CF-TCIR: A Compositor-Free Framework for Hierarchical Text-Conditioned Image Retrieval

Yuchen Yang[1,2], **Yu Wang**[2,3]✉, **Yanfeng Wang**[2,3]

[1]University of Science and Technology of China [2]Shanghai AI Laboratory [3]Shanghai JiaoTong University

Abstract

In text-conditioned image retrieval (TCIR), the combination of a reference image and modification text forms a query tuple, aiming to locate the most congruent target image within a dataset. The advantages of rich image semantic information and text flexibility are combined in this manner for more accurate retrieval. While traditional techniques often employ attention-driven compositors to craft a unified image-text representation, our paper introduces a compositor-free framework, CF-TCIR, which eschews the standard compositor. Compositor-based methods are designed to learn a joint representation of images and text, but they struggle to directly capture the correlations between attributes across the image and text modalities. Instead, we reformulate the retrieval process as a cross-modal interaction between a synthesized image feature and its corresponding text descriptor. This novel methodology offers advantages in terms of computational efficiency, scalability, and superior performance. To optimize the retrieval performance, we advocate a tiered retrieval mechanism, blending both coarse-grain and fine-grain paradigms. Moreover, to enrich the contextual relationship within the query tuple, we integrate a generative cross-modal alignment technique, ensuring synchronization of sequential attributes between image and text data.

1 Introduction

Text-conditioned image retrieval (Vo et al., 2019) (Lee et al., 2021) (Wen et al., 2021) (Yang et al., 2021) makes the retrieval system more accurate and flexible by allowing the user to enter both a reference image and a text description. The text description, also known as the modification text, describes adjustments to the attributes or layout of the reference image. Previous works typically fuse the reference image and modification text representations into a joint image-text representation by learning an attention-based compositor. The joint image-text representation can be directly compared with the representation of any potential target image.

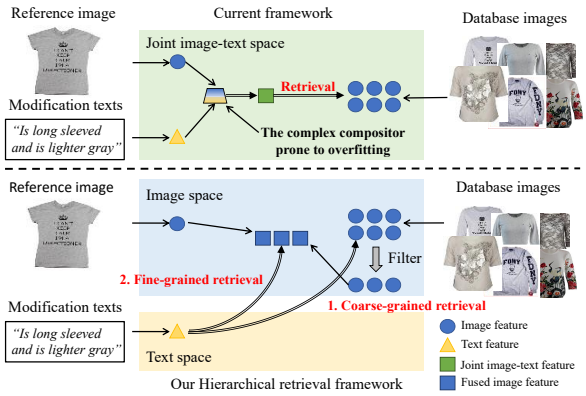

Figure 1: A comparison of our framework with current framework. The current framework utilizes a complex compositor to synthesize the image and text semantics, which is prone to overfitting. In contrast, we discard the compositor module and reformulate the retrieval pipeline as a cross-modal interaction between the fused image features and the text feature, which captures cross-modal attribute correlations more clearly.

However, existing compositor-based methods do not take full advantage of the cross-modal alignment. As shown in Fig. 1, In the task of text-conditioned image retrieval, the modification text serves as complementary information to the reference image. Specifically, textual features should emphasize areas of the image that semantically diverge from the text. In contrast, multi-modal attention-based models exhibit a tendency to allocate higher attention values to text and image features that demonstrate semantic congruence. In light of this contradiction, compositor-based (attention-based) methods attempt to aggregate these distinct semantics by learning a joint image-text representation. However, in doing so, they overlook the attribute correlations between

✉: Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 16315–16325

the text and image domains. Consequently, these compositor-based methods are prone to overfitting within the dataset and generally perform suboptimally when applied to out-of-domain scenarios.

Besides, while it is straightforward to learn a joint image-text representation in existing compositor-based methods, this does not take full advantage of the image modality and the text modality. In the text-conditioned image retrieval task, both the reference image and the modification text describe part of the attributes of the target image. Thus, both the reference image and the modification text can be used to filter out partially dissimilar target images, which cannot be exploited by using the joint image-text representation directly.

To alleviate the above issues, we propose a compositor-free framework for hierarchical text-conditioned image retrieval (CF-TCIR). In contrast to previous works, we discard the compositor module and reformulate the retrieval pipeline as a cross-modal interaction between a synthesized image feature and its corresponding text descriptor. Our method focuses on extracting differential information from images prior to aligning them through image-text attribute correlations. Our method offers distinct advantages given that the majority of modification texts in the TCIR task consist of combinations of attributes, with a lot of similar modification texts associated with different images. Therefore, our method, which centers on image-text mapping for alignment, is both a more rational and efficient choice.

