Dual-path Multimodal Optimal Transport for Composed Image Retrieval

Cairong Yan[1[0000−0003−0313−8833]], Meng Ma[1], Yanting Zhang[1[0000−0001−6317−1956)]], and Yongquan Wan[2[0000−0002−6911−0852]]

[1] Donghua University, Shanghai 200051, China cryan@dhu.edu.cn
[2] Shanghai Jian Qiao University, Shanghai 201306, China

Abstract. Unlike cross-modal retrieval tasks like text-to-image and image-to-text, which focus on one-way feature alignment, composed image retrieval emphasizes bidirectional feature alignment to differentiate between features that need to be preserved or modified. Existing methods usually map text and image modal data directly into a shared space for fusion, overlooking the issue of mismatched feature distributions between the source and target domains, resulting in limited retrieval performance. This paper presents the Dual-path Multimodal Optimal Transport (DMOT) model for composed image retrieval. It aligns features independently from both the text-to-image and image-to-text paths. During the fusion process, it explicitly calculates the preserved and modified features. Specifically, the pre-trained vision-language model BLIP is used to extract deep semantic features of both images and text. Then, we utilize two optimal transport modules to iteratively optimize and solve mapping matrices, aligning reference image features and modified text features in their respective spaces. Finally, considering the characteristics of the composed image retrieval task, we design the feature modifier module and the feature preserver module to handle the fusion of multimodal features. Extensive experiments on two public datasets, FashionIQ and CIRR, demonstrate DMOT's superior performance over state-of-the-art methods in retrieval accuracy, achieving an average improvement of 6.81% and 16.09%, respectively. The source code is available at https://github.com/AnoAuth/DMOT.

Keywords: Composed Image Retrieval · Optimal Transport · Multimodal Retrieval

1 Introduction

Image retrieval [1] aims to retrieve the most relevant images from a large-scale image gallery that match the user's intent. Traditional unimodal-based image retrieval allows users to search for target images using a single modality, such as image or text, which can sometimes fail to accurately and fully describe the user's intent. In image-based retrieval, low-quality images and redundant information like background and noise make it difficult for the model to identify the most important features to the user. Text-based retrieval may suffer from

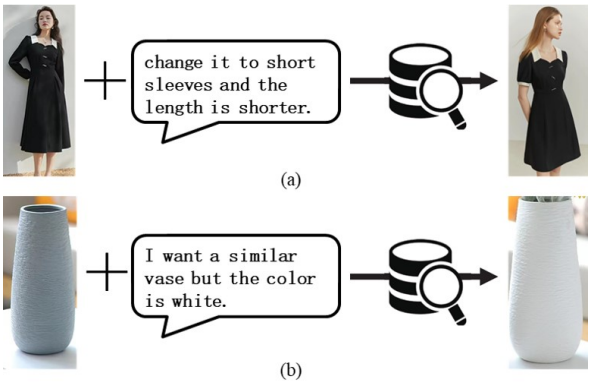

Fig. 1. Two examples of composed image retrieval (CIR). The input involves a reference image and modification text, and the output is the target image.

linguistic diversity and ambiguity. In contrast, composed image retrieval (CIR) combines both image and text modalities to specify the user's intent. The reference image provides general cues about the target image, while the text is used to express modifications to the reference image rather than describing its content. By using both modalities, users can flexibly describe the target image. Two CIR examples are provided in Fig. 1. The input of CIR is a multimodal query, consisting of a reference image and modification text. The retrieved target image needs to incorporate the specified modifications according to the given text while preserving the content of the reference image in other aspects. Due to its significant commercial value in applications like e-commerce product search, CIR has gained increasing research attention in recent years.

Existing approaches on CIR are devoted to using some classical backbones to embed the multimodal inputs and designing various mechanisms to combine multimodal features, including gated residuals [2], multilayer attention [3], and uncertainty learning [4], etc. The produced feature maps serve as unified representations of multimodal queries. However, due to the huge semantic gap between data from different modalities, these features are usually embedded in heterogeneous spaces. Directly mapping multimodal features to a joint space may overlook the differences between modalities, lose modality-specific information, and fail to fully utilize inter-modal correlation information, affecting the performance of subsequent tasks. To address this issue, we propose the Dual-path Optimal Transport (DMOT) method for aligning multimodal features, aiming to overcome the limitations of traditional methods and achieve more effective multimodal data fusion. We leverage the optimal transport algorithm to establish correlations between different modal data, compute the optimal transport plan, and utilize the mapping relationships from the transport plan to map features of one modality data to the feature space of another modality data by minimizing costs, thus achieving effective alignment of features across modalities. Moreover, unlike unidirectional feature alignment from text-to-image and image-to-text,

we design two optimal transport(OT) modules to achieve bidirectional feature alignment. It better adapts to CIR tasks by distinguishing which features need to be preserved and which need to be modified.

