Visual Image Retrieval Based on Multimodal Information Fusion

Hemachandran K*

AI Research Centre, School of Business, Woxsen University, India; hemachandran.k@woxsen.edu.in

*Corresponding Author: hemachandran.k@woxsen.edu.in

DOI: https://doi.org/10.30212/JITI.202503.015

Submitted: Aug. 25, 2025 Accepted: Oct. 06, 2025

ABSTRACT

This study proposes a multimodal information fusion approach for visual image retrieval. The model comprises three core components: a multimodal feature extraction module (MFEM), a multimodal feature fusion module (MFFM), and a unified feature retrieval module (UFRM) that process and integrate input data from different modalities. We design a Transformer-based multimodal fusion framework that combines image and text features through multi-head self-attention and cross-modal attention mechanisms, enabling joint feature representations with enhanced expressiveness and precision. Unlike existing methods that rely on simple concatenation or weighted fusion, the proposed approach learns fine-grained inter-modal interactions, thereby improving retrieval accuracy. Experimental evaluations on three public benchmarks—FashionIQ, CIRR, and Fashion200K—show that the proposed method outperforms current state-of-the-art approaches across multiple metrics. The method exhibits robust performance in both accuracy and generalization across diverse retrieval scenarios, confirming its effectiveness for complex image retrieval tasks.

Keywords: Multimodal information fusion, Visual image retrieval, Feature extraction, Transformer model, Retrieval performance optimization

1. Introduction

Visual image retrieval [1] has gained increasing significance in contemporary information systems, with applications spanning e-commerce [2], medical image analysis [3], video surveillance [4], and digital libraries [5]. Traditional retrieval approaches relying on single modalities—either images or text—prove insufficient for addressing the growing demands for precision and effectiveness. These methods encounter limitations when processing complex user queries that require relating and integrating multiple information sources, such as combining product images with textual descriptions of desired modifications. In scenarios where consumers specify style preferences alongside reference images, or clinicians integrate radiological scans with patient histories, single-modality systems fail to capture user intent adequately, necessitating multimodal fusion approaches.

Previous research has explored retrieval enhancement through low-level visual features, including color histograms, textures, and shape descriptors [6,7,8]. While effective for straightforward queries, these approaches struggle with tasks requiring advanced semantic

interpretation. With the advent of deep learning, CNN-based methods such as AlexNet, VGGNet, and ResNet have achieved substantial improvements in extracting high-level visual features. However, these methods remain single-modal and lack mechanisms for aligning visual data with textual descriptions, thereby limiting their capacity to capture the semantic richness of user queries.

Subsequent work has explored multimodal fusion through concatenation or weighted aggregation of image and text features [9,10,11,12,21]. However, these strategies lack deep semantic interaction between modalities, resulting in shallow representations that inadequately capture fine [7] grained relationships. Recent approaches employing cross-modal attention mechanisms, generative adversarial networks (GANs) [8], and cross-modal contrastive learning have advanced the field but remain constrained by computational complexity, substantial data requirements, and limited sensitivity to subtle query variations.

This study addresses these limitations by proposing a Transformer-based multimodal information fusion framework that integrates rich, fine-grained interactions between image and text modalities. The model leverages pre-trained CNNs for visual feature extraction and BERT for text encoding, with integration achieved through multi-head self-attention and cross-modal attention layers. Unlike previous approaches, our method ensures that elements from different modalities interact across multiple levels, yielding joint representations that better reflect semantic intent. Additionally, metric learning based on combined contrastive and triplet losses enhances retrieval precision and robustness.

This study demonstrates that fine-grained multimodal fusion enabled by attention-based architecture achieves substantial improvements in retrieval performance. Through comprehensive experiments on FashionIQ, CIRR, and Fashion200K datasets, we show that the proposed approach consistently outperforms state-of-the-art models across accuracy, recall, and robust metrics, establishing its viability for real-world multimodal retrieval applications.

2. Related Work

2.1 Transformer-Based Multimodal Information Fusion Model

Recent work has explored Transformer architecture for multimodal information fusion across various vision-language tasks.

TransVG [9] introduces an end-to-end visual grounding framework that leverages the Transformer architecture to integrate image and text information without requiring region proposals. By directly aligning natural language descriptions with image regions, the model addresses complex vision-language tasks with enhanced spatial-semantic correspondence. ViLT [10] presents a vision-language Transformer that eliminates the need for convolutional feature extractors or region supervision. Unlike conventional multimodal models that depend on convolutional neural networks for image encoding, ViLT processes visual and textual inputs directly within the Transformer framework, reducing computational overhead while accelerating inference and decreasing hardware requirements. The model achieves competitive performance across multiple multimodal benchmarks.

Pixel-BERT [11] employs a deep multimodal Transformer for pixel-level alignment between visual and textual features. This pixel-wise matching mechanism enables fine-grained vision-

language correspondence, yielding robust cross-modal generalization across diverse tasks and datasets. MDETR [12] presents a modulated detection framework for multimodal reasoning, utilizing Transformer-based modulation to align image and text representations. The model demonstrates effectiveness in visual grounding and object detection tasks by accurately matching textual descriptions with corresponding image regions.

Despite the efficacy of Transformer-based architectures in multimodal fusion, their substantial computational requirements and dependence on large-scale training data constrain deployment in resource-limited settings. Additionally, current approaches exhibit limitations in capturing fine-grained cross-modal interactions, indicating opportunities for further refinement.

2.2 Pre-Trained Models and Contrastive Learning

Recent advances in vision-language pre-training have explored various architectural and training strategies for cross-modal representation learning.

Oscar [13] introduces object-semantic alignment during pre-training to enhance vision-language task performance. By incorporating object-level semantic information, the model establishes tighter correspondence between visual and linguistic features, yielding improvements in image captioning, visual question answering, and image retrieval benchmarks. CLIP [14] employs natural language supervision to learn transferable visual representations from large-scale image-text pairs. The model achieves cross-task generalization by learning visual concepts directly from textual descriptions, enabling zero-shot transfer across diverse vision-language applications.