Besides, in order to fully leverage the advantages of each modality and improve the efficiency of retrieval, we use a hierarchical retrieval method by combining both coarse-grained and fine-grained retrieval. Coarse-grained retrieval leverages the similarity between target image features and text feature to filter out obviously irrelevant target images. Fine-grained retrieval takes the target images retrieved by coarse-grained retrieval, calculates the corresponding fused image features with the feature of the reference image, and then returns a ranklist in order of similarity to the text feature. Our hierarchical retrieval method not only reduces noise but also greatly improves retrieval efficiency.

To further capture the context information between the modification text, the reference image, and the target image, we propose a generative cross-modal alignment module. Specifically, we first map the features of both the reference image and the target image to sequence image features. Then we capture the cross-modal alignment between sequence image features and sequence text features through the GPT-2 model. By performing generative cross-modal alignment, our model further explores the intrinsic relationship between the query tuple and the target image.

Overall, our method achieves competitive performance while greatly reducing model parameters and increasing retrieval efficiency. We demonstrate that there is no need to design a complex compositor and only several fully connected layers are required to achieve superior performance for the text-conditioned image retrieval.

2 Related Work

2.1 Image Retrieval

Traditional content-based image retrieval (Radenović et al., 2018) (Ng et al., 2020) (Revaud et al., 2019) (Gordo et al., 2017) (Teichmann et al., 2019) takes a single image as input and extracts global or local features for retrieval. Although query images contain rich semantic information, it is difficult to accurately grasp the retrieval intention. This often leads to the intention gap problem.

The most common retrieval scenario involving text is cross-modal retrieval, which uses a textual description of the image as a query. The standard approach focuses on mapping different modalities into a common space to mitigate the domain gap (Chen et al., 2020a) (Kuang et al., 2019) (Edwards et al., 2021) (Fei et al., 2021) (Li et al., 2023) (Wu et al., 2021) (Zhan et al., 2020) (Han et al., 2023). However, the retrieval intention expressed by a single modality is still not enough to handle all scenarios.

To take advantage of the rich semantic information of the image and the simplicity and flexibility of the text, TIRG (Vo et al., 2019) first proposes the text-conditioned image retrieval task. In this setting, the query tuple is specified in the form of a reference image with a modification text that describes desired modifications to the reference image. Previous works devote to learning a joint expression of vision-language to retrieve the potential target images. VAL (Chen et al., 2020b) uses multi-scale techniques to learn the composition of image and text at both low and high semantic levels. LBF (Hosseinzadeh and Wang, 2020) explores the bidirectional correlation between the image domain and text domain. JGAN (Zhang et al., 2020) uses GCN (Kipf and Welling, 2017) to inject semantic

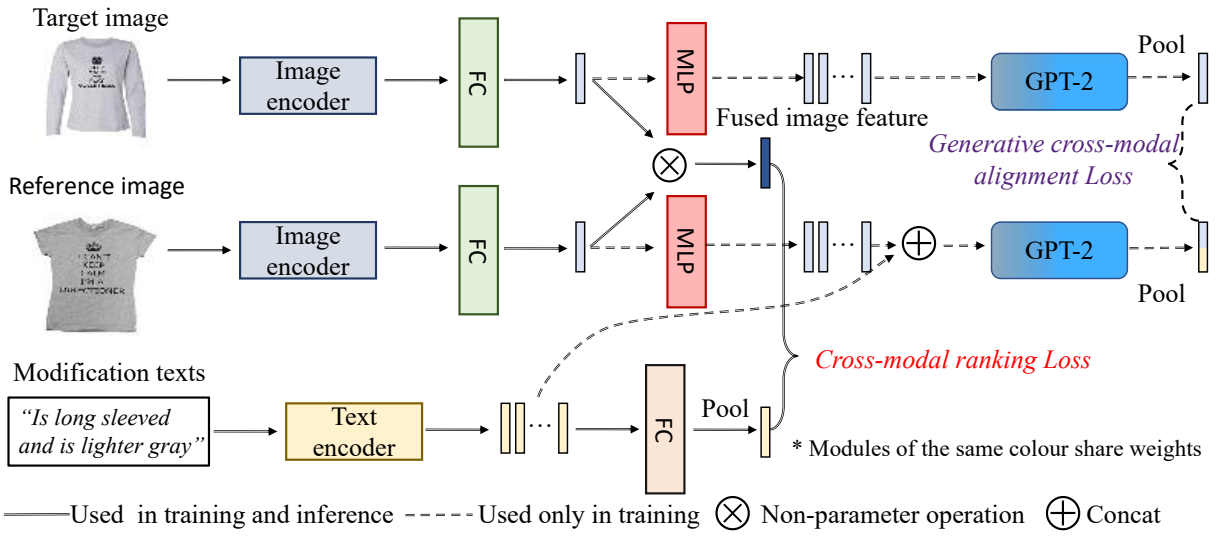

Figure 2: An overview of our framework. Our model takes the query tuple (the reference image and the modification text) and the target image as input and feeds each element into its respective encoder. For accurate and efficient cross-modal retrieval, the fused image feature is computed directly from the semantic correlation between the query image and the reference image, without the need for a fusion network. To further capture the context information and explore the intrinsic relationship between the query tuple and the target image, we build a generative cross-modal alignment loss function using the GPT-2 model during the training stage.