Another challenge in CIR is selectively modifying and preserving visual content based on textual information. For example, in the second example of Fig. 1, when the text specifies changing the color of the vase in the reference image, we are supposed to change it into the desired color according to the request while keeping other characteristics, such as shape, material, and texture, unchanged. To tackle this, we design a feature modifier (FM) module and a feature preserver (FP) module to fuse multimodal features. The FM module creates new features from the existing ones, enabling textual semantic modification of the reference image. The FP module decides which parts of the reference image should be preserved. The outputs of the two modules are combined to obtain a composite query representation. We experimentally validate our proposed method on two real-world public datasets, and the results demonstrate that our method achieves superior performance compared to previous methods, showing clear advantages in the CIR task. Our contributions can be summarized as follows:

- We propose a novel Dual-path Multimodal Optimal Transport model (DMOT) for CIR tasks. This model leverages two optimal transport modules to address cross-modal feature alignment from both text-to-image and image-to-text directions. By minimizing costs, it facilitates the mapping of multimodal features into a unified space, guided by weights. This process enhances the effectiveness of representation while aiding in distinguishing where to preserve and where to modify.
- We design the feature modifier (FM) module and feature preserver (FP) module to tackle the challenge of heterogeneous feature fusion. These modules are specifically utilized to explicitly compute preserved and modified features, ensuring precise control over the fusion process.
- We demonstrate the effectiveness of our model on two real-world image-text retrieval benchmark datasets: FashionIQ [5] and CIRR [6]. The performance surpasses that of previous research results.

2 RELATED WORK

2.1 Composed Image Retrieval

Early CIR methods relied on predefined attribute descriptions, modifying only fixed attributes when altering reference images. For instance, ASEN [7] introduces an attribute-specific embedding network to learn embedding spaces for multiple specific attributes, enabling fine-grained fashion similarity prediction. While these approaches yielded promising results, restricting user intent to predefined attributes lacked flexibility and failed to address personalized modification needs in certain scenarios.

To overcome these limitations, recent methods have explored more sophisticated techniques. TIRG [2] proposes composed image retrieval, employing a

reference image and modification text as inputs to generate a target image as the composite query representation. MAAF [3] utilizes an attention mechanism to improve retrieval relevance, while MUR [4] employs dual uncertainty learning modules to handle both coarse and fine-grained retrieval. AMC [8] designs sophisticated architectures involving gating mechanisms, residual connections, and multiple fusion nodes. These complex designs aim to handle multimodal inputs effectively but often introduce challenges in accurately differentiating between features that need to be preserved or modified. DCNet [9] uses synthesis and correction networks for bidirectional mapping of triplets. CLVC-Net [10] incorporates affine transformation-based synthesis modules with a mutual enhancement module for knowledge sharing. Although these methods enhance certain aspects of retrieval, they overlook the issue of mismatched feature distributions between source and target domains. This complexity can also lead to increased computational overhead and difficulty in model generalization. CLIP4CIR [11] leverages the pre-trained CLIP model for feature extraction and proposes a simple fusion network. Despite the power of pre-trained models, this approach primarily focuses on feature fusion without thoroughly addressing the complexities of aligning features from different modalities.

Though these methods have contributed to the progress in CIR, they predominantly focus on increasingly complex networks or feature fusion strategies without adequately addressing the critical issue of integrating heterogeneous features effectively. This limitation highlights the necessity for more robust solutions that tackle both feature alignment and integration challenges.

2.2 Optimal Transport

Optimal transport (OT) is a method used to measure the similarity between two probability distributions. It quantifies the difference between them by determining the minimum cost required to move mass from one distribution to another. OT finds wide applications in various fields such as image processing, natural language processing, and statistics. M3DN [12] proposes a multimodal, multi-instance, multi-label deep network based on optimal transport for complex object classification tasks. ADMM [13] applies optimal transport to neural decoding applications, predicting the movement direction of neuronal populations in the primary motor cortex of macaques. MOTCat [14] introduces a Co-Attention Transformer framework based on multimodal optimal transport with global structural consistency for Survival Prediction tasks. MOTKD [15] presents a multimodal optimal transport knowledge distillation method for cross-domain recommendation, utilizing optimal transport to achieve cross-domain knowledge transport. MuLOT [16] utilizes optimal transport for cross-modal correspondence to address multimodal sarcasm and humor detection from conversational videos and image-text pairs. GOT [17] uses optimal transport to graph cross-domain alignment. Given the successful applications of optimal transport in the aforementioned diverse tasks, this paper introduces OT into the CIR task to address the problem of heterogeneous feature alignment.

