Dynamic Weighted Combiner for Mixed-Modal Image Retrieval

Fuxiang Huang, Lei Zhang*, Xiaowei Fu, Suqi Song

Learning Intelligence & Vision Essential (LiVE) Group
School of Microelectronics and Communication Engineering, Chongqing University, China
{huangfuxiang, leizhang, xwfu, songsuqi}@cqu.edu.cn

Abstract

Mixed-Modal Image Retrieval (MMIR) as a flexible search paradigm has attracted wide attention. However, previous approaches always achieve limited performance, due to two critical factors are seriously overlooked. 1) The contribution of image and text modalities is different, but incorrectly treated equally. 2) There exist inherent labeling noises in describing users' intentions with text in web datasets from diverse real-world scenarios, giving rise to overfitting. We propose a Dynamic Weighted Combiner (DWC) to tackle the above challenges, which includes three merits. First, we propose an Editable Modality De-equalizer (EMD) by taking into account the contribution disparity between modalities, containing two modality feature editors and an adaptive weighted combiner. Second, to alleviate labeling noises and data bias, we propose a dynamic soft-similarity label generator (SSG) to implicitly improve noisy supervision. Finally, to bridge modality gaps and facilitate similarity learning, we propose a CLIP-based mutual enhancement module alternately trained by a mixed-modality contrastive loss. Extensive experiments verify that our proposed model significantly outperforms state-of-the-art methods on real-world datasets. The source code is available at https://github.com/fuxianghuang1/DWC.

Introduction

Image retrieval (Su, Zhong, and Zhang 2019; Brown et al. 2020; Huang et al. 2020; Shen et al. 2020; Hu et al. 2021; Wei et al. 2020; Chun et al. 2021; Yang et al. 2021a; Huang, Zhang, and Gao 2021; Dubey 2022), as a crucial computer vision task, aims to search for items of interest from the database. A key limitation of traditional image retrieval is the in-feasibility to precisely describe users' intentions (i.e., the concepts in users' minds) through a single image or a single text. Therefore, to offer a flexible and intuitive user experience, a Mixed-Modal Image Retrieval (MMIR) paradigm as shown in Fig. 1 is explored, where the search intention is expressed with mixed modalities utilized to retrieve a target image. Therefore, MMIR requires a synergistic understanding of both visual and linguistic content, which is an acknowledged challenge.

Most existing approaches (Wen et al. 2021; Gu et al. 2021; Goenka et al. 2022; Huang and Zhang 2023) mainly focus on designing complex components to learn composite image-text representations. However, these methods usually achieve poor retrieval results, with almost no more than 20% of queries retrieving the correct image in the top-1 rank. *We cannot help asking why it happened*?

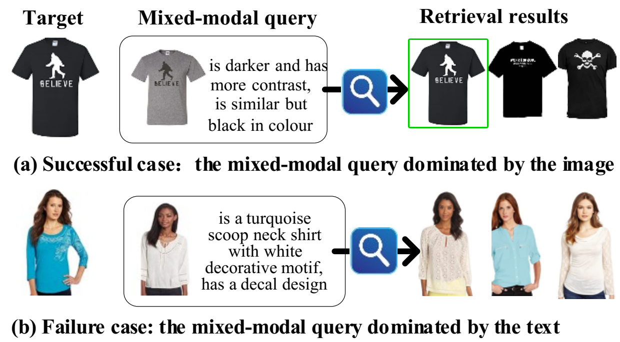


Figure 1: Overview of MMIR, i.e., image+text $\rightarrow$ image. The mixed-modal query is dominated by different modalities unequally. (a) Dominated by image modality. (b) Dominated by text modality.

Based on our observations and analysis, we have three critical findings, which are seriously overlooked. **First**, *the contribution of image and text modality differs in diverse real-world scenarios.* We observe that previous methods usually prefer to returning results similar to the query image modality, which fails when the intended image similar to the target does not exist. This is mainly due to that previous approaches depend heavily on the image modality but underestimate the text modality. Fig. 1 (a) fits this scenario, where the query image is very similar to the target image. However, in real-world scenarios and the existing datasets, the mixed-modal query is often dominated by the text modality, that is, the similarity between the query image and the target image is very low. For example, as shown in Fig. 1 (b), only query text and the concept of *shirt* in query image are conducive to retrieval, while most visual information is redundant or even detrimental. We conduct an exploratory experiment with different modality queries individually in Fig. 2 (a), which verifies modality importance disparity for different categories. **Second**, *the widely used datasets (Han et al. 2017; Guo et al. 2018; Wu et al. 2021)*

*Corresponding Author

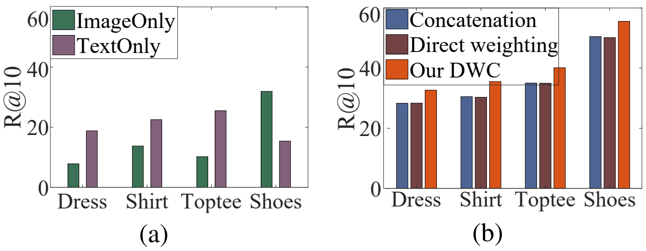


Figure 2: (a) Single-modal query. (b) Mixed-modal query with different composition methods. *Concatenation* means concatenating the image and text modality features. *Direct weighting* refers to (Zhao, Song, and Jin 2022).

collected from web are inherently noisy in labeling intention description with text and full of data bias. This is due to people from different states describe objects and concepts in distinct manners. We conduct an intriguing experiment in Table 1, in which different text descriptions (templates) show incredibly distinct performances. We observe that the previous SOTA TIRG model (Vo et al. 2019) shows significantly limited performance. This is due to noisy supervision from different templates and existing training objectives often overfit the noisy data. The relevance between the visual appearance and the text description can thus vary across diverse scenarios. **Last but not least**, *the inherent modality gaps affecting the synergistic understanding of multiple modalities are overlooked in training objectives*. Actually, the image-text modality gap makes feature combination challenging, while the mixed-modality gap makes similarity learning difficult.