information from textual feature into the visual feature. MAAF (Dodds et al., 2020) feeds the spatial image features and text embeddings as modality-agnostic tokens into a Transformer to learn the joint representation. ComposeAE (Anwaar et al., 2021) propose an autoencoder-based model and use a deep metric learning approach to learn the composition of image and text features. CoSMo (Lee et al., 2021) modulates the content and style information of reference image to generate robust image-text representation. DCNet (Kim et al., 2021) leverages both the local and global reference image features for image-text composition. ARTEMIS (Delmas et al., 2022) combines the Explicit Matching module and the Implicit Similarity module, each focusing on one of the modalities of the query. CLIP4Cir (Baldrati et al., 2022) proposes a combiner network equipped with the CLIP model to learn the joint image-text representation. FashionVLP (Goenka et al., 2022) leverages prior knowledge from large image-text corpora and multiple image features, proposing a novel attention-based approach for learning joint image-text representations.

In contrast to previous works focusing on the design of image-text compositors, we discard the compositor module and reformulate the retrieval pipeline as a cross-modal interaction between a synthesized image feature and its corresponding text descriptor. In this way, the cross-modal attribute correlations are more clearly captured.

3 Methodology

Text-conditioned image retrieval uses a reference image and a modification text as a query tuple to retrieve the target image in the database that satisfies the semantics of both the reference image and the modification text. Let (I_q, M, I_t) represent the reference image, the modification text, and a candidate target image, respectively. Most existing methods for the text-conditioned image retrieval aim to learn a joint representation $f(I_q, M)$ which is similar to the representation $f_{target}(I_t)$. In our method, we reformulate the retrieval pipeline and perform cross-modal retrieval directly using the image features of I_q, I_t and the text features of M.

In the following, we start by introducing our framework pipeline in Sec. 3.1. Then we elaborate our generative cross-modal alignment in Sec. 3.2 and the details of the training and inference procedures in Sec. 3.3 and Sec. 3.4.

3.1 Framework

Figure. 2 shows an overview of our framework. We build a lightweight model with only several fully connected layers in addition to the off-the-shelf image and text encoders. This makes our model

highly scalable while demonstrating the soundness of our model through its simple yet effective structure. Our model is divided into two parts, namely the main module and the generative cross-modal alignment module, labeled with solid and dashed lines respectively. Remarkably, the main module is used for both the training and inference stages, while the generative cross-modal alignment module is used for the training stage only. We will introduce the main module in this section and the generative cross-modal alignment module in the next section.

We first extract the reference image feature $V_q \in R^D$ and the target image feature $V_t \in R^D$ using the image encoder f_{img} and a fully connected layer:

$$V_q = FC(f_{img}(I_q)), \tag{1}$$

$$V_t = FC(f_{img}(I_t)). \tag{2}$$

In our implementation, we use ResNeXT-101 on IG (WSL (Mahajan et al., 2018)), which is followed by a pooling layer as f_{img}. Then, we extract the sequence feature of modification text $T_{seq} \in R^{L \times C}$ using the text encoder f_{text}:

$$T_{seq} = f_{text}(M). \tag{3}$$

In our implementation, we use BERT (Kenton and Toutanova, 2019) as f_{text}. To facilitate the similarity calculation with the image features, we obtain the pooled text features $T_{pool} \in R^D$ as follows:

$$T_{pool} = Pool(FC(T_{seq})), \tag{4}$$

where $Pool$ is implemented as GPO (Chen et al., 2021), a well-known pooling function in cross-modal retrieval.

To obtain the fused image feature based on the semantic correlation between the reference image feature V_q and the target image feature V_t, we design a non-parameter operation to fuse them. We do not build a fusion network for V_q and V_t because coupling the V_q and V_t in the network would make retrieval inefficient. By designing a reasonable non-parameter operation to fuse V_q and V_t, not only can V_q and V_t be decoupled from the network thus greatly improving the retrieval efficiency, but also competitive performance can be achieved. The fused image feature $V_f \in R^D$ is obtained as follows:

$$attn = 1 - sigmoid(V_q \odot V_t), \tag{5}$$

$$V_f = attn \odot V_t, \tag{6}$$

where $\odot$ refers to element-wise product. Our insight is that the fused image feature V_f should be matched to the modification text feature T_{pool}, *i.e.* the semantic content in the target image that differs from the reference image should be highlighted. Each channel in the feature reflects how well it fits a certain pattern, so by multiplying the channels of V_q and V_t, the channel attention that expresses the correlation between V_q and V_t is calculated. Thus, the reversed channel attention calculated by $1 - sigmoid(V_q \cdot V_t)$ will amplify channels in V_t that are not similar to V_q and suppress channels that are similar to V_q. Since the design of non-parameter operation has a significant impact on experimental results, we will provide more analysis in the appendix.