Zaid et al. [15] investigate vision-language model scaling through noisy text supervision, leveraging large-scale weakly-labeled data to learn broader feature representations while maintaining task efficiency. ImageBERT [16] performs cross-modal pre-training on weakly supervised image-text datasets, integrating visual and textual features through joint representation learning. The weakly-supervised pre-training strategy enables feature extraction from unlabeled or partially labeled data, improving downstream task performance. VisualBERT [17] establishes a streamlined vision-language baseline through joint pre-training and task-specific fine-tuning, demonstrating competitive performance across multiple benchmarks.

Despite these advances, real-time processing efficiency remains a challenge requiring further investigation.

2.3 Based on Graph Neural Networks and Other Models

Recent research has investigated graph neural networks and complementary architectures for multimodal information fusion in visual image retrieval through structured feature representation and alignment strategies.

Li et al. [18] introduce a visual semantic reasoning framework for image-text matching that integrates visual and semantic information through a reasoning module designed to resolve complex semantic descriptions in cross-modal retrieval. Surbhi et al. [19] propose latent semantic scaling for image-text matching by aligning visual and textual representations within a shared latent space, enhancing visual search performance. Misra et al. [20] employ graph neural networks for multimodal retrieval by constructing graph-structured representations where image and textual data serve as nodes connected through cross-modal edges, enabling structured information fusion across modalities.

VSE++ [21] refines visual-semantic embedding by incorporating hard negative mining during training, which enhances discrimination between positive and negative samples in the embedding space to improve retrieval precision. T2VLAD [22] leverages Vector of Locally Aggregated Descriptors (VLAD) to align global video features with local textual features, strengthening cross-modal retrieval robustness.

However, the computational complexity of these graph-based and embedding approaches constrains scalability in large-scale and real-time deployment scenarios. Furthermore, modeling fine-grained cross-modal interactions remains an open challenge.

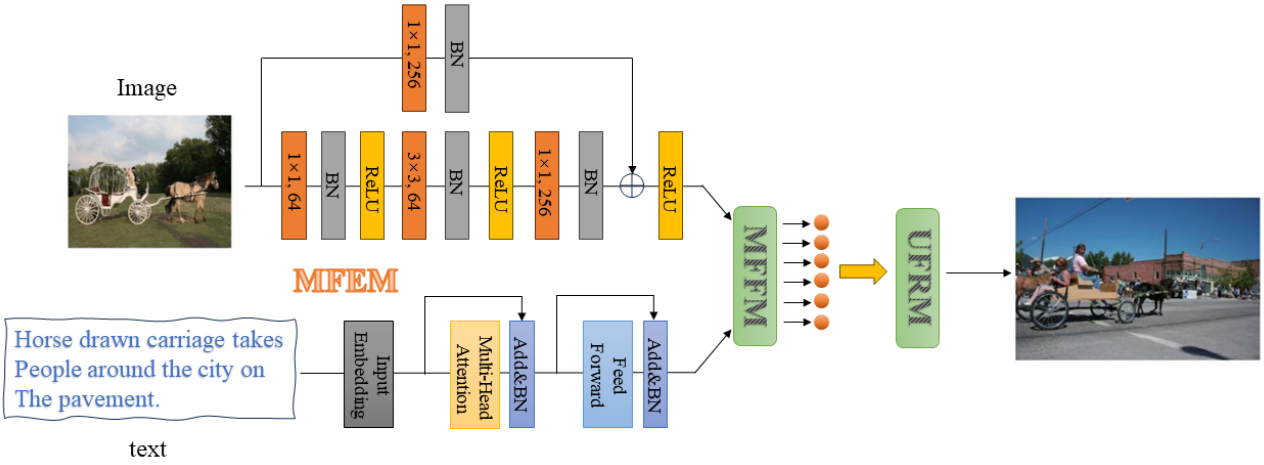

Figure 1. The overall structure of our network. The image and text modalities are extracted by the feature extractor and further fused by the multimodal fusion network to obtain our results.

3. Method

In this study, we proposed an innovative visual image retrieval method based on multimodal information fusion.

3.1 Multimodal Feature Extraction Module

We use the pre-trained ResNet-50 [24] model to extract image features. ResNet-50 is a deep convolutional neural network that solves the gradient vanishing problem.

Given an input image I, we represent it as a pixel matrix $I \in R^{H\times W\times C}$. After processing by the convolutional layers, pooling layers, and fully connected layers of the ResNet-50 model, we obtain a high-level feature representation of the image $F_I \in R^d$,, where d represents the feature dimension.

$$F_I = ResNet - 50(I) \quad \text{Formular 1}$$

Specifically, ResNet-50 contains multiple convolutional layers, pooling layers, and residual blocks. Its core calculation process can be expressed as:

$$F_{l+1} = F_l + F(F_l, W_l) \quad \text{Formular 2}$$

Where F_l is the feature representation of the lth layer, F represents the feature transformation after the convolution operation, and W_l is the weight matrix of the lth layer.

Through this residual learning framework, we can extract the high-level feature representation of the image. In addition, to further improve the expressiveness of image features, we also introduced

global average pooling and batch normalization [25] techniques when extracting features. These techniques help reduce overfitting and accelerate model convergence, thereby obtaining a more robust feature representation.

The role of global average pooling is to average all pixel values. The formula is as follows:

$$F_{GAP} = \frac{1}{H \times W} \sum_{i=1}^{H} \quad \sum_{j=1}^{W} \quad F_{i,j} \qquad \text{Formular 3}$$

The role of batch normalization is to normalize the input of each layer so that its mean is 0 and its variance is 1. The formula is as follows:

$$F_{BN} = \frac{F-\mu}{\sqrt{\sigma^2+\epsilon}} \qquad \text{Formular 4}$$

Where μ and σ represent the mean and standard deviation of the features of the current batch, respectively, and ϵ is a small constant used to avoid the denominator being zero.