To tackle the inherent modality importance disparity problem, an intuitive idea is directly assigning different weights for different modalities. We conduct an experiment as shown in Fig. 2 (b), by following (Zhao, Song, and Jin 2022), which uses direct weighting of image-only and text-only models. We find that direct weighting hardly improves the concatenation of two modality features. This is due to that this naive weighting does not delve deeper into the modality features, easily amplifying the error information implied in the dominant modality. Therefore, we suggest editing and purifying the features of different modalities to remove potential error information and dynamically weighting modalities. To tackle the inherent data bias and labeling noise, it is intuitive to enrich the training set with more diverse image and text descriptions, which, however, is labor-intensive. We suggest relaxing the noisy hard labels and exploring dynamic soft similarity labels to fully mine valuable information. Table 1 shows the potential of our proposal. Additionally, in training objectives, we propose to utilize large models (e.g., CLIP (Radford et al. 2021)) and enforce mutual enhancement training to bridge modality gaps.

Based on above critical findings and motivations, we propose a Dynamic Weighted Combiner (DWC), composed of an editable modality de-equalizer (EMD), a dynamic soft-similarity label generator (SSG), and a mixture modality-image modality contrastive loss. EMD contains two modality editors and an adaptive weighted combiner to purify modality features and unequally treat different modalities,

Text description	TIRG		Our model	
	R@1	R@10	R@1	R@10
Template 1	14.10	42.50	18.24	48.69
Template 2	18.10	52.40	33.95	60.83
Template 3 (ours)	24.50	49.60	36.49	63.58

Table 1: Impact of different text descriptions on MMIR performance, which unveils the biased labeling noise in the Fashion200k datasets.

such that modality importance disparity is amended. SSG aims to relax the biased hard labeling in text description by generating soft-similarity labels, which facilitates full utilization of valuable information in datasets, such that noisy supervision is improved. In order to bridge modality gaps and facilitate similarity learning, we introduce a CLIP-based mutual enhancement module alternately trained by a mixed-modality contrastive loss. Experiments on Fashion200K, Shoes, and FashionIQ datasets show the outstanding performances. The main contributions are as follows.

- We propose a Dynamic Weighted Combiner (DWC) to solve the inherent modality importance disparity, biased labeling noises, and modality gaps. Fig. 3 depicts the overall architecture.
- We introduce an Editable Modality De-equalizer (EMD), a dynamic soft-similarity label generator (SSG), and a CLIP-based mixed-modality contrastive training loss to meet the above challenges.

Related Work

Mixed-Modal Image Retrieval (MMIR) aims to incorporate a query image and text describing the users' intentions to navigate the visual search. Previous approaches for MMIR can be categorized into two types. First, (Vo et al. 2019; Yang et al. 2021b; Shin et al. 2021; Hou et al. 2021; Wen et al. 2021; Tian, Newsam, and Boakye 2022) focus on designing complex components for multi-modal fusion of text and image queries. (Wen et al. 2021) propose a CLVC-Net, which combines local-wise and global-wise composition modules. (Huang and Zhang 2023) propose a Language Guided Local Infiltration (LGLI), which consists of a LPVL module and a TILA module. (Huang et al. 2022) propose a plugged-and-played gradient augmentation (GA) module to improve the generalization of MMIR models. Second, (Nam, Kim, and Kim 2018; Tautkute and Trzcinski 2021) focus on generating images similar to the target. Tautkute et al. propose a SynthTripletGAN for image retrieval with synthetic query expansion.

Vision-and-Language Pre-training (VLP) (Radford et al. 2021; Zhang et al. 2021; Li et al. 2021; Dou et al. 2022) aims to learn multi-modal representations from large-scale image-text pairs, which has proven to be highly effective on various downstream tasks. Recently, VLP models attract attention to solve the MMIR problem. (Zhao, Song, and Jin 2022) propose a three-stage progressive learning method based on CLIP to acquire complex knowledge progressively, and fully exploit the open-domain and open-format resources. (Baldrati et al. 2021, 2022b,a) propose

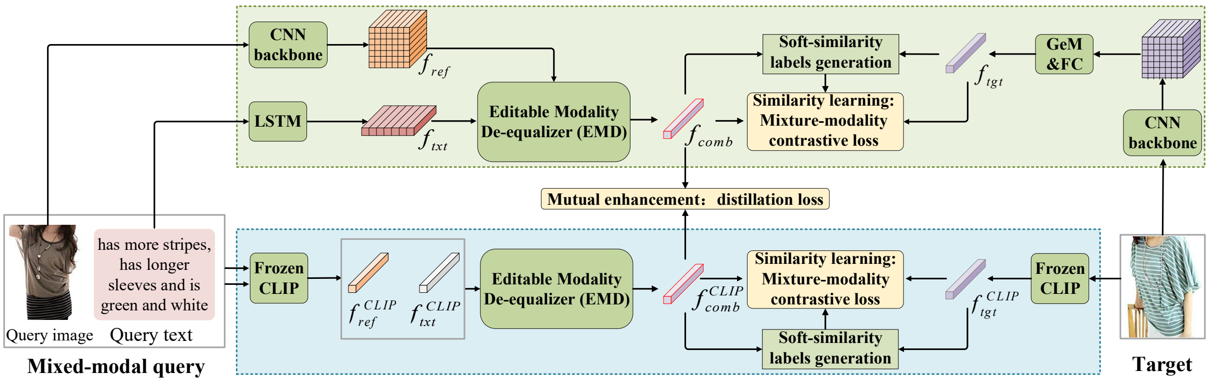


Figure 3: Overview of Dynamic Weighted Combiner (DWC), consisting of two mutually-enhanced streams.