3.2 Generative Cross-modal Alignment

To further capture context information and explore the intrinsic relationship between I_q, I_t, and M, we propose a generative cross-modal alignment loss function. First, we map both the reference image feature V_q and the target image feature V_t to sequence features $V_{qseq} \in R^{N \times C}$ and $V_{tseq} \in R^{N \times C}$, where N is a hyper-parameter indicating the sequence length of the image sequence features:

$$V_{qseq} = MLP(V_q), \tag{7}$$

$$V_{tseq} = MLP(V_t), \tag{8}$$

where MLP is implemented as a cascade of two fully connected layers. We then concatenate the reference image sequence feature V_{qseq} with the text sequence feature T_{seq} to obtain the query sequence feature $Q_{seq} \in R^{(L+N) \times C}$:

$$Q_{seq} = [V_{qseq} \oplus T_{seq}], \tag{9}$$

where $\oplus$ indicates concatenate operation. We then feed V_{tseq} and Q_{seq} into the GPT-2 model separately and adapt the output of the last hidden state of the GPT-2 model as the output feature. The rationale for this procedure is to effectively capture the context information pertaining to the reference image and the modification text. By using GPT-2 to map the image sequence feature and query sequence feature to the text domain, we not only mitigate the domain gap but also further explore the intrinsic relationship between the query tuple and the target image. To formulate the generative cross-modal alignment function, we pass the output of the last hidden state of the GPT-2 model

through the GPO pooling layer (Chen et al., 2021) to obtain the target image feature T_{gpt} and query tuple feature Q_{gpt}, respectively.

Notably, in addition to using GPT-2, we also investigated the feasibility of using randomly initialized transformers as an alternative. However, the efficacy of this substitute method pales in comparison to that of GPT-2.

3.3 Training Procedures

Training. In the training stage, the whole framework is trained with the cross-modal ranking loss and our proposed generative cross-modal alignment loss. Given a training minibatch B containing K triplets, each triplet consists of (I_{q_i}, M_i, I_{t_i}), which represents the i-th reference image, modification text, and corresponding target image, respectively. Remarkably, all features used below are normalized before feeding into the loss function.

Cross-modal Ranking Loss. For ease of expression, we use $V_{f_{ij}}$ to represent the fused image feature for the reference image I_{q_i} and a candidate target image I_{t_j}, which should be similar with the modification text feature T_{pool_i} when j equals i. Following TIRG, we consider the batch-based classification loss as the ranking loss. The batch-based classification loss takes into account all negative samples in a mini-batch and learns from both easy and difficult negative samples. The batch-based classification loss is formulated as follows:

$$L_{rank} = \frac{1}{K}\sum_{i=1}^{K} -log\{\frac{\kappa(V_{f_{ii}}, T_{pool_i})}{\sum_j^K \kappa(V_{f_{ij}}, T_{pool_i})}\}, \tag{10}$$

where κ is an arbitrary similarity kernel function. We implement κ as the dot product similar to previous works.

Generative Cross-modal Alignment Loss. We aim to capture the context information between image sequence features and text sequence features through generative cross-modal alignment loss. And by aligning the features of the query tuple and target images output by GPT-2, our model can better capture the intrinsic relationship between them.

For ease of expression, we use Q_{gpt_i} to represent the query tuple feature for the reference image I_{q_i} and the modification text I_{t_i}, which should be similar with the target image feature T_{gpt_i}. The generative cross-modal alignment loss is formulated as follows:

$$L_{align} = \frac{1}{K}\sum_{i=1}^{K} -log\{|\frac{\kappa(Q_{gpt_i}, T_{gpt_i})}{\sum_j^K \kappa(Q_{gpt_i}, T_{gpt_j})}|\}, \tag{11}$$

where κ is implemented as the dot product. The generative cross-modal alignment loss is only enabled during the training stage. As the GPT-2 model is frozen, the generative alignment loss function encourages mapping the image feature and the text feature to the same domain and learning more robust and discriminative image and text features.

The overall loss for training is formulated as follows:

$$L = L_{rank} + \lambda * L_{align} \tag{12}$$

where λ is a learnable parameter and initialized to 1.

3.4 Hierarchical Retrieval Stage

In the inference stage, we propose a hierarchical search strategy by combining both coarse-grained retrieval and fine-grained retrieval.

Coarse-grained Retrieval. In each query iteration, the user inputs a query tuple consisting of a reference image and a modification text. As the modification text is a partial description of the target image, our model first extracts the normalized modification text feature T_{pool} to filter out obviously dissimilar target images by computing dot product with each normalized target image feature V_t. We then pass the filtered target images with the query tuple to fine-grained retrieval. The filtering threshold is a hyper-parameter that will be analyzed in our ablation study.

