7%), and DvLIL (44.2% and 35.7%). These improvements highlight the advantages of our inductive learning-based fusion over traditional concatenation and gating mechanisms used in previous works. Furthermore, the incorporation of RR significantly enhances ranking consistency, resulting in better retrieval precision compared to models employing standard feature fusion techniques. Despite its strengths, the DV-XML framework has certain limitations that pave the way for future research. For instance, while the current implementation of ResNet-50 effectively captures visual features, recent advancements in vision transformers and encoder-decoder architectures offer promising avenues for further improving visual feature extraction. Similarly, the exploration of alternative linguistic models, such as Large Language Models (LLMs) with cross-modal capabilities, could further enhance the textual feature extraction and fusion processes. While the proposed DV-XML framework demonstrates significant improvements in multimodal image retrieval, several areas remain open for future exploration. One key direction is enhancing the scalability of the model for large-scale datasets. This can be achieved by exploring efficient indexing mechanisms, Approximate Nearest Neighbor (ANN) search techniques, and distributed computing frameworks to efficiently handle high-dimensional embedding.

In addition, improving computational efficiency is essential for real-world adoption. Future work will focus on optimizing the MLP-based fusion mechanism by investigating lightweight architectures, knowledge distillation techniques, and quantization methods to reduce inference time without compromising retrieval accuracy. Another important aspect is real-time application deployment. To make the proposed system viable for real-time scenarios, we aim to integrate edge computing and cloud-based acceleration to ensure low-latency

retrieval in dynamic environments. Furthermore, extending the RR strategy to real-time personalized retrieval systems could significantly improve the user experience. By addressing these challenges, future iterations of the DV-XML framework will aim to establish a scalable, computationally efficient, and real-time image retrieval system suitable for industrial and large-scale applications.

REFERENCES

[1] Z. Khan, B. Latif, J. Kim, H. K. Kim, and M. Jeon, "DenseBert4Ret: Deep bi-modal for image retrieval," *Information Sciences*, vol. 612, pp. 1171–1186, Oct. 2022, https://doi.org/10.1016/j.ins.2022.08.119.

[2] I. Ahmed, N. Iltaf, Z. Khan, and U. Zia, "Deep-view linguistic and inductive learning (DvLIL) based framework for Image Retrieval," *Information Sciences*, vol. 649, Nov. 2023, Art. no. 119641, https://doi.org/10.1016/j.ins.2023.119641.

[3] I. Ahmed, N. Iltaf, R. Latif, N. S. M. Jamail, and Z. Khan, "Dual Modality Reverse Reranking (DM-RR) Based Image Retrieval Framework," *IEEE Open Journal of the Industrial Electronics Society*, vol. 5, pp. 886–897, 2024, https://doi.org/10.1109/OJIES.2024.3435956.

[4] I. Ahmed, Z. Khan, Z. Khan, N. Iltaf, and M. Jeon, "Advanced Multi-Model Deep Learning Approach for Content-Based Image Retrieval," in *Proceedings of the 2023 Korean Software Conference of the Korean Society of Information Scientists and Engineers*, Busan, South Korea, 2023, pp. 736–738.

[5] Z. Yuan *et al.*, "Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022, https://doi.org/10.1109/TGRS.2022.3163706.

[6] M. M. Monowar, M. A. Hamid, A. Q. Ohi, M. O. Alassafi, and M. F. Mridha, "AutoRet: A Self-Supervised Spatial Recurrent Network for Content-Based Image Retrieval," *Sensors*, vol. 22, no. 6, Mar. 2022, Art. no. 2188, https://doi.org/10.3390/s22062188.

[7] N. Vo *et al.*, "Composing Text and Image for Image Retrieval - an Empirical Odyssey," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6432–6441, https://doi.org/10.1109/CVPR.2019.00660.

[8] Y. Chen, S. Gong, and L. Bazzani, "Image Search With Text Feedback by Visiolinguistic Attention Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 2998–3008, https://doi.org/10.1109/CVPR42600.2020.00307.

[9] T. Nagarajan and K. Grauman, "Attributes as Operators: Factorizing Unseen Attribute-Object Compositions," in *15th European Conference on Computer Vision, ECCV 2018*, Munich, Germany, 2018, pp. 172–190, https://doi.org/10.1007/978-3-030-01246-5_11.