a fine-tuning scheme for conditioned image retrieval using CLIP-based features. (Goenka et al. 2022) propose a FashionVLP that brings the prior knowledge contained in large image-text corpora to the domain of fashion image retrieval. However, most previous approaches usually achieve limited performance due to the inherent modality importance disparity and biased labeling noises in datasets. Based on our findings, we propose a DWC to solve these overlooked problems in MMIR.

Dynamic Weighted Combiner

Problem Definition and Model Architecture

In MMIR, given a mixed-modal query involving a query image I_r and a text T, the objective is to learn image-text combination features to retrieve the target image I_t. In other word, given an input pair (I_r, T), we aim to learn mixed-modal query features $f_{comb} = \mathcal{C}(I_r, T; \Theta)$ that are well-aligned with the target image feature $f_{tgt} = \mathcal{F}(I_t; \Theta)$ by maximizing their similarity as,

$$\max_{\mathbf{\Theta}} \quad \kappa(\mathcal{C}(I_r, T; \Theta), \mathcal{F}(I_t; \Theta)), \tag{1}$$

where $\mathcal{C}(\cdot)$ and $\mathcal{F}(\cdot)$ denote the mixed-modal feature composer and the image feature extractor, respectively. $\kappa(\cdot,\cdot)$ denotes the similarity kernel, implemented as dot product. Θ denotes the learnable parameters.

The proposed Dynamic Weighted Combiner (DWC) is illustrated in Fig. 3, which consists of two mutually-enhanced streams. Each stream is basically composed of four parts: (1) feature encoder to extract the visual and text feature, (2) Editable Modality De-equalizer (EMD) to edit and purify the modality features and unequally combine them with different contributions, (3) dynamic soft-similarity labels generator (SSG) to improve noisy supervision, and (4) mixed-modality contrastive loss to reduce the modality gaps and facilitate similarity learning.

Feature Encoder

As is shown in Fig. 3, the first stream adopts CNN and LSTM as the backbone to train the image encoder and text encoder, respectively. $f_{ref} \in \mathbb{R}^{C\times H\times W}$ and $f_{tgt} \in \mathbb{R}^{C\times H\times W}$ denote the feature maps of the query image and target image extracted by CNN, respectively, where $C\times H\times W$ is the shape of the feature maps. $f_{txt} \in \mathbb{R}^{D\times L}$ represents the text feature extracted by LSTM, where $D \times L$ is the shape of the text feature and L is the length of the text (i.e., the number of words in the text). To exploit the prior knowledge of large-scale web data and reduce the modality gap, the other stream adopts the large CLIP model as the feature encoder, pre-trained on a large-scale dataset with 400 million image-text pairs scraped from the web. f_{ref}^{CLIP}, f_{tgt}^{CLIP} and f_{txt}^{CLIP} denote the feature vectors of the query image, target image and query text extracted by pre-trained CLIP, respectively. For convenience, we name the two streams as CNN stream and CLIP stream.

Editable Modality De-equalizer

Since the contributions of image and text modalities are different in diverse real-world scenarios, an intuitive idea is assigning different weights. However, as discussed in the introduction, direct weighting will unavoidably amplify the error information in the dominant modality and lead to adverse performance. We therefore propose to purify the image and text features before assigning weights to reduce the interference of redundant error information, which gives rise to the EMD, formalized by two functionally symmetrical modality-feature editors and an adaptive weighted combiner, as shown in Fig. 4.

Image feature editor. To purify the image features, using image features as a template, we regard the image feature as $H \times W$ features of different spatial entities and propose a cross-modal spatial attention mechanism to assign different weights for these features. Specifically, we first get a text representation vector by the GeM pooling layer (Radenovic, Tolias, and Chum 2019). Then the spatial attention A^{sp} is formulated as

$$\begin{aligned} A^{sp} &= Reshape(Softmax(S^{sp})), \\ S^{sp}_{i,j} &= \kappa(f_{ref}(:, i, j), GeM(f_{txt})), \end{aligned} \tag{2}$$

where $A^{sp} \in \mathbb{R}^{1\times H\times W}$ and $S^{sp}_{i,j}$ indicate the similarity between the image feature of the $(i, j)^{th}$ spatial entity and the text feature. Accordingly, we can derive a coarse-modified image feature:

$$\widehat{f}_{ref} = A^{sp} \otimes f_{ref}, \tag{3}$$

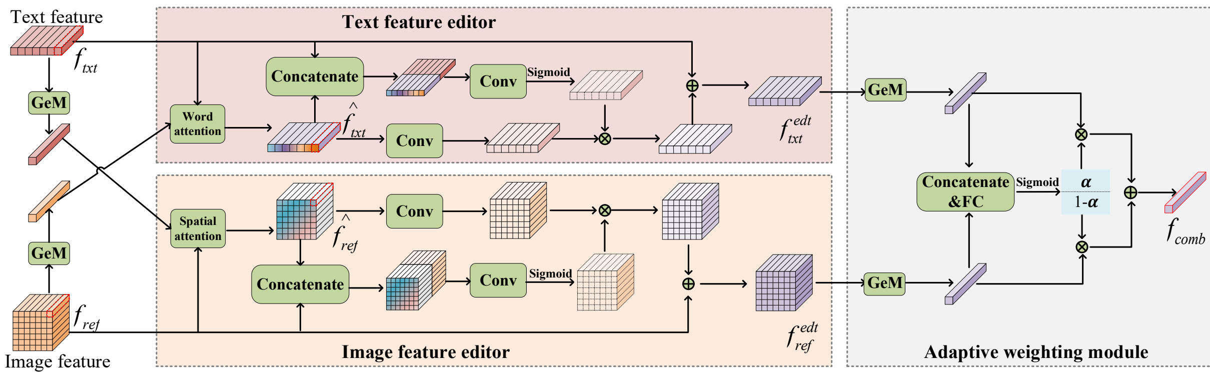


Figure 4: Overview of our Editable Modality De-equalizer (EMD) module, which contains an image feature editor, a text feature editor, and an adaptive weighted combiner. $\bigotimes$ and $\bigoplus$ stand for the Hadamard product and element-wise addition, respectively.

In order to further refine the image feature, we transform the feature map through element-level global attention, which can be formulated as

$$A^{gl}_{ref} = Sigmoid(Conv([\widehat{f}_{ref}, f_{ref}])), \tag{4}$$

where $A^{gl}_{ref} \in \mathbb{R}^{C\times H\times W}$. $Conv(\cdot)$ and $[\cdot,\cdot]$ denote 1×1 convolution layer and concatenation, respectively. Ultimately, the purified image features can be obtained as

$$f^{edt}_{ref} = (A^{gl}_{ref} \otimes Conv(\widehat{f}_{ref})) \oplus f_{ref}, \tag{5}$$

Text feature editor. Similarly, to eliminate the error information from the text, we introduce a text feature editor. Using text features as a template, we propose a cross-modal word attention mechanism to assign weights for different words. Specifically, the word attention A^w is formulated as

$$\begin{aligned} A^w &= Reshape(Softmax(S^w)), \\ S^w_l &= \kappa(f_{txt}(:l), GeM(f_{ref})), \end{aligned} \tag{6}$$

where $A^w \in \mathbb{R}^{1\times L}$ and S^w_l indicate the similarity between the text feature of the l^{th} word and the image feature. Accordingly, we can derive a coarsely-modified text feature:

$$\widehat{f}_{txt} = A^w \otimes f_{txt}, \tag{7}$$

In order to further refine the text feature, we transform the feature through element-level global attention, which can be formulated as

$$A^{gl}_{txt} = Sigmoid(Conv([\widehat{f}_{txt}, f_{txt}])), \tag{8}$$

where $A^{gl}_{txt} \in \mathbb{R}^{D\times L}$. Ultimately, the purified text features can be obtained as

$$f^{edt}_{txt} = (A^{gl}_{txt} \otimes Conv(\widehat{f}_{txt})) \oplus f_{txt}, \tag{9}$$

Adaptive weighting module. After obtaining the refined features, we introduce an adaptive weighting module to assign different modality weights for the query image and text according to their contributions. Specifically, we use α to represent the importance of the image and $1-\alpha$, the importance of the text. α is computed as

$$\alpha = Sigmoid(FC([GeM(f^{edt}_{ref}), GeM(f^{edt}_{txt})])), \tag{10}$$

where FC denotes fully-connected layers. The final combination feature is formulated as

$$f_{comb} = \alpha \cdot f^{edt}_{ref} + (1-\alpha)\cdot f^{edt}_{txt}. \tag{11}$$

Intuitively, the feature editors purify the feature maps in two steps, i.e., coarsely modifying via spatial/word attention and refining via global attention. For the CLIP stream, to reduce the computation cost, we adopt lightweight global attention to edit the CLIP features only in one step (see *Supplementary Material*).

Soft-Similarity Labels Generation

Most previous approaches overfit the biased noisy data during training, as discussed in the introduction, because the original one-hot (hard) similarity labels are usually imprecise. For example, as shown in Fig. 5 (a), the hard label for positive sample, i.e., (I_r, T) and I_t in the same triplet $< I_r, T, I_t >$, otherwise is 0 (i.e., negative sample). In fact, some negative samples can also serve as the target for some users, but re-labeling is undoubtedly a labor-intensive task.

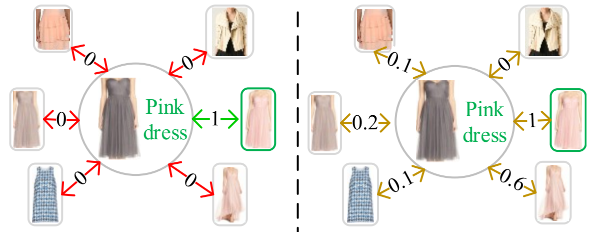
Since different language descriptions result in different similarity scores between the query and the target, we propose to relax the hard-similarity labels with soft-similarity labels, to avoid overfitting and simultaneously fully mine the valuable information in datasets, which gives rise to the soft-similarity labels generator. Specifically, we generate nonzero dynamic soft-similarity labels on-the-fly for negative pairs only. The soft-similarity label between the i^{th} query and j^{th} target in a mini-batch is generated by

$$y^s_{ij} = \begin{cases} 1, & if \;\; i=j, \\ \frac{exp(\kappa(f^i_{tgt}, f^j_{tgt})/\tau)}{\sum_{j=1}^{B} exp(\kappa(f^i_{tgt}, f^j_{tgt})/\tau)}, & if \;\; i\neq j, \end{cases} \tag{12}$$

where τ is a temperature factor. Finally, let $\mathbf{Y}^s = \{y^s_{ij}\}^{B,B}_{i=1,j=1} \in [0,1]$ denote all soft-similarity labels, and B is the batch size. A toy example of the soft-similarity labels is shown in Fig. 5 (b).

