Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval

Haokun Wen
School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
Shenzhen, China
whenhaokun@gmail.com

Xuemeng Song*
School of Computer Science and Technology, Shandong University
Qingdao, China
sxmustc@gmail.com

Xiaolin Chen
School of Software, Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University
Jinan, China
cxlicd@gmail.com

Yinwei Wei
Faculty of Information Technology, Monash University
Melbourne, Australia
weiyinwei@hotmail.com

Liqiang Nie*
School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
Shenzhen, China
nieliqiang@gmail.com

Tat-Seng Chua
School of Computing, National University of Singapore
Singapore
dcscts@nus.edu.sg

ABSTRACT

Composed image retrieval (CIR) aims to retrieve the target image based on a multimodal query, *i.e.*, a reference image paired with corresponding modification text. Recent CIR studies leverage vision-language pre-trained (VLP) methods as the feature extraction backbone, and perform nonlinear feature-level multimodal query fusion to retrieve the target image. Despite the promising performance, we argue that their nonlinear feature-level multimodal fusion may lead to the fused feature deviating from the original embedding space, potentially hurting the retrieval performance. To address this issue, in this work, we propose shifting the multimodal fusion from the feature level to the raw-data level to fully exploit the VLP model's multimodal encoding and cross-modal alignment abilities. In particular, we introduce a Dual Query Unification-based Composed Image Retrieval framework (DQU-CIR), whose backbone simply involves a VLP model's image encoder and a text encoder. Specifically, DQU-CIR first employs two training-free query unification components: text-oriented query unification and vision-oriented query unification, to derive a unified textual and visual query based on the raw data of the multimodal query, respectively. The unified textual query is derived by concatenating the modification text with the extracted reference image's textual description, while the unified visual query is created by writing the key modification words onto the reference image. Ultimately, to address diverse search intentions, DQU-CIR linearly combines the features of the two unified queries encoded by the VLP model to retrieve the target image. Extensive experiments on four real-world datasets validate the effectiveness of our proposed method.

CCS CONCEPTS

• **Information systems** → **Image search**.

*Corresponding authors: Xuemeng Song and Liqiang Nie.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07…$15.00
https://doi.org/10.1145/3626772.3657727

KEYWORDS

Composed image retrieval, Multimodal fusion, Multimodal retrieval

ACM Reference Format:
Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2024. 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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657727

1 INTRODUCTION

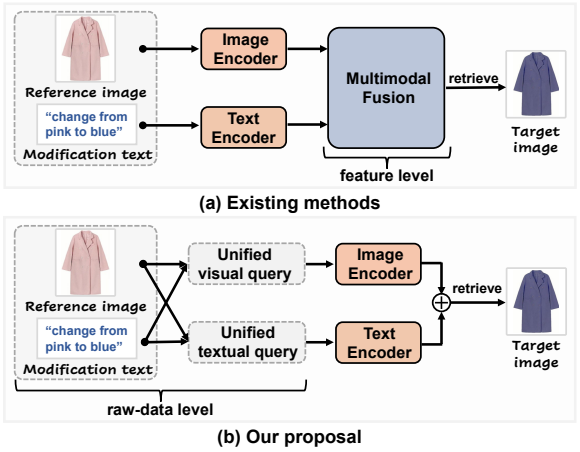

Figure 1: Comparison between existing methods and ours.

Different from traditional image retrieval that only allows users to express search intentions through a pure text or image query [24, 26, 29, 30, 46], composed image retrieval (CIR) enables users to utilize a multimodal query, *i.e.*, a reference image with a modification text expressing some modification demands, to retrieve the target image, as exemplified in Figure 1. Given its promising application potential in many real-world scenarios, including intelligent robots [35] and commercial platforms [8, 42], CIR has gained increasing research attention in recent years. The pipeline of the existing methods can be generally summarized in Figure 1(a). The reference image and the modification text are first processed by the image and text encoders, respectively. Thereafter, the extracted

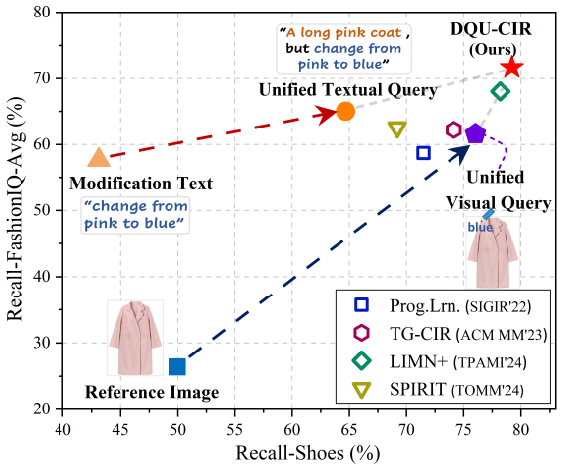

Figure 2: Performance comparison of our method with state-of-the-art baseline methods on two public datasets.

image and text features are fused through elaborately designed multimodal fusion functions to comprehend the users' search demands. Finally, the fused output is utilized to retrieve the target image.

According to the type of utilized feature extraction backbone, existing efforts can be broadly classified into two groups: traditional model-based methods and vision-language pre-trained (VLP) model-based methods. The former utilizes traditional models like ResNet [12] or LSTM [13] to extract features for the reference/target image and the modification text, while the latter adopts the VLP model, especially those with dual encoders, such as CLIP [31] and BLIP [20], as the feature extraction backbone. In contrast to traditional models, which are trained on single-modality datasets with limited scale, VLP models are pre-trained with large-scale image-text corpora, thereby achieving a more powerful capability in encoding both image and text information. Previous studies have consistently demonstrated the superiority of VLP models over the traditional feature extraction backbones in the CIR task [1, 28, 39].

Although recent CIR methods have achieved promising performance with the advances in VLP models, they overlook the following crucial point. That is, benefiting from the image-text contrastive learning pre-training task, VLP models typically map the image and text into a common embedding space with corresponding encoders. It is apparent that this property of VLP models has naturally facilitated the multimodal query fusion and the subsequent target image retrieval of the CIR task, since the reference image, modification text, and target image can be projected into the common embedding space with VLP models' encoders. Nevertheless, existing VLP model-based CIR methods further employ nonlinear multimodal fusion functions to fuse the extracted image/text features, which can potentially cause the fused multimodal query feature to deviate from the original common embedding space, thus hindering the target image retrieval.

Therefore, in this work, we propose to move the multimodal query fusion process from the feature level to the raw-data level, to harness the full potential of VLP models. The main idea is to convert the multimodal query to a unified single-modal query that encapsulates the user's original search intention. This unified single-modal query can then be directly encoded by the VLP model's encoder and its feature can be used directly for target image retrieval, eliminating the need for feature-level multimodal query fusion. In this way, the multimodal encoding and cross-modal alignment abilities of the VLP model can be fully utilized.