Text Feature Extraction: We use the pre-trained BERT model. BERT is a bidirectional Transformer [26] model that captures contextual information through large-scale text pre-training.

Given the input text T, we represent it as a word sequence {t1, t2, …, tn} where n represents the

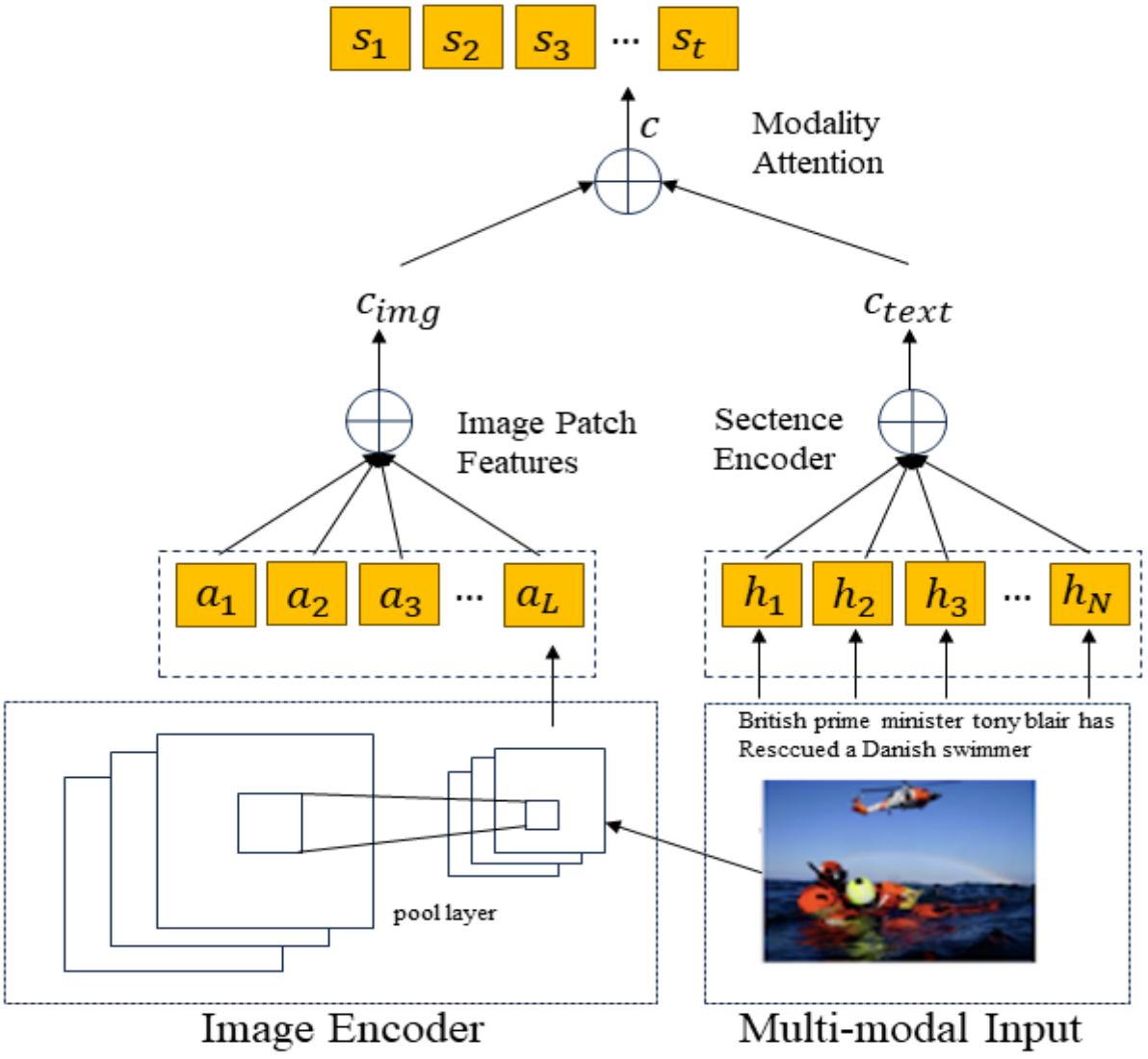

Figure 2. The structure diagram of our image and text modality fusion mechanism.

After processing the encoding layer of the BERT model, we obtain a high-level feature representation of the text $F_T \in R^{n \times d}$, where d represents the feature dimension.

$$F_T = BERT(T) \quad \text{Formular 5}$$

The BERT model processes input text through a multi-layer Transformer encoder. Each encoder layer contains a multi-head self-attention mechanism and a feedforward neural network. The calculation process of the self-attention mechanism is as follows:

$$Q = XW_Q, K = XW_K, V = XW_V \quad \text{Formular 6}$$

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad \text{Formular 7}$$

$$O = AV \quad \text{Formular 8}$$

Where X represents the input features, W_Q, W_K, and W_V are the weight matrices for the query, key, and value, respectively, and d_k represents the dimension of the key. Through a multi-layer self-attention mechanism, BERT is able to capture the complex relationships between words in a text, thereby generating high-quality text feature representations.

Subword tokenization specifically breaks words into smaller subword units, improving the model's ability to handle unknown vocabulary. For example, the word "unhappiness" can be split into "un" and "happiness." Positional encoding incorporates positional information. The formula is as follows:

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{2i/d}}\right) \quad \text{Formular 9}$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{2i/d}}\right) \quad \text{Formular 10}$$

Among them, pos represents the position, i represents the index of the feature dimension, and d represents the feature dimension.

3.2 Multimodal Feature Fusion Module

We employ a Transformer-based multimodal fusion architecture, implementing feature fusion through self-attention and cross-modal attention mechanisms.