Mixed-modality Contrastive Loss

To facilitate similarity learning, we introduce a mixed-modality contrastive loss between the combined features and



(a) Hard similarity labels (b) Soft similarity labels

Figure 5: Basic idea of the proposed soft-similarity labels.

the target image features based on the soft-similarity labels. Specifically, for each mixed-modal query and target, we first compute the softmax-normalized similarity as:

$$p_{ij} = \frac{exp(\kappa(f^i_{comb}, f^j_{tgt})/\tau)}{\sum_{j=1}^{B} exp(\kappa(f^i_{comb}, f^j_{tgt})/\tau)}, \tag{13}$$

where $\mathbf{P} = \{p_{ij}\}_{i=1,j=1}^{B,B}$ represents the predicted probability. Let $\mathbf{Y}^h$ denote the noisy hard ground-truth similarity. The cross-entropy based mixed-modality contrastive loss with soft-similarity guidance is written as

$$\begin{aligned}
\mathcal{L}_{mmc} &= \lambda \mathbb{E}[H(\mathbf{Y}^h, \mathbf{P})] + (1-\lambda)\mathbb{E}[H(\mathbf{Y}^s, \mathbf{P})] \\
&= -\frac{\lambda}{B}\sum_{i=1}^{B} \log \frac{exp(\kappa(f^i_{comb}, f^i_{tgt})/\tau)}{\sum_{j=1}^{B} exp(\kappa(f^i_{comb}, f^j_{tgt})/\tau)} \\
&- \frac{(1-\lambda)}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \frac{exp(\kappa(f^i_{tgt}, f^j_{tgt})/\tau)}{\sum_{j=1}^{B} exp(\kappa(f^i_{tgt}, f^j_{tgt})/\tau)} \\
&\cdot \log \frac{exp(\kappa(f^i_{comb}, f^j_{tgt})/\tau)}{\sum_{j=1}^{B} exp(\kappa(f^i_{comb}, f^j_{tgt})/\tau)},
\end{aligned} \tag{14}$$

where $\lambda \in (0, 1)$ is a trade-off parameter and $H(\cdot)$ is the cross-entropy loss.

Model Training

Inspired by the idea of mutual learning (Zhang et al. 2018), we introduce a mutual enhancement strategy to train two streams alternately, between which the knowledge is shared via a vanilla KL-divergence based distillation loss. Specifically, we first compute the similarity between the i^{th} query and j^{th} target of both CNN and CLIP streams as follows

$$p^z_{ij} = \frac{exp(\kappa((f^i_{comb})^z, (f^j_{tgt})^z)/\tau)}{\sum_{j=1}^{B} exp(\kappa((f^i_{comb})^z, (f^j_{tgt})^z)/\tau)}, \tag{15}$$

where $z \in$ {CNN, CLIP}. In this way, we can get the similarity distribution of the i^{th} query as $p^z_i = [p^z_{i1}, p^z_{i2}, ..., p^z_{iB}]$. Then the distillation loss is

$$\mathcal{L}^z_d = \frac{1}{B}\sum_{i=1}^{B} D_{KL}(\mathbf{p}^{\bar{z}}_i || p^z_i) = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{i=1}^{B} p^{\bar{z}}_{ij} \log \frac{p^{\bar{z}}_{ij}}{p^z_{ij}}. \tag{16}$$

Taking the optimization of the CNN stream as an example, we use $\mathcal{L}^{CNN}_d$ to transfer the knowledge from the CLIP stream to the CNN stream, and there is

$$\mathcal{L}^{CNN}_d = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{i=1}^{B} p^{CLIP}_{ij} \log \frac{p^{CLIP}_{ij}}{p^{CNN}_{ij}}. \tag{17}$$

The total loss of the CNN stream is formulated as:

$$\mathcal{L}^{CNN}_{all} = \mathcal{L}^{CNN}_{mmc} + \mathcal{L}^{CNN}_d. \tag{18}$$

$$\mathcal{L}^z_{all} = \mathcal{L}^z_{mmc} + \mathcal{L}^z_d. \tag{19}$$

Notably, the total loss of the CLIP stream can be derived similarly. In the training phase, we alternately optimize the two streams to achieve mutual enhancement. Finally, the combination features of the two streams are integrated to evaluate the similarity and rank the gallery images.

Mutual information maximization perspective. Minimizing the first term of $\mathcal{L}_{mmc}$ (Eq. 14) can be seen as maximizing the lower bound on the mutual information (MI) between the mixed-modal query and the target, i.e., maximizing a symmetric version of InfoNCE (van den Oord, Li, and Vinyals 2018). Therefore, the proposed $\mathcal{L}_{mmc}$ can not only learn the mixed-modal similarities for MMIR, but also reduce the modality gaps.

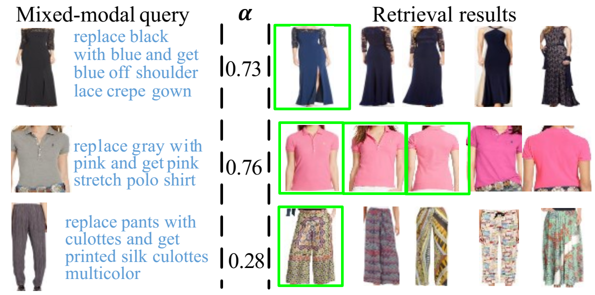


Figure 6: Qualitative results on Fashion200k.