In particular, we design two query unification strategies: text-oriented unification and vision-oriented unification. Text-oriented unification focuses on unifying the multimodal query into a pure text query, where any advanced image captioning model [19] can be used to convert the reference image into a high-quality textual caption and hence the unified textual query can be obtained by concatenating the generated reference image caption and the modification text. Vision-oriented unification targets composing the multimodal query into a unified image query by directly writing the target image description words mentioned in the modification text onto the reference image. In particular, we resort to the Large Language Model (LLM) [32] to capture the target image description words presented in the modification text, to take advantage of its outstanding natural language processing capability. Notably, the two query unification strategies are training-free, requiring no parameters and can be conducted offline.

To cope with the various search cases in real-world CIR tasks, we develop a dual query unification-based composed image retrieval framework (DQU-CIR), where the two unification strategies are jointly considered, as illustrated in Figure 1(b). The underlying philosophy is that we expect that text-oriented query unification should be more useful for handling queries with complex search intentions, whereas vision-oriented query unification excel at simpler search queries. This is because the user's complex search intention tends to be delivered by the modification text, while the unified textual query derived by the text-oriented query unification just contains the complete modification text, and should encapsulate the user's complex search intention better. In contrast, for simpler modification cases, the reference image typically plays a more significant role. Therefore, the output query of the vision-oriented unification, keeping the full content of the reference image, is more adept at capturing the user's search intention.

Towards this end, DQU-CIR first conducts the text-oriented query unification and vision-oriented query unification in parallel. Subsequently, DQU-CIR encodes the unified textual and visual queries with the text and image encoder of the VLP model, respectively. Thereafter, DQU-CIR employs the linear weighted addition strategy to combine the features of the unified textual query and the unified visual query for target image retrieval. This linear fusion strategy is designed to ensure that the fused multimodal query embedding remains within the original embedding space of the VLP model, thus benefiting the final target image retrieval. To provide an intuitive demonstration of the effectiveness of our DQU-CIR, we illustrate the performance of our DQU-CIR and several state-of-the-art baseline methods on two public datasets (*i.e.*, FashionIQ and Shoes) in Figure 2, where the performance of certain derivatives of our framework is also presented for better understanding the contribution of each key component. As can be seen, our DQU-CIR outperforms all the cutting-edge baselines even with a simple model design. Our main contribution can be summarized as follows.

- We design two training-free multimodal query unification strategies that can convert a multimodal query into a unified

textual query and a unified visual query, respectively. To the best of our knowledge, we are the first to explore the raw-data level multimodal fusion in the context of CIR, which can fully leverage the VLP model's multimodal encoding and cross-modal retrieval capabilities.

- We propose a dual query unification-based composed image retrieval framework, named DQU-CIR, which effectively integrates the two query unification strategies for handling the real-world CIR cases with diverse search intentions. Notably, our framework consists of only a text encoder, an image encoder, and an MLP for learning the weights to linearly fuse the embeddings of the two unified queries.
- We conduct extensive experiments on four real-world public datasets, and the results demonstrate the superiority of our DQU-CIR over the state-of-the-art methods. In addition, we surprisingly found that directly writing descriptive words onto the image can achieve promising multimodal fusion results, which indicates the superior Optical Character Recognition (OCR) potential of the image encoder of the VLP model. We believe this would inspire the multimodal learning community to approach multimodal fusion from a new perspective. As a byproduct, we have released the source codes to benefit other researchers[1].

2 RELATED WORK

2.1 Vision-Language Pre-trained Models

Inspired by the success of pre-trained models in the field of computer vision (CV) and natural language processing (NLP), many efforts have been dedicated to developing Vision-Language Pre-trained (VLP) models [14, 50]. Benefiting from the pre-training on large-scale image-text corpora, VLP models yield universal cross-modal encoding capabilities, which have shown remarkable performance in various downstream multimodal tasks [28, 40, 52]. Considering the inference speed for the retrieval task, in the context of CIR, we specifically focused on the VLP methods with the dual encoder structure [16, 31]. These models employ two single-modal encoders to encode images and text, respectively. Among them, the most representative method is CLIP [31], which is pre-trained on large-scale image-text pairs through the cross-modal contrastive learning task. It exhibits promising multimodal alignment and cross-modal retrieval capabilities, which are crucial for the CIR task. In this study, we have chosen CLIP as our feature extraction backbone. In addition to its exceptional multimodal encoding capabilities, we have also observed notable natural language processing proficiency in its text encoder and Optical Character Recognition (OCR) potential in its image encoder. These properties of CLIP have enabled us to achieve promising CIR performance.

2.2 Composed Image Retrieval

According to the type of utilized feature extraction backbone, existing CIR methods can be classified into two groups: traditional model-based methods [5, 7, 18, 21, 34, 37, 43, 53] and VLP model-based methods [1, 6, 22, 23, 28, 38, 39]. The former employs the traditional models (e.g., ResNet and LSTM) to encode the given image and text. Due to the limited multimodal feature encoding capabilities of traditional models, methods of this group can only achieve suboptimal performance. While the latter group leverages the advanced VLP models for image/text encoding, which has shown more promising results due to their powerful multimodal feature encoding capabilities. A typical example is CLIP4CIR [1], which first leverages CLIP to extract the image and text features, and then employs a combiner module to fuse the extracted features. Even with the simple design, CLIP4CIR also achieves satisfactory performance. Recently, Wen et al. [38] proposed LIMN+, which also employs CLIP as the feature extraction backbone. LIMN+ leveraged a masked Transformer [33] to fuse the extracted image/text features and based on that retrieve the target images. Additionally, it introduces a data augmentation technique to alleviate the issue of limited training samples are available for training the CIR models.

Although these methods have made prominent progress, they conduct multimodal fusion at the feature level with elaborate non-linear functions, which may deviate the fused feature from the original embedding space. In contrast, we propose to achieve multimodal fusion at the raw-data level, which can be directly encoded by the advanced VLP methods. Then the superior multimodal encoding and cross-modal alignment capabilities of the VLP model can be fully utilized.

3 DQU-CIR

In this section, we first formulate the problem and then detail DQU-CIR framework. As shown in Figure 3, DQU-CIR comprises three key components: text-oriented unification, vision-oriented unification, and linear adaptive fusion-based target retrieval.