Fine-grained Retrieval. For each filtered target image, we fuse them with the reference image to obtain fused image features V_f as mentioned above. As the fused image features capture semantic discrepancy between the reference image and target images, we retrieve the fused image features using the modification text feature T_{pool} and return a ranklist based on the similarity score.

4 Experiments

To verify the effectiveness of our method, we conduct experiments on three benchmarks on fashion domain including FashionIQ (Guo et al., 2019), Shoes (Berg et al., 2010) and Fashion200k (Han et al., 2017). In this section, we introduce the implementation details, the experimental results on

different datasets and ablation studies in Sec. 4.1, Sec. 4.2 and Sec. 4.3, respectively.

4.1 Implementation Details

Our model is a lightweight model, consisting of only the image encoder, the text encoder and several fully connected layers. We conduct the experiments in Pytorch (Paszke et al., 2019). For image encoder, we adopt the output from layer 4 of the backbone networks followed by a GEM pooling (Radenović et al., 2018) as image feature. The both image encoder and text encoder is followed by a single linear layer that project the output feature to a 1024-dimensional vector. To map the image features to the image sequence features, the MLP we used consisting of two linear layers with hidden sizes of $\frac{L}{2} * 768$ and $L * 768$ (L indicates the sequence length and is set to 10), respectively.

In the training stage, we use the Adam optimizer with a base learning rate of 0.0001, which decays once after 10 epochs by a factor of 10 and the batch size K is set to 32. And the GPT-2 model is freezed during training.

4.2 Experimental Results

FashionIQ Dataset. FashionIQ is a natural language-based interactive fashion product retrieval dataset. It contains 77,684 images, covering three categories: Dress, Toptee and Shirt. There are 18,000 image pairs in the 46,609 training images and each training tuple consisting of a reference image and two relative captions produced by two different human annotators.

Table 1 shows our results on FashionIQ. It is observed from the table that our simple but effective method achieves a competitive performance with FashionVLP that uses additional side information. Our proposed method obtain a 3.61% performance improvement in terms of the AvgRecall@50 metric compared to CLIP4Cir, which use the powerful CLIP model as the image and text encoder. This proves that our reformulated retrieval pipeline is more natural and achieves better performance.

Shoes Dataset. The Shoes dataset is originally proposed for attribute discovery. It consists of 10,000 training queries and 4,658 validation examples. Guo et al. (Guo et al., 2018) tagged the images with captions in natural language for fashion image retrieval.

According to Table 2, our method outperforms ARTEMIS (Delmas et al., 2022) on Recall@1 and Recall@10, respectively. This further validates our motivation that reformulated retrieval pipeline is effective.

Fashion200K Dataset. Fashion200K is a diverse dataset consisting of about 200K clothes images of various styles. It contains around 172k images for training and 33,480 test queries for evaluation. Each image is equipped with some tags describing attributes. Notably, the modification text of this dataset is automatically generated rather than human-annotated. To be specific, when generating training triplets, if the tags of two fashion images differ by one word, we choose them as the reference image and the target image, and the modification text is formulated as "change A to B".

As shown in Table 3, we achieve competitive results with ComposeAE. In the experiments, we found that Fashion200K is a sensitive dataset and that small changes in parameters or settings can have a significant impact on the results. For a fair comparison, we repeat each experiment five times and report the mean results. The performance of this dataset differs obviously from the other datasets. We believe the reason for that is the modification text in the Fashion200K dataset is automatically generated with only two meaningful words. This makes cross-modal alignment more dependent on the correspondence of attributes between the image domain and the text domain. The experimental results demonstrate the good generalization ability of our method, which consistently achieves good results on all three datasets.

4.3 Ablation Studies

In this subsection, we conduct ablation studies to analyze the effect of the design of non-parameter operation, the generative cross-modal alignment module and the proportion of coarse retrieval. For a fair comparison, we conduct experiments on FashionIQ and use the same evaluation metric as before.

Effect of the Design of Non-parameter Operation. In our method, we design a non-parameter operation to obtain fused image feature V_f based on the semantic correlation between the reference image feature V_q and the target image feature V_t, as in Eq. (5) and (6). We now study variants of this design. For ease of the following discussion, we use V_{f_1}-V_{f_4} to denote the original design and these variants, where V_{f_1} corresponds to the original design.