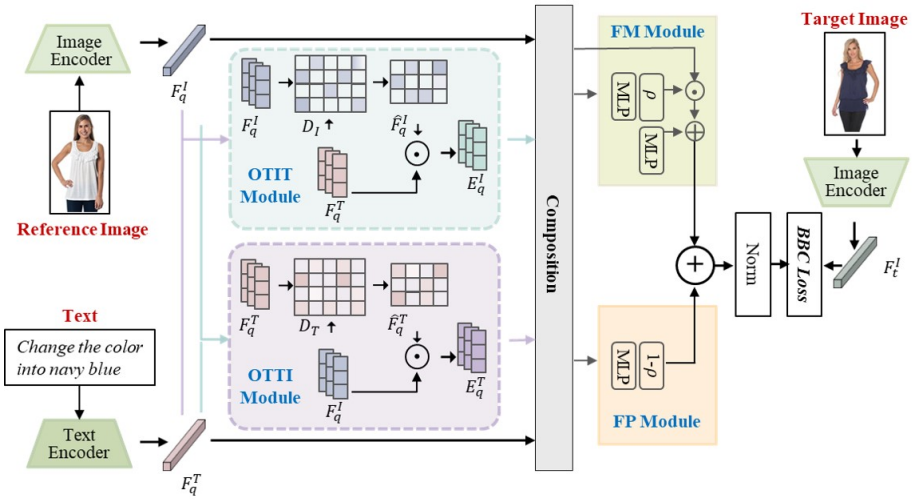

Fig. 2. The architecture of the proposed model DMOT. It mainly contains four modules: the image-guided optimal transport (IOT) aligning the text modality to the image modality space, the text-guided optimal transport (TOT) aligning the image modality to the text modality space, the feature modifier (FM), and the feature preserver (FP).

3 METHOD

3.1 Problem Definition

In this work, our objective is to address the CIR problem, which involves retrieving the corresponding target image from a database given a multimodal query consisting of a reference image and modification text. Assume that we have a triplet input, including a reference image I_q, a modification text sentence T_q, and a target image I_t. It is denoted as $T = \{(I_q, T_q, I_t)_i\}_{i=1}^{N}$, where N is the total number of triples. We utilize an image encoder to extract the visual features F_q^I for I_q, the target features F_t^I for I_t, and a text encoder to extract the text features F_q^T for T_q. The same encoder and encoding rules are used for both the reference and target images. Our goal is to design and optimize a multimodal composition method that combines the reference image features F_q^I and modification text features F_q^T into a composite feature F_q. This method maps the multimodal query features F_q^I to a unified metric space so that F_q is as close as possible to the corresponding target image F_t^I (the positive sample), and far away from negative samples. The definition is as follows,

$$F_q = G(F_q^I, F_q^T) \rightarrow F_t^I, \tag{1}$$

where G represents the multimodal feature composition method.

3.2 The Proposed Model DMOT

Fig. 2 provides an overview of our model framework. Our model primarily consists of three key components: (a) image and text encoders, (b) optimal transport, which includes two modules: the image-guided optimal transport (IOT) aligning the text modality to the image modality space and the text-guided optimal transport (TOT) aligning the image modality to the text modality space, and (c) feature fusion, which includes two parts: the feature modifier (FM) module and the feature preserver (FP) module. The first component extracts the features of the reference/target images and modification text. The second component focuses on the cross-modal alignment of the reference image and modification text features. The purpose of the third component is to address the multimodal feature fusion problem; specifically, the former achieves the modification of the reference image according to textual semantics, while the latter determines which parts of the reference image should be preserved without modification, thereby realizing the CIR task. Now, we explain each component of our model in detail.

Visual and Textual Representation. We first obtain the features of the reference image, the modification text, and the target image. Specifically, we adopt the language-image pre-training model BLIP [18] as the backbone for attribute feature extraction. BLIP is a new VLP framework designed for effective multi-task pre-training and flexible transfer learning. It is based on a multimodal hybrid encoder-decoder architecture, achieving superior performance on multiple vision-language downstream tasks such as image captioning, visual question answering, visual reasoning, and visual dialogue. Moreover, BLIP is trained using large-scale noisy image-text paired datasets, which enables better feature extraction for images and text in CIR datasets. We represent the features as follows,

$$F_q^I = \text{BLIP}(I_q), F_q^T = \text{BLIP}(T_q), F_t^I = \text{BLIP}(I_t), \tag{2}$$

where $F_q^I, F_q^T, F_t^I \in R^L$. To obtain better performance of the pre-trained model on the downstream CIR datasets, we fine-tune the text encoder before training. During fine-tuning, the extracted image features and text features are element-wise added, and the multimodal combination is compared with the target image embedding through cosine similarity. Due to cost constraints, we do not fine-tune the image encoder.