3.1 Problem Formulation

In this work, we aim to address the task of CIR, which is to retrieve the target image that meets the multimodal query (*i.e.,* a reference image, and a modification text). Formally, suppose we have a set of N training triplets, denoted as $\mathcal{T} = \{(x_r, t_m, x_t)_i\}_{i=1}^{N}$, where x_r, t_m, and x_t refer to the reference image, modification text, and target image, respectively. Given a multimodal query (x_r, t_m), our goal is to first derive a unified textual query q_t and a unified visual query q_v, each containing the user's essential search demand, through raw-data level multimodal fusion operations. Subsequently, we seek to optimize an embedding function for the dual query (q_t, q_v) and an image embedding function for the target image x_t, which can ensure the derived representations of the dual query and the target image are as close as possible within the embedding space. This can be formally expressed as follows,

$$\mathcal{F}\left(q_t, q_v | x_r, t_m\right) \rightarrow \mathcal{H}\left(x_t\right), \tag{1}$$

where $\mathcal{F}$ and $\mathcal{H}$ denote the to-be-learned functions for embedding the unified dual queries and the target image, respectively.

3.2 Text-oriented Query Unification

In this component, we aim to employ an image captioning model to derive a textual description for the reference image. Then by concatenating the reference image caption and the modification text, the user's multimodal query can be converted into a pure sentence query. Specifically, inspired by the remarkable success

[1] https://github.com/haokunwen/DQU-CIR.

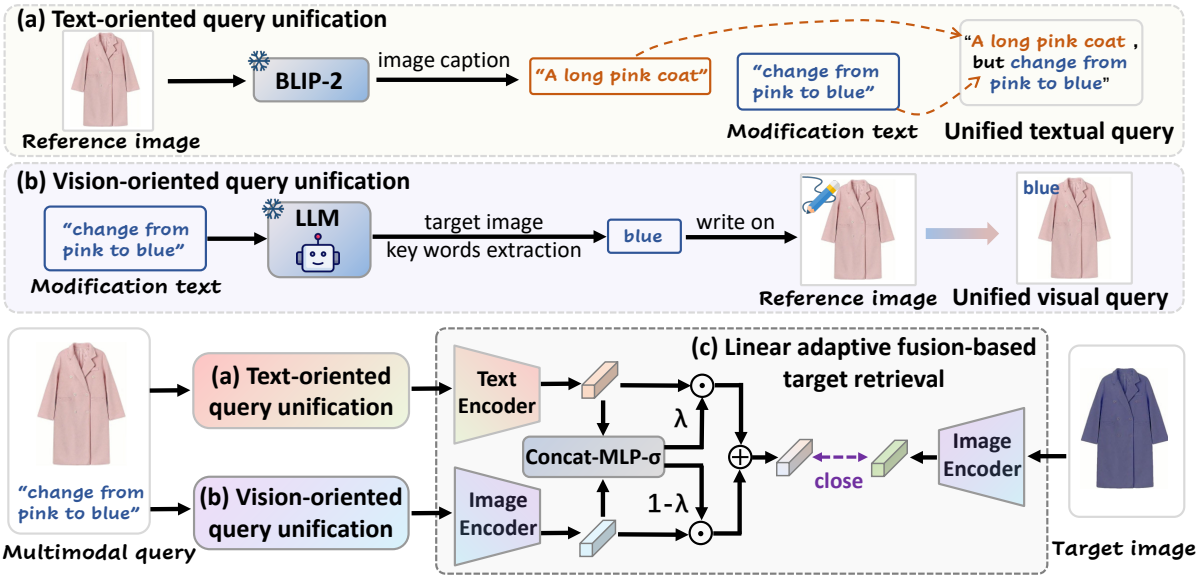

Figure 3: The proposed DQU-CIR consists of three components: (a) text-oriented query unification, (b) vision-oriented query unification, and (c) linear adaptive fusion-based target retrieval.

of the VLP models in image captioning [3, 4], we resort to the BLIP-2 [19] model to generate a high-quality description for each given reference image[2]. Formally, we have,

$$t_r = \text{BLIP-2}\,(x_r)\,, \tag{2}$$

where t_r denotes the generated caption for the reference image x_r.

Subsequently, as illustrated in Figure 3(a), we derive the unified textual query q_t by concatenating the reference image caption and the modification text through the template "t_r, but t_m". In this way, the derived textual query contains the essential textual description of the reference image and the complete information of the modification text, and leaves the task of user intention reasoning to the powerful CLIP text encoder. Essentially, the text-oriented unification should benefit the CIR cases with complex modification requests, where the modification text plays the dominant role in delivering the user's search demand.

3.3 Vision-oriented Query Unification

In this part, we aim to derive a unified image query that encapsulates the user's search demand based on the input multimodal query. One straightforward solution for deriving the unified image query is to synthesize a new image that depicts the user's overall search demand with the image manipulation technology [17, 45], where the modification text is used to manipulate the reference image. However, to ensure the image manipulation effect, the off-the-shelf generative image manipulation models need to be re-trained on the CIR dataset, which would lead to additional computational costs. Therefore, we ingeniously design a simple but effective vision-oriented query unification method, that is first extracts the key words of the modification text that indicate the desired attributes of the target image, and directly writes them on the reference image. This approach ensures that the visual details of the reference image

[2]Any other advanced image captioning model is applicable.

Prompt: "Analyze the provided text, which details the differences between two images. Extract and list the distinct features of the target image mentioned in the caption, separating each feature with a comma. Ensure to eliminate any redundancies and correct typos. The input text for this task is: #[Modification Text] "

Example:

Prompt + "is blue in color and has a boat neck instead of round"

blue, boat neck

Figure 4: Illustration of our designed prompt and an example for the key words extraction.

are completely retained, and the crucial target desired attributes mentioned in the modification text are embedded in the image. In fact, it has been proven that CLIP image encoder has great potential in OCR [36, 47]. Accordingly, we expect the CLIP image encoder to fulfil the user intention reasoning given the unified image query.

It is worth mentioning that the modification text usually contains both the target desired attributes as well as the original attributes of the reference image that should be replaced. Therefore, to extract the target image description words from the modification text, it is essential to discern which parts pertain to the reference image and should be discarded, and which parts are relevant to the target image and thus should be extracted. This process involves careful reasoning to accurately separate and identify the relevant segments of the modification text. Inspired by the compelling success of the large language models (LLM) in various natural language reasoning tasks, we resort to the Gemini-pro [32] to automatically

extract the target image description words from the modification text. Specifically, we design a prompt [25] to enable LLM to assist us in accomplishing this goal. The designed prompt and an example are shown in Figure 4.

Thereafter, as depicted in Figure 3(b), we directly write the extracted key words describing the target image on the blank part of the reference image with the help of the computer vision library OpenCV [2]. We regard the reference image with modification keywords on it as the unified visual query q_v. Intuitively, the obtained visual query contains the visual details of the reference image and the key description of the target image, which is friendly to the CIR cases with simple modification requests.