Given the input features, we first compute the query, key, and value matrices Q, K, and V: Next, we compute the attention weight matrix A:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad \text{Formular 11}$$

Where d_k is the dimension of the key. Finally, calculate the self-attention output:

$$O = AV \quad \text{Formular 12}$$

The self-attention mechanism calculates the similarity between input features, thereby weighting and summing important information, achieving information fusion and interaction within features.

In our architecture, each multi-head self-attention layer includes multiple independent attention heads, each of which independently calculates attention weights and output features. These are then concatenated and integrated through a linear transformation layer:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W_O \quad \text{Formular 13}$$

Where h represents the number of attention heads, and W_O is the weight matrix used for the linear transformation.

The purpose of the cross-modal attention mechanism is to facilitate information exchange and

fusion between features from different modalities. Given image features F_I and text features F_T:

$$Q_I = F_I W_{Q_I}, K_T = F_T W_{K_T}, V_T = F_T W_{V_T} \quad \text{Formular 14}$$

Where W_{V_T} is the weight matrices for text key and value respectively. Next, calculate the cross-modal attention weight matrix:

$$A_{IT} = softmax\left(\frac{Q_I K_T^T}{\sqrt{d_k}}\right) \quad \text{Formular 15}$$

Finally, the cross-modal attention output is calculated:

$$O_{IT} = A_{IT} V_T \quad \text{Formular 16}$$

The cross-modal attention mechanism computes similar scores between image and text feature representations, weighting and integrating textual information into visual features to construct joint embeddings.

Textual descriptions encode semantic details that may be ambiguous in visual data, while images provide spatial and structural information absent from text alone. The integration of these modality-specific representations yields joint features with enhanced discriminative capacity.

Multi-Level Fusion: Feature fusion is extended through multi-level attention mechanisms within the Transformer architecture, enabling hierarchical interaction between visual and textual representations. Multi-head self-attention operates within each modality to capture intra-modal dependencies, while multi-head cross-modal attention facilitates inter-modal alignment across multiple projection subspaces, extracting features at varying semantic granularities.

Given image features F_I and text features F_T, we first fuse the features within each modality:

$$O_I = MultiHead(Q_I, K_I, V_I), O_T = MultiHead(Q_T, K_T, V_T) \quad \text{Formula 17}$$

Next, the features is fused through the multi-head cross-modal attention mechanism:

$$O_{IT} = MultiHead(Q_I, K_T, V_T) \quad \text{Formula 18}$$

Finally, we concatenate and transform the fused features to generate a joint feature representation:

$$F_{fused} = Concat(O_I, O_{IT}) \quad \text{Formula 19}$$

$$F_{final} = F_{fused} W_{final} \quad \text{Formula 20}$$

$W_{final}$is the weight matrix used to transform the joint feature representation.

Through a multi-level attention mechanism and cross-modal feature fusion, we are able to generate high-quality joint feature representations, providing a solid foundation for subsequent retrieval.

3.3 Joint Feature Retrieval Module

We employ multiple similarity metrics and efficient retrieval techniques to ensure accurate and fast retrieval on large datasets.

Given the fused joint feature representation Ffinal, we first normalize it:

$$F_{norm} = \frac{F_{final}}{\|F_{final}\|} \quad \text{Formula 21}$$

Next, given the query image features Fquery and the database image features Fdb, we calculate the cosine similarity between them:

$$s = cos(F_{query}, F_{db}) = \frac{F_{query} \cdot F_{db}}{\|F_{query}\| \|F_{db}\|} \quad \text{Formula 22}$$

By calculating the cosine similarity between a query image and all images in the database, we can find the images most similar to the query image.

Nearest Neighbor Search: To improve retrieval efficiency, we employ a nearest neighbor search algorithm. Specifically, we utilize techniques such as a k-d tree (k-dimensional tree) and LSH (Locality-Sensitive Hashing) for fast nearest neighbor searches.

Given the query image feature Fquery and the database image feature set , we first build a k-d tree or LSH index structure, and then perform a fast nearest neighbor search through the index structure:

$$\{F_{nn_1}, F_{nn_2}, \dots, F_{nn_k}\} = NearestNeighbors(F_{query}, k) \quad \text{Formula 23}$$

Where k is the number of nearest neighbors.

Comprehensive Similarity Calculation: To further improve search accuracy, we combine multiple similarity metrics. In addition to cosine similarity, we can also combine other similarity metrics such as Euclidean distance and Hamming distance:

$$d_{euclidean} = \| F_{query} - F_{db} \|_2 \quad \text{Formula 24}$$

$$d_{hamming} = \sum_{i=1}^{d} |F_{query_i} - F_{db_i}| \quad \text{Formula 25}$$

Finally, we use the weighted summation method to combine the results of multiple similarity measurement methods to obtain a more accurate similarity evaluation:

$$s_{final} = \lambda_{cosine} \cdots_{cosine} - \lambda_{euclidean} \cdot d_{euclidean} - \lambda_{hamming} \cdot d_{hamming} \quad \text{Formula 26}$$

4. Experiment

4.1 Experimental Setup

4.1.1 FashionIQ dataset

FashionIQ constitutes a benchmark for evaluating composed image retrieval models that utilize natural language descriptions. The dataset comprises fashion images across multiple clothing categories paired with human-annotated relative captions, providing infrastructure for multimodal fusion research. Visual content spans diverse garment styles, designs, colors, and patterns, ensuring comprehensive representation of fashion-domain characteristics.

Human annotators supply textual descriptions specifying garment attributes including color, style, pattern, and material composition. These annotations exhibit semantic diversity necessary for models to learn nuanced linguistic variations. The dataset structure employs image-text-image triplets where a reference image, modification text, and target image form the training and evaluation framework. This compositional format enables models to learn correspondence between visual features and linguistic modifications.

FashionIQ's annotation quality and standardized evaluation protocols establish its utility for advancing composed retrieval architectures and multimodal alignment techniques.