Multimodal Alignment. As mentioned before, We design two Optimal Transport (OT) modules: the image-guided optimal transport (IOT) and the text-guided optimal transport (TOT). Although both modules utilize optimal transport, their functions and the type of information they process differ significantly. IOT translates textual modifications into the visual domain, ensuring that the semantic integrity of the reference image is preserved while incorporating the

modifications suggested by the text. Conversely, TOT reflects visual modifications in the resulting image features, maintaining the semantic integrity of the text while integrating visual elements from the reference image.

Taking IOT module as an example, the OT alignment strategy involves the following steps. First, based on the feature representations of different modal data, we construct a cost matrix, which measures the distance between different features $F_q^I[j]$ and $F_q^T[k]$, denoted as $Cost[k, j]$. We use the Wasserstein Distance with the 2-norm here. Formally, the cost representations can be obtained as follows,

$$Cost[k, j] = ||F_q^T[k] - F_q^I[j]||_2^2, \tag{3}$$

where $\forall j \in [m]$, $\forall k \in [n]$, and m, n are the dimensions of F_q^I and F_q^T, respectively. Now we estimate the distribution i of the image embedding F_q^I and the distribution t of the text embedding F_q^T. Then compute the discrete distributions of the image embedding and the text embedding based on i and t, which are denoted by ϕ_q^I, ϕ_q^T, respectively.

$$\phi_q^I = \sum_{j=1}^{m} i_j \delta_{F_j^I}, \tag{4}$$

$$\phi_q^T = \sum_{k=1}^{n} t_k \delta_{F_k^T}, \tag{5}$$

where δ is the Dirac function. We set $i_j = 1/m$, $t_k = 1/n$.

Then we denote the joint distribution as follows,

$$\Gamma(\phi_q^T, \phi_q^I) = \{D \in R_+^{n \times m} | D1_m = t, D^T 1_n = i\}, \tag{6}$$

where $\Gamma(\phi_q^T, \phi_q^I)$ represents the joint distribution and 1 denotes an all-one vector. Further, to measure the joint probability of observing ϕ_q^I and ϕ_q^T, we compute the optimal transport matrix D between the two modal data through the optimal transport algorithm, where $D[k, j]$ gives the probability of transmitting the k-th dimension of F_q^T to the j-th dimension of F_q^I. The OT coupling D is computed as follows,

$$W(\phi_q^T, \phi_q^I) = \min_{D \in \Gamma(\phi_q^T, \phi_q^I)} \sum_{k=1}^{n} \sum_{j=1}^{m} D[k, j] Cost[k, j], \tag{7}$$

where $W(\phi_q^T, \phi_q^I)$ measures the distribution discrepancy between ϕ_q^T and ϕ_q^I.

Based on the computed optimal transport plan D, we get the weight of the linguistic embedding $\hat{F_q^T}$ as follows,

$$\hat{F_q^T} = diag(1/\phi_q^I) D^T + \Delta D, \tag{8}$$

where D^T denotes the transpose of matrix D, and ΔD is an adjustable parameter. And then the linguistic embedding F_q^T is mapped into the feature space

of the image F_q^I and transformed into a new target-aligned embedding E_q^T as follows,

$$E_q^T = \hat{F_q^T} \bigodot F_q^I, \tag{9}$$

where $\bigodot$ represents the Hadamard element-wise product. By this step, the text embedding E_q^T is located in the same space as the image embedding F_q^I. Similarly, we can obtain the image embedding E_q^I aligned with the text embedding through the TOT module by the above steps.

Multimodal Fusion. In this branch, we first input the extracted image and text features into multilayer perception for preliminary processing to obtain the intermediate representations ϕ^I and ϕ^T of the image and text.

$$\phi^I = \rho(\partial(W_{i1} * F_q^I)), \tag{10}$$

$$\phi^T = \rho(\partial(W_{t1} * F_q^T)). \tag{11}$$

Here, ∂ represents the ReLU function. ρ represents the dropout layer, which is used to reduce overfitting and improve model stability and generalization ability. * denotes the Fully Connected (FC) layer, and W_{i1}, W_{t1} are the learnable parameters. We do the same for E_q^I, E_q^T obtained from the above steps to obtain the intermediate representation.

$$\xi^I = \rho(\partial(W_{i2} * E_q^I)), \tag{12}$$

$$\xi^T = \rho(\partial(W_{t2} * E_q^T)). \tag{13}$$

Next we connect the intermediate representations ϕ^T, ξ^T, ϕ^I, ξ^I as a composition F_C.

$$F_C = concat[\phi^T, \xi^I, \phi^I, \xi^T]. \tag{14}$$

Then we achieve modification and preservation through two modules.