Experiments

To evaluate our model, we chose three real-world datasets: Fashion200K (Han et al. 2017), Shoes (Guo et al. 2018), and FashionIQ (Wu et al. 2021). We compare our DWC with many SOTA MMIR methods, such as **TIRG** (Vo et al. 2019), **JAMMAL** (Zhang et al. 2020), **LBF** (Hosseinzadeh and Wang 2020), **JVSM** (Chen and Bazzani 2020), **SynthTripletGAN** (Tautkute and Trzcinski 2021), **VAL** (Chen, Gong, and Bazzani 2020), **DCNet**(Kim et al. 2021), **JPM** (Yang et al. 2021b), **DATIR** (Gu et al. 2021), **ComposeAE** (Anwaar, Labintcev, and Kleinsteuber 2021), **CoSMo** (Lee, Kim, and Han 2021), **CLVC-Net** (Wen et al. 2021), **ARTEMIS** (Delmas et al. 2022), **SAC** (Jandial et al. 2022), **GA** (Huang et al. 2022), ***CIRPLANT*** (Liu et al. 2021), ***Combiner w/ CLIP*** (Baldrati et al. 2022b), and ***FashionVLP*** (Goenka et al. 2022), where the methods in italic are based on VLP models.

Experimental Results

Quantitative Results. Table 2 shows comparisons with existing methods on the FashionIQ dataset. We observe that

Method	Dress		Shirt		Toptee		Average		
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Mean
TIRG	14.87	34.66	18.26	37.89	19.08	39.68	17.40	37.41	27.41
VAL	22.53	44.00	22.38	44.15	27.53	51.68	24.15	46.61	35.38
ComposeAE	14.03	35.10	13.88	34.59	15.80	39.26	14.57	36.32	25.44
JVSM	10.70	25.90	12.00	27.10	13.00	26.90	11.90	26.63	19.27
SynthTripletGAN	22.60	45.10	20.50	44.08	28.01	52.10	23.70	47.09	35.40
CoSMo	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
JPM	21.38	45.15	22.81	45.18	27.78	51.70	23.99	47.34	35.67
DATIR	21.90	43.80	21.90	43.70	27.20	51.60	23.67	46.37	35.02
CLVC-Net	29.85	56.47	28.75	54.76	33.50	64.00	30.70	28.41	44.56
ARTEMIS	27.34	44.18	21.05	49.87	24.91	48.59	24.43	47.55	35.99
SAC	26.13	52.10	26.20	50.93	31.16	59.05	27.83	54.03	40.93
CIRPLANT	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
Combiner w/ CLIP	26.82	51.31	31.80	53.38	33.40	57.01	30.67	53.90	42.29
FashionVLP	32.42	**60.29**	31.89	58.44	38.51	**68.79**	34.27	**62.51**	48.39
DWC	**32.67**	57.96	**35.53**	**60.11**	**40.13**	66.09	**36.11**	61.39	**48.75**

Table 2: Interactive image retrieval performance (%) on the FashionIQ dataset. The best results are in bold.

Method	R@1	R@10	R@50	Average
TIRG	14.10	42.50	63.80	40.13
JVSM	19.00	52.10	70.00	47.03
JAMMAL	17.30	45.30	65.70	42.77
LBF	17.80	48.40	68.50	44.90
VAL	22.90	50.80	72.70	48.80
DCNet	–	46.90	67.60	–
JPM	19.80	46.50	66.60	44.30
DATIR	21.50	48.80	71.60	47.30
ComposeAE	22.80	55.30	73.40	50.50
CoSMo	23.30	50.40	69.30	47.67
CLVC-Net	22.60	53.00	72.20	49.27
ARTEMIS	21.50	51.10	70.50	47.70
GA	24.00	57.20	75.70	52.30
Combiner w/ CLIP	20.56	52.07	71.35	47.99
FashionVLP	–	49.90	70.50	–
DWC	**36.49**	**63.58**	**79.02**	**59.70**

Table 3: Performance Comparison (%) on Fashion200k.

Method	R@1	R@10	R@50	Average
TIRG	12.60	45.45	69.39	42.48
VAL	17.18	51.52	75.83	48.18
SynthTripletGAN	–	47.6	73.6	–
ComposeAE	4.37	19.36	47.58	23.77
CoSMo	16.72	48.36	75.64	46.91
DATIR	17.20	51.10	75.60	47.97
DCNet	–	53.8	79.3	–
CLVC-Net	17.60	54.40	79.50	50.50
ARTEMIS	17.6	51.05	76.85	48.50
SAC	18.11	52.41	75.42	48.65
Combiner w/ CLIP	8.12	33.28	62.42	34.61
FashionVLP	–	49.08	77.32	–
DWC	**18.94**	**55.55**	**80.19**	**51.56**

Table 4: Performance Comparison (%) on Shoes.

the performance improvement of DWC over the second-best method is 0.25%, 3.64% and 1.98% on R@10 for three subsets, i.e., dress, shirt, and top-tee, respectively. Our DWC is slightly lower than FashionVLP in R@50, a more complex model equipped with auxiliary modules such as object detection module and landmark module. Table 3 shows our comparison with existing methods on the Fashion200k dataset. As can be seen, our model demonstrates compelling results compared to all other alternatives. The performance improvement of DWC over the second-best method is 12.49% and 7.4% on R@1 and the average, respectively. Table 4 shows comparisons with existing methods on the Shoes dataset. Our DWC model still achieves the best performance with improvements of 0.83% and 1.06% on R@1 and the average, respectively.