3.4 Linear Adaptive Fusion-based Target Retrieval.

As aforementioned, modification demands in the context of CIR are diverse, including both complex and simple modifications. In cases with complex modification, many visual properties of the reference image need to be altered, thus the modification text becomes crucial. As a result, unifying the multimodal query into a textual query is expected to exhibit superior retrieval performance. In contrast, in cases with simple modifications, only minor aspects of the reference image need to be changed, thus the reference image plays a more dominant role. Accordingly, the unified visual query is more suitable to handle these cases. In light of this, we propose to adaptively consider the unified textual query and visual query. Specifically, we first encode the unified textual query q_t and visual query q_v through the CLIP text encoder and image encoder, respectively. Formally, we have,

$$\begin{cases} \mathbf{f}_{textual} = \text{CLIP}_{\text{text}}\left(q_t\right), \\ \mathbf{f}_{visual} = \text{CLIP}_{\text{image}}\left(q_v\right), \end{cases} \tag{3}$$

where $\mathbf{f}_{textual} \in \mathbb{R}^D$ and $\mathbf{f}_{visual} \in \mathbb{R}^D$ refer to the extracted textual query feature and visual query feature, respectively.

We then adaptively fuse the features of the unified queries q_t and q_v with a linear weighted addition, where the weights for fusion are learnable based on the given context. It is worth mentioning that in the previous query unification, we have fulfilled the multimodal query fusion at the raw-data level, which is distinct from previous studies that conduct the multimodal query fusion at the embedding level. In this part, the fusion process works on fusing the two kinds of unified queries for adaptively handling CIR cases with various modification demands. The reason for designing the linear fusion operation is to maintain the fused query embedding still residing in the VLP's original embedding space, in which the target image embedding resides, and hence facilitates the final target retrieval.

Specifically, we fuse the features of the two unified queries with the following linear weighted addition,

$$\begin{cases} \mathbf{f}_q = \lambda * \mathbf{f}_{textual} + (1-\lambda) * \mathbf{f}_{visual}, \\ \lambda = \sigma\left(\text{MLP}\left(\left[\mathbf{f}_{textual} \| \mathbf{f}_{visual}\right]\right)\right), \end{cases} \tag{4}$$

where $\mathbf{f}_q \in \mathbb{R}^D$ denotes the final composed query feature. λ is the trade-off weight for balancing the importance of the unified textual query and unified visual query. $[\cdot\|\cdot]$ denotes the feature concatenation operation, MLP refers to the multi-layer perceptron, and σ represents the Sigmoid activate function.

Finally, the task boils down to pushing the composed query feature close to the corresponding target image feature in the embedding space. Here we resort to the commonly used batch-based classification loss as the optimization function. The essence of batch-based classification loss is to closely align the composed query feature with the target image feature within a mini-batch, while simultaneously distancing it from the features of other images. Here we also derive the target image feature through the CLIP image encoder as follows,

$$\mathbf{f}_t = \text{CLIP}_{\text{image}}\left(x_t\right), \tag{5}$$

where $\mathbf{f}_t \in \mathbb{R}^D$ is the target image feature. Mathematically, we have the loss function for optimizing the composed query feature and target image feature as follows,

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\cos\left(\mathbf{f}_{qi}, \mathbf{f}_{ti}\right)/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\cos\left(\mathbf{f}_{qi}, \mathbf{f}_{tj}\right)/\tau\right\}}\right\}, \tag{6}$$

where the subscript i refers to the i-th triplet sample in the mini-batch, B is the batch size, $\cos\left(\cdot,\cdot\right)$ denotes the cosine similarity function, and τ is the temperature factor. For optimization, we parameterize the final objective function for our DQU-CIR as follows,

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}}\left(\mathcal{L}\right), \tag{7}$$

where $\mathbf{\Theta}$ denotes the to-be-learned parameters of our DQU-CIR.

4 EXPERIMENT

4.1 Experimental Settings

4.1.1 Datasets. We chose four public datasets to evaluate our DQU-CIR, including three fashion-domain datasets including FashionIQ [41], Shoes [9], and Fashion200K [10], as well as an open-domain dataset CIRR [27].

4.1.2 Implementation Details. DQU-CIR employs the pre-trained CLIP (ViT-H/14 version) for feature extraction and is trained by the AdamW optimizer. For the FashionIQ, Fashion200K, and CIRR datasets, CLIP parameters are fine-tuned at a learning rate of $1e-6$, while other DQU-CIR parameters are optimized at $1e-4$ for effective convergence. For the Shoes dataset, these learning rates are adjusted to $5e-6$ for CLIP and $5e-5$ for other parameters. The feature dimension D is set to 1024, the batch size B is fixed at 16, and the temperature factor τ in Eqn.(6) is set to 0.1 for all four datasets. All experiments are conducted using PyTorch on a server equipped with a single A100-40G GPU.

4.1.3 Evaluation. We adhered to the standard evaluation protocols for each dataset and utilized the Recall@k (R@k) metric for a fair comparison. For the FashionIQ dataset, we followed the challenge's evaluation criteria [41], focusing on R@10 and R@50 across its three categories. Notably, FashionIQ comprises two evaluation protocols: the VAL-Split [5] and the Original-Split [41]. The VAL-Split is introduced by the early-stage CIR study, which builds the candidate image set for testing based on the union of the reference images and target images in all the triplets of the validation set. The Original-Split is recently adopted, which directly uses the original candidate image set provided by the FashionIQ dataset for testing. We reported results for both dataset splits to provide a comprehensive evaluation. In the case of the Shoes and Fashion200K datasets,

Table 1: Performance comparison on FashionIQ with respect to R@k(%). Averages and standard deviations are reported from 5 random seed experiments. The best results are colored in blue, while the second-best results are underlined.