$$V_{f_1} = (1 - sigmoid(V_q \cdot V_t)) \cdot V_t, \qquad (13)$$

Method	Dress		Toptee		Shirt		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
TIRG (Vo et al., 2019)	14.87	34.66	19.08	39.62	18.26	37.89	17.40	37.39
TIRG† (WSL+BERT)	29.11	57.24	32.09	60.74	27.88	53.13	29.69	57.03
VAL (Chen et al., 2020b)	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04
MAAF (Dodds et al., 2020)	23.80	48.60	27.90	53.60	21.30	44.20	24.30	48.80
RTIC (Shin et al., 2021)	27.37	52.95	27.33	53.60	22.03	45.29	25.58	50.61
RTIC-GCN (Shin et al., 2021)	27.71	53.50	29.63	56.30	22.72	44.16	26.69	51.32
ComposeAE (Anwaar et al., 2021)	11.99	31.38	11.01	27.48	11.04	26.49	11.34	28.45
TRACE (Jandial et al., 2020)	26.13	52.10	31.16	59.05	26.20	50.93	27.83	54.02
TRACE w/ BERT (Jandial et al., 2020)	26.52	51.01	32.70	61.23	28.02	51.86	29.08	54.70
CIRR (Liu et al., 2021)	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53
CoSMo (Lee et al., 2021)	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31
CoSMo† (WSL+BERT)	29.20	56.48	33.15	63.90	28.47	53.89	30.27	58.09
DCNet (Kim et al., 2021)	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89
CLVC-Net (Wen et al., 2021)	29.85	56.47	33.50	64.00	28.75	54.76	30.70	58.41
ARTEMIS (Delmas et al., 2022)	27.16	52.40	29.20	54.83	21.78	43.64	26.05	50.29
CLIP4Cir (Baldrati et al., 2022)	31.63	56.67	38.19	62.42	**36.36**	58.00	**35.39**	59.03
FashionVLP⋆ (Goenka et al., 2022)	**32.42**	**60.29**	**38.51**	**68.79**	31.89	**58.44**	34.27	**62.51**
AACL (Tian et al., 2023)	24.82	48.85	30.88	56.85	29.89	55.85	28.53	53.85
CF-TCIR	**31.97**	**58.37**	**40.64**	**70.00**	**32.39**	**59.56**	**35.00**	**62.64**

Table 1: Retrieval performance on the FashionIQ official validation set under VAL evaluation protocols. † means our re-implementation. ⋆ denotes the use of additional side information (*e.g.* landmark detection) during training. The "Avg" column refers to the average results on three categories. Overall 1^{st}/2^{nd} in **black**/**blue**

Method	Shoes		
	R@1	R@10	R@50
TIRG (Vo et al., 2019)	7.89	26.53	51.05
VAL (Chen et al., 2020b)	16.49	49.12	73.53
RTIC (Shin et al., 2021)	–	43.66	72.11
RTIC-GCN (Shin et al., 2021)	–	43.38	72.09
ComposeAE (Anwaar et al., 2021)	3.46	20.84	52.58
TRACE (Jandial et al., 2020)	18.11	52.41	75.42
CoSMo (Lee et al., 2021)	16.72	48.36	75.64
DCNet (Kim et al., 2021)	–	**53.82**	79.33
CLVC-Net (Wen et al., 2021)	17.64	**54.39**	**79.47**
ARTEMIS (Delmas et al., 2022)	**18.72**	53.11	79.31
FashionVLP⋆ (Goenka et al., 2022)	–	49.08	77.32
CF-TCIR	**18.95**	53.17	**79.59**

Table 2: Retrieval performance on the Shoes dataset. ⋆ denotes the use of additional side information during training. Overall 1^{st}/2^{nd} in **black**/**blue**

Method	Fashion200k		
	R@1	R@10	R@50
TIRG (Vo et al., 2019)	14.1	42.5	63.8
JGAN (Zhang et al., 2020)	17.3	45.2	65.7
LBF (Hosseinzadeh and Wang, 2020)	17.8	48.4	68.5
VAL (Chen et al., 2020b)	21.2	49.0	68.8
MAAF (Dodds et al., 2020)	18.9	–	–
ComposeAE (Anwaar et al., 2021)	22.8	**55.3**	**73.4**
CoSMo (Lee et al., 2021)	**23.3**	50.4	69.3
DCNet (Kim et al., 2021)	–	46.9	67.6
CLVC-Net (Wen et al., 2021)	22.6	**53.0**	72.2
FashionVLP⋆ (Goenka et al., 2022)	–	49.9	70.5
AACL (Tian et al., 2023)	19.64	52.3	71.0
CF-TCIR	**23.5**	52.7	**72.5**

Table 3: Retrieval performance on the Fashion200K dataset. ⋆ denotes the use of additional side information during training. Overall 1^{st}/2^{nd} in **black**/**blue**

$$V_{f_2} = (1-sigmoid(V_q \cdot V_t)) \cdot V_t + sigmoid(V_q \cdot V_t) \cdot V_q, \quad (14)$$

$$V_{f_3} = 2(sigmoid(|(V_q - V_t)|) - 0.5) \cdot V_t, \quad (15)$$

$$V_{f_4} = ReLU(V_t) - ReLU(V_q), \quad (16)$$

Method	AvgR@10	AvgR@50
V_{f_4}	30.09	54.62
V_{f_3}	32.50	58.79
V_{f_2}	31.28	58.14
V_{f_1}	**35.00**	**62.64**