4.1.2 CIRR dataset

The Composed Image Retrieval on Real-world images (CIRR) dataset extends compositional retrieval to general visual domains beyond fashion. Each query comprises a reference image and modification instruction, simulating real-world search scenarios such as "find objects similar to this reference but with different colors".

Image content encompasses diverse domains including human subjects, wildlife, landscapes, urban environments, and interior scenes. Query formulations capture fine-grained semantic variations requiring models to parse complex linguistic instructions and identify subtle visual differences. Human-generated modification texts describe both simple attribute transformations (color, shape) and complex scene alterations or object interactions.

CIRR's compositional query structure provides data infrastructure for personalized retrieval systems requiring interpretation of relative rather than absolute descriptions.

4.1.3 Fashion200K dataset

Fashion200K aggregates over 200,000 garment images with detailed textual metadata across categories including dresses, trousers, tops, outerwear, and footwear. Each image receives multiple attribute labels and human-written descriptions specifying color, style, pattern, and material properties. This annotation scheme yields rich visual-linguistic feature sets supporting multimodal retrieval research.

The dataset facilitates development of retrieval architectures that process natural language queries and supports personalized recommendation systems interpreting user preferences expressed through textual descriptions. Fashion200K serves as infrastructure for investigating multimodal fusion techniques in fashion-domain applications.

4.1.4 Evaluation protocols

FashionIQ Evaluation: The dataset employs dual evaluation protocols using the original candidate pool and the VAL-refined candidate set. The VAL method eliminates redundant images and unifies reference-target pairs, reducing candidate pool size and computational overhead while maintaining evaluation validity.

CIRR Evaluation: Beyond standard R@K metrics, CIRR introduces RSubset@K, which evaluates retrieval performance on subsets containing visually similar negative samples. This metric assesses model capacity to discriminate fine-grained differences under challenging conditions.

These complementary metrics provide standardized frameworks for comparing model architectures and quantifying performance across retrieval scenarios.

4.1.5 Implementation configuration

Visual Encoding: ResNet-50 extracts hierarchical image representations, with features from layers 4, 10, and 12 providing multi-scale visual information. Residual connections within each block preserve gradient flow during backpropagation while capturing high-level semantic features.

Text Encoding: BERT-base-uncased generates contextual text embeddings through its bidirectional Transformer architecture, encoding modification instructions as token sequences.

Multimodal Fusion: A cross-modal Transformer comprising 4 layers with 8-head attention mechanisms fuses visual and textual representations into unified embeddings.

Training Regime: Optimization employs AdamW with initial learning rate 3×10^{-4} across 150 epochs. Learning rate decay to 0.1× occurs at epoch 75 to stabilize convergence.

Retrieval Mechanism: The Unified Feature Retrieval Module (UFRM) computes cosine similarity between query and candidate embeddings. Incorporating multiple distance metrics (Euclidean, Hamming) enhances retrieval robustness.

This configuration enables efficient processing of compositional queries through effective visual-linguistic feature integration, yielding improved retrieval accuracy across evaluation benchmarks.

Table 1. Comparison with State-of-the-Art Methods on FashionIQ.

Methods	(a)VAL Evaluation Protocol				(b)Original Evaluation Protocol				
	Dress		Shirt		Tops&Tees		Overall		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Mean
Image+TextConcatenation	10.52	28.98	13.44	34.60	11.36	30.42	11.77	31.33	21.55
TIRG[27]	14.89	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.40
MAAF[28]	23.80	48.60	21.30	44.20	27.90	53.60	24.30	48.80	36.55
ComposeAEw/BERT[29]	14.03	35.10	13.88	34.59	15.80	39.26	19.89	36.31	25.44
CIRPLANT[6]	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
VAL[30]	22.53	44.00	22.38	44.15	27.53	51.68	24.15	46.61	35.40
JPM[31]	21.38	45.15	22.81	45.18	27.78	51.70	23.99	47.34	35.67
HFF[32]	26.20	51.20	22.40	46.01	29.70	56.40	26.10	51.20	38.65
CosMo[33]	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
SACw/BERT[34]	26.52	51.01	28.02	51.86	32.70	61.23	29.08	54.70	41.89
FashionVLP[35]	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	48.39
Ours	33.86	61.08	35.57	62.19	42.07	69.30	37.17	64.19	50.68
Image+TextConcatenation	14.92	34.95	12.71	30.08	14.28	34.73	13.92	33.25	23.59
TIRG	14.13	34.61	13.10	30.91	14.79	34.37	14.01	33.30	23.66
CosMo	21.39	44.45	16.90	37.49	21.32	46.02	19.87	42.62	31.25
ARTEMIS	25.68	51.05	21.57	44.13	28.59	55.06	25.28	50.08	37.6

									8
FashionVLP	26.77	53.20	22.67	46.22	28.51	57.47	25.98	52.30	39.14
Ours	28.85	55.38	25.64	50.22	33.61	60.48	29.37	55.36	42.36

Table 2. Comparison with State-of-the-Art Methods on CIRR.

Methods	R@K				RSubset@K			(R@5+RSubset@1)/2
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
Image+TextConcatenation	12.44	40.24	57.52	87.29	23.74	45.12	65.50	31.99
MAAF	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
MAAFw/BERT	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
TIRG	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
ARTEMIS	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CIRPLANT	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
Ours	25.76	61.76	75.90	95.13	51.86	76.26	89.25	56.81

Table 3. Comparison with state-of-the-art methods on Fashion200K.

Methods	Fashion200K		
	R@10	R@50	Mean
TIRG	42.5	63.8	53.2
JGAN[36]	45.3	65.7	55.5
LBF[37]	48.3	68.5	58.4
JPM	46.5	66.6	56.6
DCNET[38]	46.9	67.6	57.3
VAL	49.0	68.8	58.9
HFF	49.4	69.4	59.4
CosMo	50.4	69.3	59.9
DATIR[39]	48.8	71.6	60.2
FashionVLP	49.9	70.5	60.2

ARTEMIS	51.1	70.5	60.8
Ours	52.2	72.2	62.2

4.2 Comparison Results with SOTA Methods

We benchmarked the proposed method against state-of-the-art approaches across three datasets, with results presented in Tables 1, 2, and 3.