Feature Preserver (FP) Module. First, we project F_c through the linear layer and the ReLU function. The projected features are then concatenated and fed into two different multilayer perceptron (MLP) components with very similar structures. The first MLP component is dedicated to generating a mixture of the learned reference image and modification text features, while the second MLP component generates retention coefficients to learn multimodal query combinations:

$$\psi_c = W_{12} * \rho(\partial(W_{11} * F_c)), \tag{15}$$

$$\psi_p = w_i W_{22} * \rho(\partial(W_{21} * F_c)), \tag{16}$$

where W_{11}, W_{12}, W_{21}, W_{22} are the learnable parameters, and w_i is a learnable weight. ψ_c is the mixture of multimodal features, and ψ_p is the learned retention coefficients. Next, we use ψ_c, ψ_p to get the part of the reference image that should be retained by calculating.

$$F_p = \psi_c + \psi_p \bigodot F_q^I, \tag{17}$$

where F_p is the feature that should be retained in the reference image.

Feature Modifier (FM) Module. Derive the modification coefficient ψ_m from the retained coefficient ψ_p generated in FP:

$$\psi_m = (1 - w_i)\psi_p. \tag{18}$$

Further computation yields the feature F_m which is modified according to the corresponding semantics of the modified text.

$$F_m = \psi_m \bigodot F_q^T. \tag{19}$$

The output of the FP module is added with the output of the FM module and regularized to get the final query representation F_q.

$$F_q = Norm(F_p + F_m). \tag{20}$$

Loss Function. To measure the degree of error between the model's predictions and ground truths, making the model-generated multimodal query representation converge towards the ground truth image, i.e., maximizing the matching degree between the multimodal query representation and the target image while distancing it from mismatched images in the potential space, we follow previous work and adopt the batch-based classification (BBC) loss as the basic optimization loss for metric learning. Given a training batch size B for the query, where $G(F_q^I{}^{-p}, F_q^T{}^{-p})$ is the final modified representation from the last layer of the network for the p-th image-text query pair. Only the ground truth target image $F_t^I{}^{-p}$ of that query is considered as a positive sample, while other target images $F_t^I{}^{-j}$ in the batch are treated as negative samples, where j is not p. The loss is computed as follows:

$$Loss = 1/B \sum_{p=1}^{B} -log\{\frac{exp\{\kappa(G(F_q^I{}^{-p}, F_q^T{}^{-p}), F_t^I{}^{-p})/\tau\}}{\sum_{j=1}^{B} exp\{\kappa(G(F_q^I{}^{-p}, F_q^T{}^{-p}), F_t^I{}^{-j})/\tau\}}\}, \tag{21}$$

where τ is the temperature parameter, and $\kappa(,)$ is an arbitrary similarity kernel implemented as cosine similarity in our work.

4 EXPERIMENT

4.1 Experimental Settings

Datasets. We evaluate our approach on two public datasets: FashionIQ, a fashion domain dataset, and CIRR, an open-domain dataset. FashionIQ is a natural language-based interactive fashion retrieval dataset comprising 77,684 fashion product images across three categories gathered from Amazon.com: Dress, Toptee, and Shirt. The training set comprises 46,609 images, including 18,000 training triplets consisting of a reference image, a pair of related captions, and a target image. The validation set comprises 15,537 images and 6,017 triplets,

while the test set consists of 15,538 images and 6,119 triplets. CIRR is sampled from the NLVR2 dataset and includes 21,552 real-life images, with 36,554 randomly assigned triplets. 80% of these triplets are utilized for training, 10% for validation, and 10% for testing. The images in the dataset are grouped into six subsets that are semantically and visually similar, with reference-target pairs drawn from these subsets. Compared to other datasets, CIRR places more emphasis on distinguishing visually similar images and enables evaluation on fully labeled subsets, addressing the issue of false negatives due to incomplete labeling.

Evaluation metric. Following previous work, we use Recall@k (R@k for short) to report our results, which indicates the percentage of queries for which the true target is ranked among the top k candidates. For FashionIQ, we report R@10 and 50 to evaluate the performance for each category, as well as the average results across all three categories to reflect the overall performance. For CIRR, we report R@1, 5, 10, and 50, and the subset metrics R_{subset}@1, 2, and 3, which only consider images within the subset and are unaffected by false negative samples. Among these metrics, R@5 reflects the potential false negative issue in the corpus, while R_{subset}@1 captures fine-grained image-text modification scenarios. Therefore, we also report the average of R@5 and R_{subset}@1.