Table 4: Ablation study on effect of the design of non-parameter operation. We use V_{f_1}-V_{f_4} to denote the original design and these variants, where V_{f_1} corresponds to the original design

As shown in Table 3, the design of non-parameter operation matters a lot in our method. Our core idea is to amplify channels in V_t that are not similar to V_q and suppress channels that are similar to V_q. V_{f_2} additionally retains part of the V_q on top of V_{f_1}, and the performance drop 3.75% in terms of AvgR@10. Since there is no semantic overlap between modification texts and reference images in the FashionIQ dataset, this result meets our expectations. For V_{f_3} and V_{f_4}, performance decreased by 2.50% and 4.91% in terms of AvgR@10 respectively, which further shows the importance of choosing an appropriate non-parameter operation.

Effect of Generative Cross-modal Alignment Module. In our method, we propose a generative cross-modal alignment module to further capture context information and explore the intrinsic relationship between the reference image, the target image and the modification text. We make an ablation experiment to study on its impact.

Method	L_{align}	AvgR@10	AvgR@50
CF-TCIR	×	33.52	60.98
CF-TCIR	✓	**35.00**	**62.64**

Table 5: Ablation study on generative cross-modal alignment module. × denotes the absence of generative cross-modal alignment module while ✓ is opposite.

As shown in Table 5, the L_{align} consistently improves the performance of our model. This validates the effectiveness of our generative cross-modal alignment module.

Effect of the Proportion of Coarse Retrieval. As our hierarchical retrieval strategy consists of coarse-grained retrieval and fine-grained retrieval, the coarse-grained retrieval filters out obviously dissimilar target images and retains a certain proportion of the target images for fine-grained retrieval. We now investigate the effect of the proportion of retained target images on accuracy.

Proportion	AvgR@10	AvgR@50
1.0	34.57	62.20
0.8	34.42	61.98
0.5	34.76	62.15
0.3	34.89	62.33
0.1	34.85	62.39
0.05	**35.00**	**62.64**

Table 6: Ablation study on the proportion of coarse retrieval. The proportion column means the proportion of target images retained by coarse-grained retrieval.

As shown in Table 6, different coarse-retrieval proportions hardly affect retrieval accuracy. We finally set the proportion to 0.05, which maintains both accuracy and retrieval speed.

Attention Visualization.

In Figure 3, we provide some attention visualisation heatmaps by calculating the similarity of target image feature maps to reference image feature maps and modification text features. Specifically, we aodpt the image activations at the last convolutional layer of the third blocks of WSL (Mahajan et al., 2018) as image feature maps. The "img2img attention" denotes the heatmaps calculated from the reference image feature map to the target image feature map. The "img2txt attention" denotes the heatmaps calculated from the modification text feature to the target image feature map.

It appears obviously that the modification text focus on contents of the target image that correspond to specific attributes while the reference image focus more on the subject content in the target image and less on the detail differences.

Concretely, as shown in the "img2txt attention" column of Figure 1, when the modification text contains "short sleeves", the cuff of the target image is highlighted. And when the modification text contains "white background", the background areas of the target image are highlighted instead of the floral area. When the modification text contains "graphic", patterns in the target image are highlighted while other parts receive less attention. When the modification text contains "checkered button", the button and nearby areas in the target image receive more attention. Besides, as shown in the "img2img attention" column of Figure 1, all parts of the target image that are related to clothing are highlighted.

Since our insight is to distinguish the semantic information of the target and reference images from the channel weights and thus align the fused feature with the modification text, these visualization results meet our expectations.

5 Conclusions

We propose a compositor-free framework for hierarchical text-conditioned image retrieval (CF-TCIR). In contrast to previous works that design an attention-based compositor to learn the joint image-text representation, we discard the compositor module and reformulate a more natural and efficient retrieval pipeline. In our framework, we utilize the semantic correlation of the query image and the target image and then perform cross-modal interaction with the modification text feature. Through extensive experiments, it is demonstrated that our method achieves competitive performance, which can greatly reduce the model parameters and also increase retrieval efficiency.