Split	Method	Dresses		Shirts		Tops&Tees		Average		Avg.
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
VAL-Split	*Traditional Model-Based Methods*									
	TIRG [34] (CVPR'19)	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	12.60
	VAL [5] (CVPR'20)	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	16.49
	CIRPLANT [27] (ICCV'21)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
	CLVC-Net [37] (SIGIR'21)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	44.56
	ARTEMIS [7] (ICLR'22)	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	38.17
	EER [49] (TIP'22)	30.02	55.44	25.32	49.87	33.20	60.34	29.51	55.22	42.37
	CRR [48] (MM'22)	30.41	57.11	30.73	58.02	33.67	64.48	31.60	59.87	45.74
	AMC [53] (TOMM'23)	31.73	59.25	30.67	59.08	36.21	66.60	32.87	61.64	47.26
	CRN [44] (TIP'23)	32.67	59.30	30.27	56.97	37.74	65.94	33.56	60.74	47.15
	CMAP [21] (TOMM'24)	36.44	64.25	34.83	60.06	41.79	69.12	37.64	64.42	51.03
	VLP Model-Based Methods									
	Prog. Lrn. [51] (SIGIR'22)	38.18	64.50	48.63	71.54	52.32	76.90	46.37	70.98	58.68
	TG-CIR [39] (MM'23)	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	62.21
	LIMN+ [38] (TPAMI'24)	52.11	75.21	57.51	77.92	62.67	82.66	57.43	78.60	68.02
	SPIRIT [6] (TOMM'24)	43.83	68.86	52.50	74.19	56.60	79.25	50.98	74.10	62.54
	DQU-CIR	**57.63**$^{\pm 0.24}$	**78.56**$^{\pm 0.50}$	**62.14**$^{\pm 0.66}$	**80.38**$^{\pm 0.15}$	**66.15**$^{\pm 0.50}$	**85.73**$^{\pm 0.25}$	**61.97**$^{\pm 0.28}$	**81.56**$^{\pm 0.22}$	**71.77**$^{\pm 0.17}$
Original-Split	*Traditional Model-Based Methods*									
	TIRG [34] (CVPR'19)	14.13	34.61	13.10	30.91	14.79	34.37	14.01	33.30	23.66
	ARTEMIS [7] (ICLR'22)	25.68	51.05	21.57	44.13	28.59	55.06	25.28	50.08	37.68
	VLP Model-Based Methods									
	CLIP4CIR [1] (CVPR'22)	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	47.21
	Prog. Lrn. [51] (SIGIR'22)	33.60	58.90	39.45	61.78	43.96	68.33	39.02	63.00	51.01
	FAME-ViL [11] (CVPR'23)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.30
	SPIRIT [6] (TOMM'24)	39.86	64.30	44.11	65.60	47.68	71.70	43.88	67.20	55.54
	BLIP4CIR+Bi [28] (WACV'24)	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.40
	DQU-CIR	**51.90**$^{\pm 0.64}$	**74.37**$^{\pm 0.39}$	**53.57**$^{\pm 0.27}$	**73.21**$^{\pm 0.34}$	**58.48**$^{\pm 0.46}$	**79.23**$^{\pm 0.29}$	**54.65**$^{\pm 0.38}$	**75.60**$^{\pm 0.18}$	**65.13**$^{\pm 0.14}$

aligned with prior studies [5, 7, 37, 39], we reported R@1, R@10, R@50, and their calculated averages. For CIRR, following the precedent set by [7, 27], we reported R@k for $k = 1, 5, 10, 50$, R_{subset}@k for $k = 1, 2, 3$, and the average of R@5 and R_{subset}@1.

4.2 Performance Comparison

We compared DQU-CIR with the following two baseline groups: traditional model-based methods and VLP model-based methods. The first group methods utilize conventional models like ResNet and LSTM for feature extraction, including TIRG [34], VAL [5], CLRPLANT [27], CLVC-Net [37], ARTEMIS [7], EER [49], CRR [48], AMC [53], CRN [44], CMAP [21]. While the second group leverages advanced VLP models, such as CLIP and BLIP, as the feature extraction backbones, including CLIP4CIR [1], Prog.Lrn. [51], FAME-ViL [11], TG-CIR [39], LIMN+ [38], SPIRIT [6], and BLIP4CIR+Bi [28]. Tables 1, 2, 3, and 4 summarize the performance comparison on the four datasets. Note that some baseline methods only report results for selected datasets in their original papers. Accordingly, results for datasets not covered by the baseline methods are omitted from our comparison.

Our observations are presented as follows. 1) Traditional model-based methods are generally outperformed by VLP model-based

Table 2: Performance comparison on Shoes regarding R@k(%). Averages and standard deviations are reported from 5 random seed experiments. The best results are colored in blue, while the second-best results are underlined.

Method	R@1	R@10	R@50	Avg.
Traditional Model-Based Methods				
TIRG [34] (CVPR'19)	12.60	45.45	69.39	42.48
VAL [5] (CVPR'20)	16.49	49.12	73.53	46.38
CLVC-Net [37] (SIGIR'21)	17.64	54.39	79.47	50.50
ARTEMIS [7] (ICLR'22)	18.72	53.11	79.31	50.38
EER [49] (TIP'22)	20.05	56.02	79.94	52.00
CRR [48] (MM'22)	18.41	56.38	79.92	51.57
AMC [53] (TOMM'23)	19.99	56.89	79.27	52.05
CRN [44] (TIP'23)	18.92	54.55	80.04	51.17
CMAP [21] (TOMM'24)	21.48	56.18	81.14	52.93
VLP Model-Based Methods				
Prog. Lrn. [51] (SIGIR'22)	22.88	58.83	84.16	55.29
TG-CIR [39] (MM'23)	25.89	63.20	85.07	58.05
LIMN+ [38] (TPAMI'24)	–	68.37	88.07	–
SPIRIT [6] (TOMM'24)	–	56.90	81.49	–
DQU-CIR	**31.47**$^{\pm 1.31}$	**69.19**$^{\pm 0.99}$	**88.52**$^{\pm 0.31}$	**63.06**$^{\pm 0.69}$

Table 3: Performance comparison on Fashion200K regarding R@k(%). Averages and standard deviations are reported from 5 random seed experiments. The best results are colored in blue, while the second-best results are underlined.

Method	R@1	R@10	R@50	Avg.
Traditional Model-Based Methods				
TIRG [34] (CVPR'19)	14.1	42.5	63.8	40.1
VAL [5] (CVPR'20)	22.9	50.8	72.7	48.8
CLVC-Net [37] (SIGIR'21)	22.6	53.0	72.2	49.3
ARTEMIS [7] (ICLR'22)	21.5	51.1	70.5	47.7
EER [49] (TIP'22)	–	55.3	73.4	–
CRR [48] (MM'22)	24.9	56.4	73.6	51.6
CRN [44] (TIP'23)	–	53.5	74.5	–
CMAP [21] (TOMM'24)	24.2	56.9	75.3	52.1
VLP Model-Based Methods				
LIMN [38] (TPAMI'24)	–	57.2	76.6	–
SPIRIT [6] (TOMM'24)	–	55.2	73.6	–
DQU-CIR	$36.8^{\pm 3.8}$	$67.9^{\pm 2.1}$	$87.8^{\pm 0.3}$	$64.1^{\pm 1.7}$

methods across four datasets. This trend underscores the advantages of the VLP model's superior multimodal encoding capabilities. Such capabilities are pivotal for effective multimodal fusion and cross-modal retrieval in CIR tasks. 2) DQU-CIR consistently surpasses all baseline methods on the three fashion domain datasets by a significant margin. Specifically, DQU-CIR achieves absolute improvements of 5.51%, 11.72%, 8.63%, and 23.0% over the best baseline for the Avg. metric on FashionIQ VAL-Split, FashionIQ Original-Split, Shoes, and Fashion200K, respectively. These advancements highlight the effectiveness of fusing the multimodal query at raw-data level, and reflects the remarkable natural language reasoning capability of the CLIP text encoder and the impressive OCR capability of the CLIP image encoder. 3) DQU-CIR exhibits comparable performance on the open-domain CIRR dataset, unlike the notable performance improvements observed on fashion-domain datasets. To delve deeper into the underlying reasons, we conducted supplementary experiments with two derivatives of our model: using the unified textual query (TQU-CIR) and using the unified visual query only (VQU-CIR). The results are shown in Table 4. As can be seen, TQU-CIR outperforms with a large margin compared to VQU-CIR, with 14.36% towards the (R@5 + R_{subset}@1) / 2 metric. This may be due to the fact that the modification demands in the open-domain CIRR dataset are considerably more complex than those in the fashion-domain datasets. This modification complexity emphasizes the essential role of modification text in target image retrieval, while the reference image's contribution is negligible. Accordingly, the unified visual query proves inadequate in capturing the intricate modification requirements of the CIRR dataset, leading to suboptimal performance. In fact, TQU-CIR even exceeds the full model DQU-CIR in the $R_{subset}@k$ metrics, indicating that in cases with extremely complex modification demands, the unified visual query may have a detrimental effect, and relying solely on the unified textual query can yield better outcomes.