[10] M. Aboali, I. Elmaddah, and H. E.-D. Hassan, "Augmented TIRG for CBIR Using Combined Text and Image Features," in *2021 International Conference on Electrical, Computer and Energy Technologies*, Cape Town, South Africa, 2021, pp. 1–6, https://doi.org/10.1109/ICECET52533.2021.9698617.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks, " *Communications of the ACM,* vol. 60, no. 6, pp. 84–90, May 2017, https://doi.org/10.1145/3065386.

[12] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *13th European Conference on Computer Vision, ECCV 2014*, Zurich, Switzerland, 2014, pp. 818–833, https://doi.org/10.1007/978-3-319-10590-1_53.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015, https://doi.org/10.48550/arXiv.1409.1556.

[14] S. Merugu, R. Yadav, V. Pathi, and H. R. Perianayagam, "Identification and Improvement of Image Similarity using Autoencoder," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15541–15546, Aug. 2024, https://doi.org/10.48084/etasr.7548.

[15] H. Wen, X. Song, X. Chen, Y. Wei, L. Nie, and T.-S. Chua, "Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington, DC, USA, 2024, pp. 229–239, https://doi.org/10.1145/3626772.3657727.

[16] H. Wen, X. Zhang, X. Song, Y. Wei, and L. Nie, "Target-Guided Composed Image Retrieval," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, Canada, 2023, pp. 915–923, https://doi.org/10.1145/3581783.3611817.

[17] A. Baldrati, M. Bertini, T. Uricchio, and A. del Bimbo, "Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features." arXiv, Aug. 22, 2023, https://doi.org/10.48550/arXiv.2308.11485.

[18] H. Wu *et al.*, "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 11302–11312, https://doi.org/10.1109/CVPR46437.2021.01115.

[19] X. Han *et al.*, "Automatic Spatially-Aware Fashion Concept Discovery," in *2017 IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 1472–1480, https://doi.org/10.1109/ICCV.2017.163.

[20] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models," in *2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 2105–2114, https://doi.org/10.1109/ICCV48922.2021.00213.

[21] M. Haq and M. A. R. Khan, "DNNBoT: Deep Neural Network-Based Botnet Detection and Classification," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1729–1750, 2021, https://doi.org/10.32604/cmc.2022.020938.

[22] M. Suresh, A. S. Shaik, B. Premalatha, V. A. Narayana, and G. Ghinea, "Intelligent & Smart Navigation System for Visually Impaired Friends," in *12th International Advanced Computing Conference, IACC 2022, Part I*, Hyderabad, India, 2022, pp. 374–383, https://doi.org/10.1007/978-3-031-35641-4_30.

[23] S. Merugu, A. Tiwari, and S. K. Sharma, "Spatial–Spectral Image Classification with Edge Preserving Method," *Journal of the Indian Society of Remote Sensing*, vol. 49, no. 3, pp. 703–711, Mar. 2021, https://doi.org/10.1007/s12524-020-01265-7.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019, https://doi.org/10.48550/arXiv.1810.04805.

[25] I. Annamoradnejad and G. Zoghi, "ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor," *Expert Systems with Applications*, vol. 249, no. B, Sep. 2024, Art. no. 123685, https://doi.org/10.1016/j.eswa.2024.123685.

[26] Y. Zhu *et al.*, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 19–27, https://doi.org/10.1109/ICCV.2015.11.

[27] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching Images and Text with Multi-modal Tensor Fusion and Re-ranking." arXiv, Jul. 29, 2020, https://doi.org/10.48550/arXiv.1908.04011.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 815–823, https://doi.org/10.1109/CVPR.2015.7298682.

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3156–3164, https://doi.org/10.1109/CVPR.2015.7298935.

[30] H. Noh, P. H. Seo, and B. Han, "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 30–38, https://doi.org/10.1109/CVPR.2016.11.

[31] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4974–4983.

[32] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 3942–3951, Apr. 2018, https://doi.org/10.1609/aaai.v32i1.11671.

[33] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber, "Compositional Learning of Image-Text Query for Image Retrieval," in *2021 IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2021, pp. 1139–1148, https://doi.org/10.1109/WACV48630.2021.00118.

[34] Y. Tian, S. Newsam, and K. Boakye, "Fashion Image Retrieval with Text Feedback by Additive Attention Compositional Learning," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023, pp. 1011–1021, https://doi.org/10.1109/WACV56688.2023.00107.