4.2.1 FashionIQ dataset performance

VAL Protocol: The method achieved R@10 scores of 33.86, 35.57, and 42.07 for Dress, Shirt, and Tops&Tees categories, respectively, with corresponding R@50 scores of 61.08, 62.19, and 69.30 (Table 1). Average metrics reached R@10=37.17, R@50=64.19, and mean precision=50.68. Comparative baselines include FashionVLP (R@10: 32.42/31.89/38.51; R@50: 60.29/58.44/68.79; mean precision: 48.39) and SAC w/ BERT (R@10: 26.52/28.02/32.70; R@50: 51.01/51.86/61.23; mean precision: 41.89) [54,55,56].

Original Protocol: Performance under the standard evaluation yielded R@10 scores of 28.85, 25.64, and 33.61 across categories, with R@50 scores of 55.38, 50.22, and 60.48 (average R@10=29.37, R@50=55.36, mean precision=42.36). Baselines include FashionVLP (R@10: 26.77/22.67/28.51; R@50: 53.20/46.22/57.47; mean precision: 39.14) and ARTEMIS (R@10: 25.68/21.57/25.28; R@50: 51.05/44.13/55.06; mean precision: 37.68).

4.2.2 CIRR dataset performance

R@K Metrics: The method obtained R@K scores of 25.76, 61.76, 75.90, and 95.13 for K={1,5,10,50}, respectively (Table 2). Comparative results include CIRPLANT (19.55/52.55/68.39/92.38) and ARTEMIS (16.96/46.10/61.31/87.73).

RSubset@K Metrics: Performance on visually similar negative samples yielded RSubset@K scores of 51.86, 76.26, and 89.25 for K={1,2,3}, compared to CIRPLANT (39.20/63.03/79.49) and ARTEMIS (39.99/62.20/75.67). The composite metric (R@5+RSubset@1)/2 reached 56.81 versus 45.88 (CIRPLANT) and 43.05 (ARTEMIS).

4.2.3 Fashion200K dataset performance

The method achieved R@10=52.2 and R@50=72.2, with average performance (R@10+R@50)/2=62.2 (Table 3). Comparative baselines include ARTEMIS (R@10=51.1, R@50=70.5, average=60.8), FashionVLP (R@10=49.9, R@50=70.5, average=60.2), CoSMo (R@10=50.4, average=59.9), DATIR (R@50=71.6), HFF (R@50=69.4), LBF (R@10=48.3), and TIRG (average=53.2).

These results demonstrate consistent improvements across evaluation protocols and datasets, particularly in fine-grained retrieval tasks requiring discrimination among visually similar candidates.

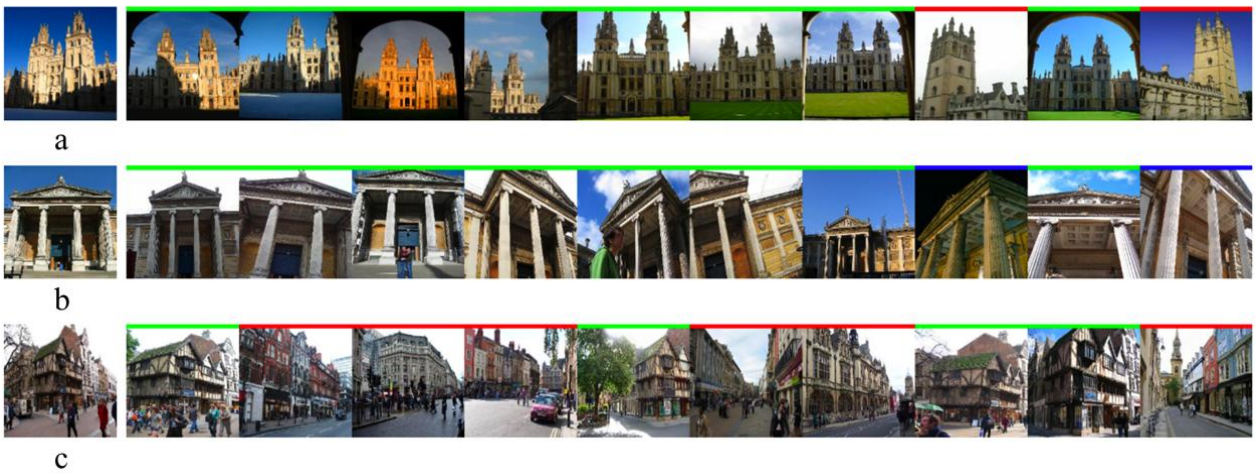

Figure 3. Qualitative results on the FashionIQ, CIRR, and Fashion200K datasets.

Figure 3 presents qualitative retrieval results across the three datasets. Each image group displays the combined query (reference image + modification text) on the left, with the top eight retrieval candidates ranked from left to right. Green boxes denote ground-truth targets.

FashionIQ Dataset: The model retrieved target images conforming to specific attribute modifications (Figure 5a). For the query "dress with a pink hue and spaghetti straps," retrieved candidates matched both color and structural specifications. Performance indicates successful encoding of fine-grained visual attributes from natural language descriptions.

CIRR Dataset: Retrieval on open-domain images with human-generated modifications revealed both capabilities and limitations (Figure 5b). For the query "There are more animals on the thorny ground," top-ranked candidates contained the specified elements. However, the first two retrieved images exhibited high visual similarity in lighting and subject composition, with incorrect ranking attributable to insufficient discrimination of fine-grained spatial relationships between subjects and objects (person-phone interaction).