Table 1. Quantitative comparison on the FashionIQ dataset with R@k(%). The best results are in bold, and the second-best results are underlined. The last row indicates the performance improvements of DMOT relative to the best baseline.

Method	Dress		Shirt		Topee		Average		Avg
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
TIRG [2]	14.87	34.66	18.62	37.89	19.08	39.62	17.40	37.39	27.40
CIRPLANT [6]	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
DCNet [9]	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	40.84
CoSMo [19]	26.45	52.43	26.94	52.99	31.95	62.09	28.45	55.84	39.45
MCR [20]	26.20	51.20	22.40	46.00	29.70	56.40	26.10	51.20	38.65
CLVC-Net [10]	29.85	56.47	28.75	54.76	33.50	64.00	30.07	58.41	44.24
FashionVLP [21]	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	48.39
MAAF [3]	23.80	48.60	21.30	44.20	27.90	53.60	24.30	48.80	36.55
ARTEMIS [22]	27.16	52.40	21.78	54.83	29.20	43.64	26.05	50.29	38.17
AACL [23]	29.89	55.85	24.82	48.85	30.88	56.85	28.53	53.85	41.18
AMC [8]	31.73	59.25	30.67	59.08	36.21	66.60	32.87	61.64	47.25
MUR [4]	32.61	61.34	33.23	<u>62.55</u>	41.40	**72.51**	35.75	65.47	50.61
Css-Net [24]	33.65	63.16	35.96	61.96	<u>42.65</u>	70.70	37.42	65.27	51.35
CLIP4CIR [11]	33.81	59.40	<u>39.99</u>	60.45	41.41	65.37	<u>38.32</u>	61.74	50.03
DFN [25]	<u>35.59</u>	<u>63.51</u>	34.49	62.51	42.22	<u>71.03</u>	37.43	<u>65.68</u>	<u>51.56</u>
DMOT(ours)	**41.52**	**67.37**	**40.78**	**63.69**	**46.62**	70.45	**42.97**	**67.17**	**55.07**
Impr.(%)	+16.66	+6.08	+1.98	+1.82	+9.31	-2.84	+12.13	+2.27	+6.81

Baselines. To validate the effectiveness of our approach in composed image retrieval, we compare our method with the following methods on the FashionIQ

dataset: TIRG [2], CIRPLANT [6], DCNet [9], CoSMo [19], MCR [20], CLVC-Net [10], FashionVLP [21], MAAF [3], ARTEMIS [22], AACL [23], AMC [8], MUR [4], Css-Net [24], CLIP4CIR [11], and DFN [25]. CIRR is a newly released dataset, and only a few baseline results are available. We report the following representative state-of-the-art methods as baselines: TIRG [2], MAAF [3], CIRPLANT [6], ARTEMIS [22], CLIP4CIR [11], RUTIR [26].

Implementation details. Our code is based on the PyTorch deep learning framework and all experiments are conducted with an NVIDIA 3090 GPU. We adopt the pre-trained BLIP encoder as the feature extractor. Following [18], when fine-tuning the text encoder, we optimize with AdamW for 20 epochs with a learning rate of 5×10^{-5}, a weight decay of 0.05, and employ a cosine learning rate schedule. During training, we train and optimize the model with AdamW for 100 epochs, setting the base learning rate to 2×10^{-5}, the text and image feature dimensions to 256, the temperature factor τ in the equation to 0.01, and the dropout layer parameter to 0.5. When experimenting on the FashionIQ dataset, we set the batch size to 512. However, due to GPU limitations, we halve the batch size on CIRR.

Table 2. Quantitative comparison on the CIRR dataset with R@k(%) and R_{subset}@k(%). We follow [6] to report Avg. as (R@5+R_{subset}@1)/2. The best results are in bold, and the second-best results are underlined. The last row indicates the performance improvements of DMOT relative to the best baseline.