On Feature Extraction Backbone. To provide more insight into the effect of feature extraction backbones, we illustrated the performance of DQU-CIR with different versions of CLIP [15] on FashionIQ and Shoes in Figure 5. This includes ViT-based backbones

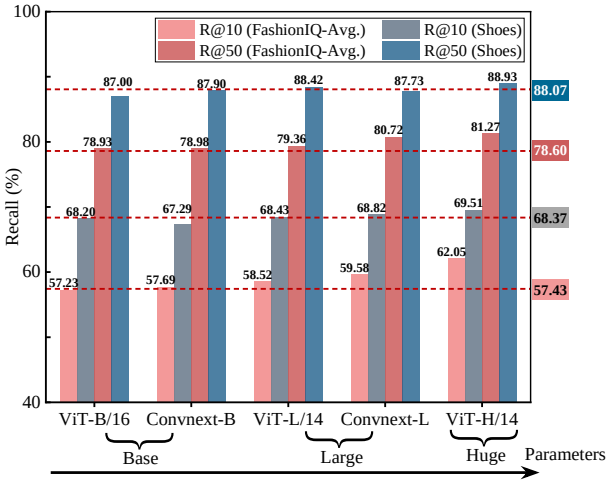

Figure 5: Performance on FashionIQ-Avg (VAL-Split) and Shoes of DQU-CIR with different versions of CLIP. The horizontal dashed lines denote the best baseline performance. The random seeds are fixed at 42 across the experiments.

"ViT-B/16", "ViT-L/14", and "ViT-H/14", as well as convolutional neural network (CNN)-based CLIP backbones like "Convnext-B" and "Convnext-L". As a reference, we also included the results of the best baseline LIMN+ [38] for the two datasets, represented by horizontal dashed lines. Notably, LIMN+ [38] adopts CLIP ViT-L/14 as the backbone. From Figure 5, we obtained the following observations. 1) The performance of our model generally improves with larger backbones, this is reasonable as the larger backbones typically have the more powerful multimodal encoding capabilities. Accordingly, utilizing a larger backbone could potentially enhance the performance. Specifically, we selected the "ViT-H/14" version as the feature extraction backbone in our experiments. Notably, since we move the multimodal query fusion from the feature level to the raw-data level, which saves computational and memory costs, making it possible for us to deploy the larger backbone under the same computing resources. 2) With the same "ViT-L/14" backbone, our method outperforms the best baseline, LIMN+, which involves complicated feature-level multimodal fusion and sophisticated data augmentation. This reflects the superiority of our well-designed raw-data level multimodal query fusion.

4.3 Ablation Study

To verify the influence of each component in our model, we conducted ablation study experiments on FashionIQ and Shoes, as shown in Table 5. Specifically, we compared DQU-CIR with the following ablation methods on three components.

- **On component (a) :** In the text-oriented unification component, we concatenated the generated reference image caption and the modification text as the unified text query. To gain deep insights of this component, we conducted the target image retrieval with three types of queries, including modification text only (**AM1**), unified textual query only (**AM2**), and reference image caption only (**AM3**).
- **On component (b) :** In the vision-oriented unification component, we directly write the target image description on the

Table 4: Performance comparison on CIRR with respect to R@k(%) and R_{subset}@k(%).The random seed is set to 42 because excessive submissions were being rejected by the evaluation server during multiple experiments. The best results are colored in blue, while the second-best results are underlined.

Method	R@k				R_{subset}@k			(R@5 + R_{subset}@1) / 2
	$k=1$	$k=5$	$k=10$	$k=50$	$k=1$	$k=2$	$k=3$	
Traditional Model-Based Methods								
TIRG [34] (CVPR'19)	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
CIRPLANT [27] (ICCV'21)	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
ARTEMIS [7] (ICLR'22)	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
VLP Model-Based Methods								
CLIP4CIR [1] (CVPR'22)	33.59	65.35	77.35	95.21	62.39	81.81	92.02	63.87
TG-CIR [39] (MM'23)	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
LIMN+ [38] (TPAMI'24)	43.33	75.41	85.81	97.21	69.28	86.43	94.26	72.35
SPIRIT [6] (TOMM'24)	40.23	75.10	84.16	96.88	73.74	89.60	95.93	74.42
BLIP4CIR+Bi [28] (WACV'24)	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
TQU-CIR (unified textual query only)	44.05	75.21	85.01	96.63	76.41	90.53	95.90	75.81
VQU-CIR (unified visual query only)	32.87	64.80	77.06	95.40	58.10	78.72	89.76	61.45
DQU-CIR	46.22	78.17	87.64	97.81	70.92	87.69	94.68	74.55

Table 5: Ablation study on FashionIQ and Shoes towards three key components of DQU-CIR. The "Ref-Img", "Mod-Text", and "Ref-Cap" denote the reference image, modification text, and the reference image caption obtained by BLIP-2, respectively. AM# is the abbreviation of the ablation method number for simplicity. The random seeds are fixed at 42 across the experiments.

AM#	Comp.	Ref-Img	Mod-Text	Ref-Cap	FashionIQ (VAL-Split)						Shoes		Avg.
					Dresses		Shirts		Tops&Tees				
					R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
1	(a)		✓		41.70	66.04	44.95	65.85	51.81	75.42	31.01	55.25	54.00
2			✓	✓	45.91	70.30	56.77	77.33	58.18	81.64	50.54	78.82	64.94
3				✓	3.47	10.36	15.90	29.20	9.38	20.45	10.51	30.38	16.21
4	(b)	✓			12.89	28.01	23.45	40.19	18.71	35.14	37.37	62.52	32.29
5		✓	✓(Blue)		43.83	68.86	51.91	72.37	55.43	77.00	66.38	85.69	65.18
6		✓	✓(Green)		44.42	68.82	51.96	72.18	54.26	76.70	65.59	85.75	64.96
7		✓	✓(Red)		44.22	67.92	52.31	72.42	53.90	77.00	66.33	84.89	64.87
8		✓	✓(Black)		44.52	68.42	51.42	72.87	54.26	76.95	65.81	85.29	64.94
9	(c)	Average addition			54.98	77.59	61.24	80.23	64.97	84.80	68.60	88.07	72.56
DQU-CIR					**56.67**	**78.14**	**62.46**	**80.42**	**67.01**	**85.26**	**69.51**	**88.93**	**73.55**

reference image. To investigate this component, we conducted the target image retrieval with the reference image only (**AM4**) as a base comparison. In addition, we further investigated the influence of different font colors utilized for writing on the reference image (**AM5-AM8**).