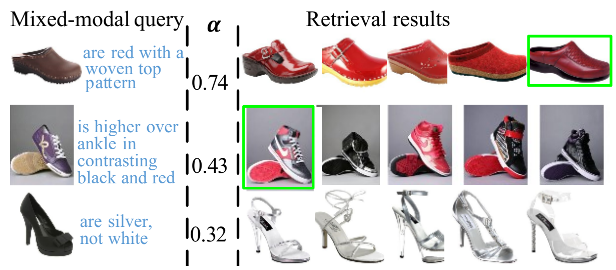


Figure 7: Qualitative results on the Shoes dataset.

Figure 8: Qualitative results on the FashionIQ dataset.

Qualitative Results and Analysis. Fig. 6, Fig. 7 and Fig. 8 shows our qualitative results on the Fashion200k, Shoes and FashionIQ datasets.Fig. 9 provides the attention visualizations, which show the proposed spatial attention and word attention are effective.

Ablation Study and Model Analysis

Ablation study of losses. The results are provided in Table 5. DWC w/o $\mathcal{L}^h_{mmc}$ and DWC w/o $\mathcal{L}^s_{mmc}$ represent remove

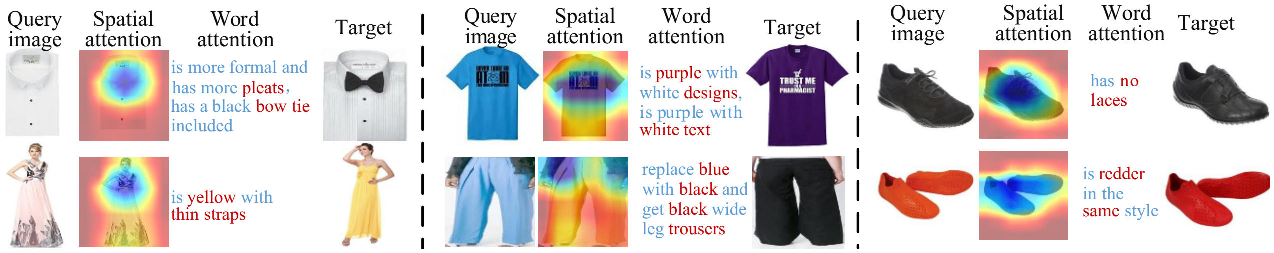


Figure 9: Attention visualization of our model on different datasets. Words with high attention value are in red.

Method	Dress		Shirt		Toptee	
	R@10	R@50	R@10	R@50	R@10	R@50
w/o $\mathcal{L}_{mmc}^{h}$	31.82	57.11	34.89	61.19	39.37	65.37
w/o $\mathcal{L}_{mmc}^{s}$	31.73	57.80	34.99	59.91	39.42	65.63
w/o $\mathcal{L}_{d}$	29.74	55.82	32.04	59.13	36.21	65.12
DWC	**32.67**	**57.96**	**35.53**	**60.11**	**40.13**	**66.09**

Table 5: Ablation study of losses.

Method	Dress	Shirt	Toptee	Shoes
DWC w/o EMD	23.55	26.40	29.37	50.72
DWC w/o SSG	31.73	34.99	39.42	55.44
DWC w/o CLIP	24.79	24.48	29.37	52.32
DWC w/o ME	28.01	27.82	31.00	53.41
EMD w/o TFE	30.49	32.34	36.92	52.54
EMD w/o IFE	28.01	30.72	34.12	51.86
EMD w/o FE	27.37	29.39	33.61	49.65
DWC	**32.67**	**35.53**	**40.13**	**55.55**

Table 6: Ablation study (R@10).

$\mathbb{E}[H(\mathbf{Y}^h, \mathbf{P})]$ and $\mathbb{E}[H(\mathbf{Y}^s, \mathbf{P})]$ from Eq. (14), respectively. DWC w/o $\mathcal{L}_d$ means remove the distillation loss, i.e., Eq. (16). Results reveal the effectiveness of the losses and mutual enhancement strategy.

Analysis of different components. To demonstrate the contributions from different components in our proposed model, we conduct ablation studies in Table 6, in which *DWC w/o EMD*, *DWC w/o CLIP*, *DWC w/o SSG* and *DWC w/o ME* indicate the variants of our DWC by removing EMD, CLIP stream, SSG and mutual enhancement, respectively. Contribution of different components in EMD is also evaluated, *EMD w/o TFE*, *EMD w/o IFE* and *EMD w/o FE* are the variants of DWC by removing text feature editor, image feature editor and two feature editors from EMD, respectively. Experiments show that each component plays a significant role in improving the MMIR performance. This further verifies our intuition to meet the challenges of inherent modality importance disparity, biased labeling noises in datasets and modality gaps.

Analysis of modality importance disparity. The modality importance disparity, including its impact and the effectiveness of our EMD, is presented in Table 7. The first two rows indicate that different modalities play different roles. The third and fourth rows indicate that mixed-modality feature fusion (sum. vs. concat.) can effectively improve the performance. The fifth row shows that naive weighting of two modalities cannot improve modality importance disparity because error of the important modality is also amplified. In contrast, the proposed EMD can significantly improve the performances to a large margin by taking into account the purification of mixed-modality features with an adaptive weighting combiner.

Method	Dress	Shirt	Toptee	Shoes
Image Only	7.83	13.74	10.20	31.92
Text Only	18.79	22.52	25.50	15.39
Summation	23.55	26.40	29.37	50.72
Concatenation	28.31	30.52	35.03	50.50
Weighting	28.36	30.30	34.98	50.16
Our EMD	**32.67**	**35.53**	**40.13**	**55.55**

Table 7: Impact of different combination methods (R@10).