Method	R@k				R_{subset}@k			Avg
	k=1	k=5	k=10	k=50	k=1	k=2	k=3	
TIRG [2]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
TIRG+LastConv [2]	11.04	35.68	51.27	83.29	23.82	45.65	64.55	29.75
MAAF [3]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
MAAF+BERT [3]	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
MAAF-IT [3]	9.90	32.86	48.83	80.27	21.17	42.04	60.91	27.02
MAAF-RP [3]	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.37
CIRPLANT [6]	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
ARTEMIS [22]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CLIP4CIR [11]	<u>33.59</u>	65.35	77.35	95.21	<u>62.39</u>	<u>81.81</u>	<u>92.02</u>	63.87
RUTIR [26]	32.24	<u>66.63</u>	<u>79.23</u>	<u>96.43</u>	61.25	81.33	<u>92.02</u>	<u>63.94</u>
DMOT(ours)	**41.55**	**74.25**	**84.67**	**96.49**	**74.20**	**89.55**	**95.58**	**73.70**
Impr.(%)	+23.70	+11.44	+6.87	+0.06	+18.93	+9.46	+3.87	+16.09

4.2 Performance Comparison

Results on FashionIQ. Table 1 presents our results on the Fashion IQ dataset. It can be observed that our model achieves an overall average R@10 and R@50 of 42.97% and 67.17%, respectively, with improvements of approximately 12.13%

and 2.27%. Our approach outperforms state-of-the-art methods and surpasses the most competitive method by an average of 6.81%.

Results on CIRR. Table 2 presents the performance comparison between our method and other approaches on the CIRR dataset. Compared to other methods, ours surpasses the previous state-of-the-art and achieves a significant performance boost. Particularly, notable performance advantages of our method can be observed when k is relatively small. Our model achieves R@1 and R@5 of 41.55% and 74.25%, respectively, and reaches 74.20% on R_{subset}@1. Compared to the previous best method, our approach enhances the performance by 23.70%, 11.44%, and 18.93% in terms of R@1, R@5, and R_{subset}@1, respectively. These results confirm that our model not only performs well on fashion datasets but also demonstrates excellent performance on open-domain datasets, further validating the effectiveness of our approach.

Table 3. Ablation studies on FashionIQ and CIRR.

Method	FashionIQ(Avg)		CIRR	
	R@10	R@50	R@5	R_{subset}@1
w/o TOT	42.43	66.73	73.89	72.97
w/o IOT	42.17	66.35	73.36	72.53
w/o OT_both	42.03	65.96	73.06	72.36
w/o FP	40.75	65.22	72.43	71.29
w/o FM	41.48	66.19	73.08	72.61
w/o FP&FM	39.17	63.62	69.92	68.36
DMOT(ours)	**42.97**	**67.17**	**74.25**	**74.20**

4.3 Ablation Studies

To help analyze our proposed method and gain insights into the role of each module, we design several variants of the model and conduct ablation studies. For a fair comparison and analysis, all ablation studies are conducted under the same experimental settings. The different experimental variables are as follows:

- w/o TOT, w/o IOT, and w/o OT_both: To investigate the effectiveness of optimal transport, we remove IOT, TOT, and both modules simultaneously. From the results in Table 3, we observe that removing either OT module leads to a performance drop, demonstrating the effectiveness of OT. Furthermore, we find that removing the IOT module has a greater impact on performance compared to removing the TOT module. We hypothesize that this is because removing IOT may result in text features not being effectively represented in the image feature space, preventing accurate matching of key

Fig. 3. Qualitative results on FashionIQ. We show reference images on the left and top-10 retrievals with descending scores on the right. Ground-truths are highlighted with green boxes.

textual information with image features and thus affecting retrieval performance. On the other hand, removing the TOT module may have a smaller impact on the representation of image features and may not significantly affect the matching of key textual information, as image features typically possess richer structural information.

- w/o FP, w/o FM, and w/o FP&FM: As shown in Table 3, our model has two modules FP and FM for combining multimodal features. The fourth to sixth rows demonstrate that removing either module leads to a performance drop. These observations highlight the complementary roles of the FP and FM modules in our DMOT model. The FP module ensures the preservation of key features from the reference image, while the FM module effectively modifies features according to the text input. Together, they enable the DMOT model to achieve superior performance by balancing the preservation and modification of features during the fusion process.

4.4 Qualitative Results

Fig. 3 presents sample retrievals, illustrating some quantitative results of our model on the FashionIQ dataset. Each row of the figure includes the reference image, the text, and the top ten retrieval results, with the target image highlighted by a green box. As depicted in the figure, the modification text sentences in FashionIQ are complex and often contain multiple modification requests simultaneously, such as color, decoration, and style. We observe that our model effectively captures both concrete and abstract modification requirements from the text. For instance, in the second row, "longer" and "dark velvet" specify length, color, and material, while "sexier" in the fourth row and "casual" in the fifth row represent abstract style concepts. This highlights our model's ability to comprehend high-level semantic information. Furthermore, our model seamlessly

integrates multiple modifications from the text descriptions, adjusting certain visual content in the reference image based on textual cues to retrieve the target image. The retrieved images display a certain degree of visual similarity, suggesting that our method not only retrieves the target image but also offers similar candidates for users to consider. Though this may not perfectly match strict evaluation metrics, it represents a more realistic and practical approach.