- **On component (c) :** In the linear adaptive fusion-based target retrieval component, we targeted at validating the effectiveness of the linear weighted addition function. Specifically, we employed the average addition for fusing the two unified queries by setting $\lambda = 0.5$ in Eqn.(4) (**AM9**).

From the ablation study results listed in Table 5, we gained the following observations. 1) **AM2** performs much better than both **AM1** and **AM3**. This is reasonable as both the reference image and the modification text convey partial user search intention. 2) **AM4** shows much inferior results than **AM5-AM8**. The significant performance gap demonstrates the effectiveness of directly writing target image description words on the reference image to fuse the multimodal query. This also reflects the potential OCR ability of the CLIP image encoder. 3) **AM5-AM8** achieve comparable performance, which implies that the font color does not significantly influence the vision-oriented query unification performance. Notably, in all the other experiments, we adopted the blue font, which achieves slightly higher performance. 4) **AM9** performs consistently worse than the full model. This proves that the designed linear weighted addition can adaptively fuse the unified textual query and unified visual query for handling cases with various modification demands. 5) In fact, **AM2** and **AM5** correspond to the two derivatives TQU-CIR and VQU-CIR, which only use the text-oriented unified query and the vision-oriented unified query, respectively. We noticed that **AM2** and **AM5** perform closely on the FashionIQ and Shoes dataset, unlike that on the CIRR dataset, where only using the unified textual query performs significantly better than solely using the unified visual query. This suggests that the modification demands in these two datasets are moderate, and the retrieval demands are not highly dependent on the modification text as in the CIRR dataset. Meanwhile, this also confirms the result

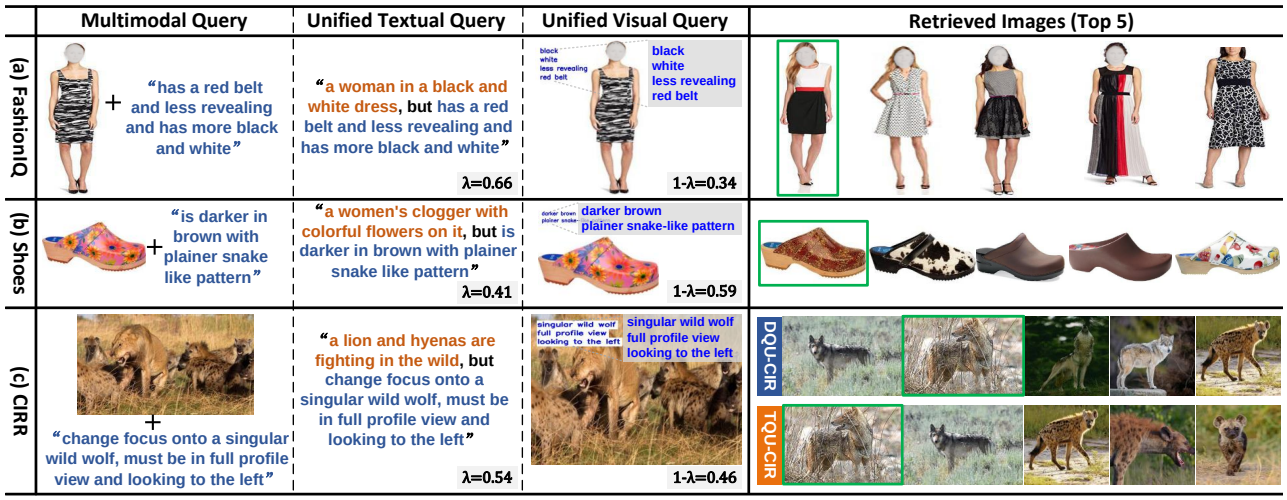

Figure 6: Case study on (a) FashionIQ, (b) Shoes, and (c) CIRR datasets.

that our linear weighted addition strategy performs well in balancing the unified textual query and unified visual query to address various modification cases.

4.4 Case Study

Figure 6 illustrates three examples by DQU-CIR across three datasets. The top 5 retrieved images are listed, where the green boxes denote the target images. To provide a more intuitive demonstration of our method, we also included the unified textual and visual queries. Specifically, for each unified textual query, the generated reference image caption part is in orange, while the modification text part is in blue. As the key words written on the reference image are too blurry in the top-left of the unified visual query due to limited space, we enlarged them on the top-right for better readability. As can be seen, the reference image captions generated by BLIP-2 and the key modification words extracted with Gemini-pro are generally accurate and reasonable. The numbers in the bottom-right denote the linear fusion weights. As shown in Figure 6(a), DQU-CIR successfully ranks the target image first given a dress modified in terms of belt, style, and color. The weights show that the unified textual query contributes more than the visual query. This is reasonable as this modification is a little bit complex, and the unified textual query can deliver full modification details by maintaining the original modification text. In Figure 6(b), our method also places the target image at the top. While in this case, the unified visual query plays a slightly more important role. This is also reasonable since the modification involves simpler changes, just the color and pattern. Accordingly, the reference image becomes crucial, and the visual query, which retains all the visual details of the reference image and recognizes the essential modification words, is assigned a higher weight. Figure 6(c) shows a failure case. It can be found that the extensive modification demands render the reference image less influence for the target image retrieval. To gain more insights, we provided the retrieval results of our derivative TQU-CIR that only utilizes the unified textual query for the target image retrieval. In this case, TQU-CIR successfully ranks the target image at the first place. This may be due to the fact that it better understands the modification demand "looking to the left". This suggests that in cases where the search demands are too unique that they can be directly delivered by the modification text, utilizing only the single unified textual query can achieve promising results. Overall, these observations validate the effectiveness of DQU-CIR in moving the multimodal fusion from the feature level to the raw-data level in the context of CIR.