6 Limitations

Our method is highly scalable and will be more dependent on the performance of the image and text encoders due to the small number of parameters. Therefore, it is necessary to choose powerful image

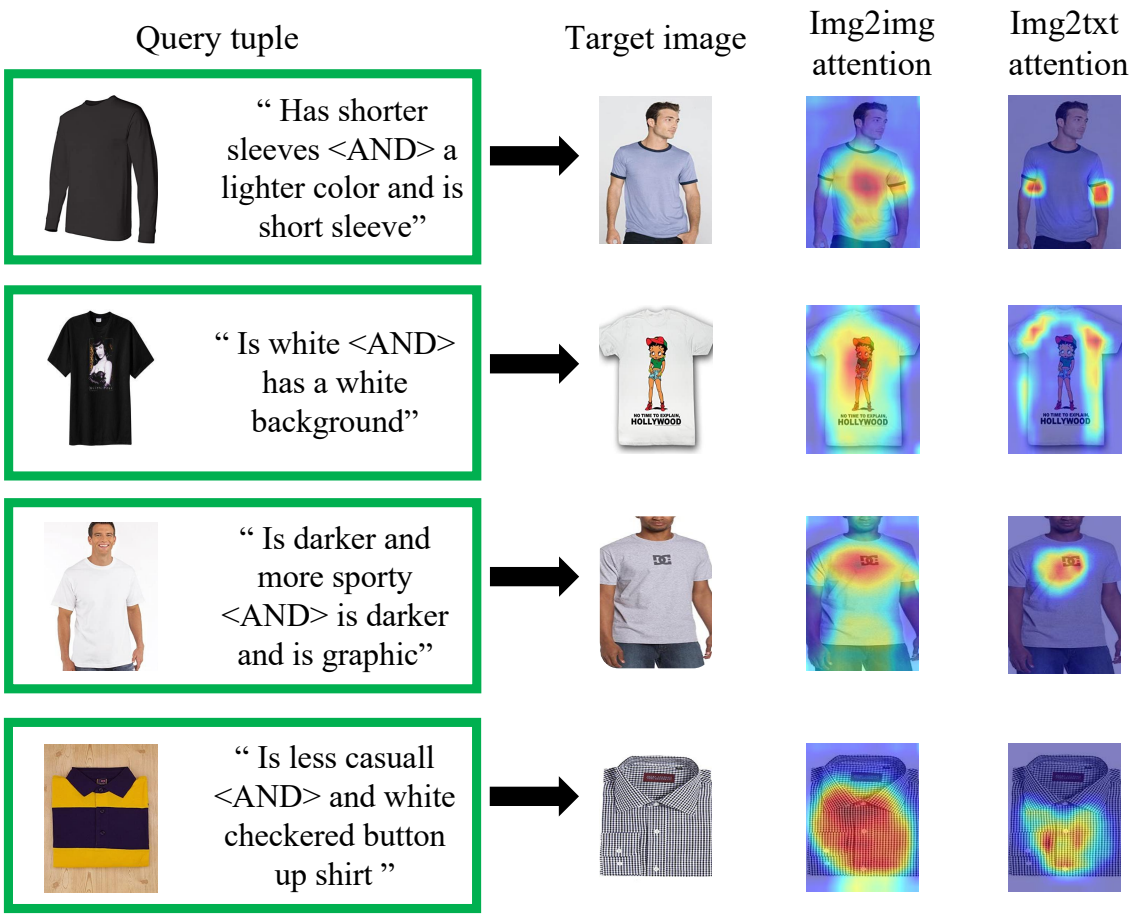

Figure 3: Attention visualization of our method. The “img2img attention” denotes the heatmaps calculated from the reference image to the target image. The “img2txt attention” denotes the heatmaps calculated from the modification text to the target image. It is observed that the modification text focus on contents of the target image that correspond to specific attributes while the reference image focus more on the subject content in the target image.

and text encoders with our CF-TCIR framework for real-world applications.

Acknowledgements

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), National Natural Science Foundation of China (No. 62106140) and STCSM (No. 21511101100, No. 22DZ2229005).

References

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1140–1149.

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/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474.

Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the European Conference on Computer Vision*, pages 663–676.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020a. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10638–10647.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020b. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3001–3011.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115.

Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. 2017. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, pages 237–254.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems*.

Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019. Fashion iq: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*.

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*, pages 1463–1471.

Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. 2023. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038.

Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3596–3605.

Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. *arXiv preprint arXiv:2009.01485*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*, pages 1–14.

Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. 2019. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3066–3075.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-sstyle modulation for image retrieval with text feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 802–812.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuan-Jing Huang, and Zhongyu Wei. 2023. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958.

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*, pages 2125–2134.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision*, pages 181–196.

Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. 2020. SOLAR: Second-order loss and attention for image retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 253–270.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.

Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116.

Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. 2021. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*.

Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. 2019. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5109–5118.

Yuxin Tian, Shawn Newsam, and Kofi Boakye. 2023. Fashion image retrieval with text feedback by additive attention compositional learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1011–1021.

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*, pages 6439–6448.

Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1369–1378.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738.

Yuchen Yang, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Cross-modal joint prediction and alignment for composed query image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3303–3311.

Yu-Wei Zhan, Xin Luo, Yongxin Wang, and Xin-Shun Xu. 2020. supervised hierarchical deep hashing for cross-modal retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3386–3394.

Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu. 2020. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3367–3376.