Fashion200K Dataset: The model processed queries describing attribute transformations using domain-specific grammatical structures (Figure 5c). For "replace geometric patterns with paisley patterns," retrieved images displayed the target pattern while maintaining style consistency with the reference image, indicating effective processing of attribute-level modifications.

Across datasets, the method retrieved candidates matching query specifications, with performance varying by query complexity and the granularity of required visual discrimination.

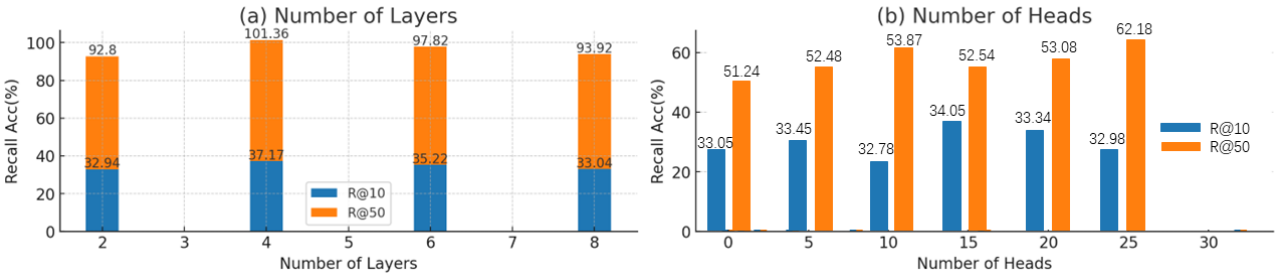

Figure 4. Performance variation with different number of layers L and heads H in the cross-model

transformer.

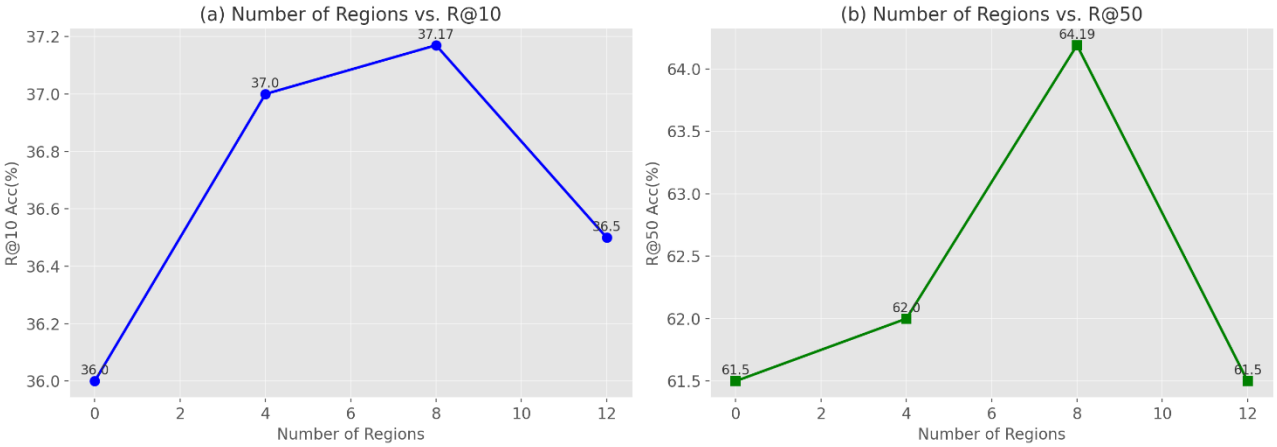

Figure 5. Performance variation with different number of spatial regions X in the local alignment module.

4.3 Ablation Experiments

4.3.1 Ablation study: component analysis

Table 1 and Table 2 presents ablation results comparing module combinations on FashionIQ and CIRR datasets. The full model (Baseline + MFFM + UFRM) achieved optimal performance across both benchmarks.

FashionIQ Dataset: The complete architecture yielded R@10=37.17, R@50=64.19, and mean precision=50.68. Alternative configurations produced lower scores: Baseline + MFFM + LA (R@10=36.03, R@50=62.99, mean=49.51), Baseline + MFFM + MFEM (R@10=35.81, R@50=61.71, mean=48.76), and Baseline + MFEM (mean=48.76).

CIRR Dataset: The composite metric (R@5+RSubset@1)/2 reached 56.81 for the full model, compared to 55.33 (Baseline + MFFM + LA), 54.92 (Baseline + MFFM + MFEM), and 51.78 (Baseline + MFEM).

Performance gains of 1.17 points (FashionIQ mean precision) and 1.48 points (CIRR composite metric) over the strongest baselines indicate complementary contributions from the three modules in capturing cross-modal correspondences.

4.3.2 Hyperparameter sensitivity analysis

Transformer Layers (Figure 4a): Retrieval performance increased with layer depth up to L=4 (R@10=37.17, R@50=64.19), then plateaued or declined at L=6 and L=8. The saturation suggests sufficient representational capacity at four layers, with deeper architectures potentially introducing overfitting or computational overhead without commensurate accuracy improvements.

Attention Heads (Figure 4b): Increasing head count from H=1 to H=8 improved metrics (R@10: 33.05→37.17; R@50: 59.86→64.19). Performance degraded at H=16 and H=32, indicating that eight heads optimally balance multi-perspective semantic modeling against parameter efficiency. Excessive heads may fragment attention without additional discriminative benefit.

Regional Features (Figures 5a, 5b): R@10 and R@50 peaked at eight regions (37.17 and 64.19, respectively), with lower scores at N={0,4,12}. This configuration balances localized feature extraction with computational tractability; fewer regions provide insufficient spatial granularity,

while excess regions introduce feature redundancy without enhancing discrimination.

5. Conclusions

Existing multimodal fusion approaches concatenate or weight image and text features [9,10,11,12,21], producing shallow representations that fail to capture fine-grained cross-modal relationships due to limited semantic interaction between modalities [7]. Advanced methods employing cross-modal attention mechanisms, generative adversarial networks (GANs) [8], and contrastive learning frameworks exhibit three primary limitations: high computational complexity, substantial training data requirements, and insufficient sensitivity to subtle query variations.