5 CONCLUSION

This paper introduces a Dual-path Multimodal Optimal Transport model (DMOT) for CIR tasks. Our approach focuses on facilitating knowledge transport between the source and target domains through two optimal transport modules designed to construct feature alignment. Additionally, we employ an FM module and an FP module to efficiently fuse the multimodal features by explicitly calculating the modified and preserved features. Extensive experiments on two real-world datasets validate the effectiveness of our proposed method. Ablation studies further confirm the impact of each designed module. Future work includes exploring model optimization from multiple directions and investigating the extension and application of this method.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China (No. 62477006 and 62206046).

References

1. Ma, C., Gu, C., Li, W., Cui, S.: Large-scale image retrieval with sparse binary projections. In: SIGIR. pp. 1817-1820 (2020)
2. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L. J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: CVPR. pp. 6439-6448 (2019)
3. Dodds, E., Culpepper, J., Herdade, S., Zhang, Y., Boakye, K.: Modality-agnostic attention fusion for visual search with text feedback. arXiv preprint arXiv:2007.00145 (2020)
4. Chen, Y., Zheng, Z., Ji, W., Qu, L., Chua, T. S.: Composed image retrieval with text feedback via multi-grained uncertainty regularization. arXiv preprint arXiv:2211.07394 (2022)
5. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion iq: A new dataset towards retrieving images by natural language feedback. In: CVPR. pp. 11307-11317 (2021)
6. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV. pp. 2125-2134 (2021)
7. Ma, Z., Dong, J., Long, Z., Zhang, Y., He, Y., Xue, H., Ji, S.: Fine-grained fashion similarity learning by attribute-specific embedding network. In: AAAI. pp. 11741-11748 (2020)
8. Zhu, H., Wei, Y., Zhao, Y., Zhang, C., Huang, S.: AMC: Adaptive multi-expert collaborative network for text-guided image retrieval. ACM Transactions on Multimedia Computing, Communications and Applications **19**(6), 1-22 (2023)

9. Kim, J., Yu, Y., Kim, H., Kim, G.: Dual compositional learning in interactive image retrieval. In: AAAI. pp. 1771-1779 (2021)
10. Wen, H., Song, X., Yang, X., Zhan, Y., Nie, L.: Comprehensive linguistic-visual composition network for image retrieval. In: SIGIR. pp. 1369-1378 (2021)
11. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: CVPR. pp. 21466-21474 (2022)
12. Yang, Y., Wu, Y. F., Zhan, D. C., Liu, Z. B., Jiang, Y.: Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In: SIGKDD. pp. 2594-2603 (2018)
13. Lee, J., Dabagia, M., Dyer, E., Rozell, C.: Hierarchical optimal transport for multimodal distribution alignment. NeurIPS **32** (2019)
14. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: ICCV. pp. 21241-21251 (2023)
15. Yang, W., Yang, J., Liu, Y.: Multimodal Optimal Transport Knowledge Distillation for Cross-domain Recommendation. In: CIKM. pp. 2959-2968 (2023)
16. Pramanick, S., Roy, A., Patel, V. M.: Multimodal learning using optimal transport for sarcasm and humor detection. In: WACV. pp. 3930-3940 (2022)
17. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: ICML. pp. 1542-1553. PMLR (2020)
18. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888-12900. PMLR (2022)
19. Lee, S., Kim, D., Han, B.: CoSMo: Content-style modulation for image retrieval with text feedback. In: CVPR. pp. 802-812 (2021)
20. Zhang, G., Wei, S., Pang, H., Zhao, Y.: Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In: ACMMM, pp. 5353-5362 (2021)
21. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: CVPR. pp. 14105-14115 (2022)
22. Delmas, G., Rezende, R. S., Csurka, G., Larlus, D.: ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In: ICLR. (2021)
23. Tian, Y., Newsam, S., Boakye, K.: Fashion image retrieval with text feedback by additive attention compositional learning. In: WACV. pp. 1011-1021 (2023)
24. Zhang, X., Zheng, Z., Wang, X., Yang, Y.: Relieving Triplet Ambiguity: Consensus Network for Language-Guided Image Retrieval. arXiv preprint arXiv:2306.02092 (2023)
25. Li, H., Wu, Y., Wang, F.: Dynamic network for language-based fashion retrieval. In: Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval, pp. 49-57 (2023)
26. Chen, J., Lai, H.: Ranking-aware Uncertainty for Text-guided Image Retrieval. arXiv preprint arXiv:2308.08131 (2023)