5 CONCLUSION

In this work, we present a novel dual query unification-based composed image retrieval framework to address the challenging CIR task, which shifts the multimodal fusion from the conventional feature level to the raw-data level. In particular, we unified the multimodal query by concatenating the reference image caption with the modification text to derive the textual query, and directly writing the key modification words on the reference image to derive the visual query. The derived unified dual queries can be processed by the VLP model's dual encoders, respectively, fully leveraging their multimodal encoding capabilities. Furthermore, the unified textual query excels at handling complex modifications as it preserves the entire modification text. The unified visual query is effective for tackling simple modifications, as it contains the essential visual details of the reference image. To cater to diverse modification demands, we employed a linear weighted addition to combine the dual queries while ensuring that the fused features remain in the original embedding space. Extensive experiments have been conducted on four public datasets, and the results demonstrate the effectiveness of our method. In the future, we plan to explore more application scenarios with the raw data-level multimodal fusion strategy.

ACKNOWLEDGMENTS

This work is supported in part by Shenzhen College Stability Support Plan (No.:GXWD20220817144428005), Natural Science Foundation of China (No.:62376137), and Shandong Provincial Natural Science Foundation (No.:ZR2022YQ59). Haokun Wen and Xiaolin Chen acknowledge the support from the China Scholarship Council for pursuing their joint Ph.D. studies at NUS.

REFERENCES

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 21434–21442.

[2] Gary Bradski. 2000. The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25, 11 (2000), 120–123.

[3] Xiaolin Chen, Xuemeng Song, Liqiang Jing, Shuo Li, Linmei Hu, and Liqiang Nie. 2024. Multimodal Dialog Systems with Dual Knowledge-enhanced Generative Pretrained Language Model. *ACM Transactions on Information Systems* 42, 2 (2024), 53:1–53:25.

[4] Xiaolin Chen, Xuemeng Song, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. 2023. Dual Semantic Knowledge Composed Multimodal Dialog Systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1518–1527.

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2998–3008.

[6] Yanzhe Chen, Jiahuan Zhou, and Yuxin Peng. 2024. SPIRIT: Style-guided Patch Interaction for Fashion Image Retrieval with Text Feedback. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 6 (2024), 1–17.

[7] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–12.

[8] Weili Guan, Haokun Wen, Xuemeng Song, Chun Wang, Chung-Hsing Yeh, Xiaojun Chang, and Liqiang Nie. 2022. Partially Supervised Compatibility Modeling. *IEEE Transactions on Image Processing* 31 (2022), 4733–4745.

[9] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogério Schmidt Feris. 2018. Dialog-based Interactive Image Retrieval. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 676–686.

[10] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1472–1480.

[11] Xiaoping Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2023. FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2669–2680.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[14] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. 2021. GilBERT: Generative Vision-Language Pre-Training for Image-Text Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1379–1388.

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4904–4916.

[17] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2416–2425.

[18] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 802–812.

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning*. PMLR, 19730–19742.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 12888–12900.

[21] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. 2024. Cross-Modal Attention Preservation with Self-Contrastive Learning for Composed Query-Based Image Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 6 (2024), 1–22.

[22] Haoqiang Lin, Haokun Wen, Xiaolin Chen, and Xuemeng Song. 2023. CLIP-Based Composed Image Retrieval with Comprehensive Fusion and Data Augmentation. In *Proceedings of the Australasian Joint Conference on Artificial Intelligence*. Springer, 190–202.

[23] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Fine-grained Textual Inversion Network for Zero-Shot Composed Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1–10.

[24] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 15–24.

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (2023), 195:1–195:35.

[26] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. 2024. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 14052–14060.

[27] Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2105–2114.

[28] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-Directional Training for Composed Image Retrieval via Text Prompt Learning. In *Proceedings of the IEEE Conference on Winter Conference on Applications of Computer Vision*. IEEE, 5753–5762.

[29] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1104–1113.

[30] Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. 2024. Discriminative Probing and Tuning for Text-to-Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–11.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR* abs/2312.11805 (2023).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*. MIT Press, 5998–6008.

[34] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6439–6448.

[35] Junyan Wang, Peng Zhang, Cheng Zhang, and Dawei Song. 2019. SCSS-LIE: A Novel Synchronous Collaborative Search System with a Live Interactive Engine. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1309–1312.

[36] Zixiao Wang, Hongtao Xie, Yuxin Wang, Jianjun Xu, Boqiang Zhang, and Yongdong Zhang. 2023. Symmetrical Linguistic Feature Distillation with CLIP for Scene Text Recognition. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 509–518.

[37] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1369–1378.

[38] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. 2024. Self-Training Boosted Multi-Factor Matching Network for Composed Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 5 (2024), 3665–3678.

[39] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-Guided Composed Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 915–923.

[40] Zhoufutu Wen, Xinyu Zhao, Zhipeng Jin, Yi Yang, Wei Jia, Xiaodong Chen, Shuanglong Li, and Lin Liu. 2023. Enhancing Dynamic Image Advertising with Vision-Language Pre-training. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 3310–3314.

[41] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11307–11317.

[42] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, and Maarten de Rijke. 2020. Modeling User Behavior for Vertical Search: Images, Apps and Products. In *Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval*. ACM, 2440–2443.

[43] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-Modal Transformer With Global-Local Alignment for Composed Query Image Retrieval. *IEEE Transactions on Multimedia* 25 (2023), 8346–8357.

[44] Qu Yang, Mang Ye, Zhaohui Cai, Kehua Su, and Bo Du. 2023. Composed Image Retrieval via Cross Relation Network With Hierarchical Aggregation Transformer. *IEEE Transactions on Image Processing* 32 (2023), 4543–4554.

[45] Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, and Liqiang Nie. 2020. Generative Attribute Manipulation Scheme for Flexible Fashion Search. In *Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval*. ACM, 941–950.

[46] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video Moment Retrieval With Cross-Modal Neural Architecture Search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.

[47] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Turning a CLIP Model into a Scene Text Detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6978–6988.

[48] Feifei Zhang, Ming Yan, Ji Zhang, and Changsheng Xu. 2022. Comprehensive Relationship Reasoning for Composed Query Based Image Retrieval. In *Proceedings of the International ACM Conference on Multimedia*. ACM, 4655–4664.

[49] Gangjian Zhang, Shikui Wei, Huaxin Pang, Shuang Qiu, and Yao Zhao. 2022. Composed Image Retrieval via Explicit Erasure and Replenishment With Semantic Alignment. *IEEE Transactions on Image Processing* 31 (2022), 5976–5988.

[50] Xin Zhang, Wen Xie, Ziqi Dai, Jun Rao, Haokun Wen, Xuan Luo, Meishan Zhang, and Min Zhang. 2023. Finetuning Language Models for Multimodal Question Answering. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 9420–9424.

[51] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1012–1021.

[52] Xiaoyang Zheng, Fuyu Lv, Zilong Wang, Qingwen Liu, and Xiaoyi Zeng. 2023. Delving into E-Commerce Product Retrieval with Vision-Language Pre-training. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 3385–3389.

[53] Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. AMC: Adaptive Multi-Expert Collaborative Network for Text-Guided Image Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 6 (2023), 1–22.