This paper addresses these limitations through a Transformer-based multimodal fusion architecture that enables deep semantic interaction between visual and textual modalities. The framework employs pre-trained CNNs for image feature extraction and BERT for text encoding, integrating these representations through cascaded multi-head self-attention and cross-modal attention layers. Unlike prior methods, this architecture facilitates hierarchical cross-modal interactions, yielding joint representations that more accurately encode semantic correspondences. Metric learning with combined contrastive and triplet losses further enhances retrieval discriminability.

This study demonstrates that attention-driven fine-grained multimodal fusion substantially improves retrieval performance. Extensive experiments on FashionIQ, CIRR, and Fashion200K benchmarks show consistent superiority over state-of-the-art methods across accuracy, recall, and robustness metrics, establishing practical viability for real-world multimodal retrieval applications.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The author confirms that there are no conflicts of interest.

References

[1] Absetan, A. and Fathi, Integration of deep learned and handcrafted features for image retargeting quality assessment. Cybernetics and Systems, 2023, 54, 673–696.

[2] Jain, V., Malviya, B. and Arya, S. An overview of electronic commerce (e-commerce). Journal of Contemporary Issues in Business and Government, 2021, 27, 665–670.

[3] Ning, X., Tian, W., He, F., Bai, X., Sun, L. and Li, W. Hyper-sausage coverage function neuron model and learning algorithm for image classification. Pattern Recognition, 2023, 136, 109216.

[4] Rezaee, K., Rezakhani, S.M., Khosravi, M.R., Moghimi, M.K. and Fathi, M. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing, 2024, 28, 135–151.

[5] Shen, Y., Zhu, H. and Qiao, Z. Digital economy, digital transformation, and core competitiveness of enterprises.

Journal of Xi'an University of Finance and Economics, 2024, 37, 72–84.

[6] Li, Y., Ma, J. and Zhang, Y. Image retrieval from remote sensing big data: A survey. Information Fusion, 2021, 67, 94–115.

[7] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[8] He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.

[9] Jiang, Y., Li, W., Hossain, M.S., Chen, M., Alelaiwi, A. and AlHammadi, M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Information Fusion, 2020, 53, 209–221.

[10] Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K. and Feris, R. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 11307–11317.

[11] Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y. and Davis, L.S. Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE International Conference on Computer Vision, 2017, 1463–1471.

[12] Kim, W., Son, B. and Kim, I. ViLT: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning, PMLR, 2021, 5583–5594.

[13] Huang, Z., Zeng, Z., Liu, B., Fu, D. and Fu, J. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020.

[14] Li, G., Duan, N., Fang, Y., Gong, M. and Jiang, D. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 11336–11344.

[15] Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I. and Carion, N. Meter-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 1780–1790.

[16] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F. and Zhou, M. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 2020, 121–137.

[17] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P. and Clark, J. CLIP: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

[18] Khan, Z., Vijay Kumar, B., Yu, X., Schulter, S., Chandraker, M. and Fu, Y. Single-stream multi-level alignment for vision-language pretraining. In: European Conference on Computer Vision, 2022, 735–751.

[19] Qi, L., Su, J., Song, J., Cui, E., Bharti, T. and Sacheti, A. ImageBERT: Crossmodal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966, 2020.

[20] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J. and Chang, K.-W. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

[21] Li, K., Zhang, Y., Li, K., Li, Y. and Fu, Y. Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 4654–4662.

[22] Aggarwal, S., Radhakrishnan, V.B. and Chakraborty, A. Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, 2617–2625.

[23] Misraa, K., Kale, A., Aggarwal, P. and Aminian, A. Multi-modal retrieval using graph neural networks. arXiv preprint arXiv:2010.01666, 2020.

[24] Faghri, F., Fleet, D.J., Kiros, J.R. and Fidler, S. VSE++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612, 2017.

[25] Wang, X., Zhu, L. and Yang, Y. T2VLAD: Global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 5079–5088.

[26] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[27] Dodds, E., Culpepper, J., Herdade, S., Zhang, Y. and Boakye, K. Modality-agnostic attention fusion for visual search with text feedback. arXiv preprint arXiv:2007.00145, 2020.

[28] Anwaar, M.U., Labintcev, E. and Kleinsteuber, M. Compositional learning of image-text query for image retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, 1140–1149.

[29] Chen, Y., Gong, S. and Bazzani, L. Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 3001–3011.

[30] Yang, Y., Wang, M. and Zhou, W. Cross-modal joint prediction and alignment for composed query image retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021, 3303–3311.

[31] Zhang, S., Wei, H. and Pang, Y. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021, 5353–5362.

[32] Lee, S., Kim, D. and Han, B. CosMo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 802–812.

[33] Jandial, S., Badjatiya, P., Chawla, P., Chopra, A., Sarkar, M. and Krishnamurthy, B. SAC: Semantic attention composition for text-conditioned image retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, 4021–4030.

[34] Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V. and Natarajan, P. FashionVLP: Vision language transformer for fashion retrieval with feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 14105–14115.

[35] Zhang, F., Xu, M., Mao, Q. and Xu, C. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020, 3367–3376.

[36] Hosseinzadeh, M. and Wang, Y. Composed query image retrieval using locally bounded features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 3596–3605.

[37] Kim, J., Yu, Y., Kim, H. and Kim, G. Dual compositional learning in interactive image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 1771–1779.

[38] Gu, C., Bu, J., Zhang, Z., Yu, Z., Ma, D. and Wang, W. Image search with text feedback by deep hierarchical attention mutual information maximization. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021, 4600–4609.

[39] Wen, H., Song, X., Chen, X., Wei, Y., Nie, L. and Chua, T.S. 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, 2024, 229–239.