CLIP model	Dress	Shirt	Toptee	Shoes
RN50	32.67	35.53	40.13	55.55
RN101	32.42	35.97	39.93	54.25
ViT-B32	31.63	34.99	38.55	**56.26**
ViT-B16	**33.61**	**37.09**	**40.80**	54.92

Table 8: Impact of pre-trained CLIP models (R@10).

Impact of pre-trained CLIP models under different backbones. We consider four versions of the pre-trained CLIP models in the CLIP stream to conduct experiments. As presented in Table 8, we observe slight differences among them, which, instead, indicate that the modality gap can be largely bridged by a vanilla CLIP model.

Notably, **the impact of different soft labels** and many other experimental analyses are provided in the *supplementary material* at https://arxiv.org/pdf/2312.06179.pdf.

Conclusion

We have two critical findings that are seriously overlooked in MMIR community. 1) There exists significant modality importance disparity, leading to degradation of model training. 2) There exists inherent labeling noises and data bias due to diverse text descriptions crawled from web scenarios. Based on the findings, we propose a DWC, which includes three merits. First, we propose an EMD by taking into account the contribution disparity between mixed modalities. Second, we propose a dynamic SSG to relax the biased hard labeling and implicitly improve noisy supervision. Finally, to bridge modality gaps and facilitate similarity learning, we propose a CLIP-based mutual enhancement module alternately trained by a mixed-modality contrastive loss. Experiments and analysis verify the superiority of our approach.

Acknowledgements

This work was partially supported by National Key R&D Program of China (2021YFB3100800), National Natural Science Fund of China (62271090, 61771079), Chongqing Natural Science Fund (cstc2021jcyj-jqX0023) and National Youth Talent Project. This work is also supported by Huawei computational power of Chongqing Artificial Intelligence Innovation Center.

References

Anwaar, M. U.; Labintcev, E.; and Kleinsteuber, M. 2021. Compositional Learning of Image-Text Query for Image Retrieval. In *WACV*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2021. Conditioned image retrieval for fashion using contrastive learning and CLIP-based features. In *ACMMM Asia*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2022a. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *CVPRW*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2022b. Effective conditioned and composed image retrieval combining CLIP-based features. In *CVPR*.

Brown, A.; Xie, W.; Kalogeiton, V.; and Zisserman, A. 2020. Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval. In *ECCV*.

Chen, Y.; and Bazzani, L. 2020. Learning Joint Visual Semantic Matching Embeddings for Language-guided Retrieval. In *ECCV*.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *CVPR*, 2998–3008.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *CVPR*, 8415–8424.

Delmas, G.; Rezende, R. D.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *ICLR*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; Liu, Z.; and Zeng, M. 2022. An Empirical Study of Training End-to-End Vision-and-Language Transformers. In *CVPR*.

Dubey, S. R. 2022. A Decade Survey of Content Based Image Retrieval Using Deep Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2687–2704.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback. In *CVPR*.

Gu, C.; Bu, J.; Zhang, Z.; Yu, Z.; Ma, D.; and Wang, W. 2021. Image Search with Text Feedback by Deep Hierarchical Attention Mutual Information Maximization. In *ACM MM*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris., R. S. 2018. Dialog-based Interactive Image Retrieval. In *NeurIPS*.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; uan Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *ICCV*.

Hosseinzadeh, M.; and Wang, Y. 2020. Composed Query Image Retrieval Using Locally Bounded Features. In *CVPR*, 3593–3602.

Hou, Y.; Vig, E.; Donoser, M.; and Bazzani, L. 2021. Learning Attribute-Driven Disentangled Representations for Interactive Fashion Retrieval. In *ICCV*, 12147–12157.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning Cross-Modal Retrieval With Noisy Labels. In *CVPR*, 5403–5413.

Huang, F.; and Zhang, L. 2023. Language Guided Local Infiltration for Interactive Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Huang, F.; Zhang, L.; and Gao, X. 2021. Domain Adaptation Preconceived Hashing for Unconstrained Visual Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*.

Huang, F.; Zhang, L.; Yang, Y.; and Zhou, X. 2020. Probability Weighted Compact Feature for Domain Adaptive Retrieval. In *CVPR*, 9579–9588.

Huang, F.; Zhang, L.; Zhou, Y.; and Gao, X. 2022. Adversarial and Isotropic Gradient Augmentation for Image Retrieval with Text Feedback. *IEEE Transactions on Multimedia*.

Jandial, S.; Badjatiya, P.; Chawla, P.; Chopra, A.; Sarkar, M.; and Krishnamurthy, B. 2022. SAC: Semantic Attention Composition for Text-Conditioned Image Retrieval. In *WACV*.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual Compositional Learning in Interactive Image Retrieval. In *AAAI*.

Lee, S.; Kim, D.; and Han, B. 2021. CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback. In *CVPR*.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S.; Xiong, C.; and Hoi, S. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*.

Nam, S.; Kim, Y.; and Kim, S. J. 2018. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language. In *NeurIPS*.

Radenovic, F.; Tolias, G.; and Chum, O. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7): 1655–1668.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; and Clark, J. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Shen, Y.; Qin, J.; Chen, J.; Yu, M.; Liu, L.; Zhu, F.; Shen, F.; and Shao, L. 2020. Auto-Encoding Twin-Bottleneck Hashing. In *CVPR*, 2815–2824.

Shin, M.; Cho, Y.; Ko, B.; and Gu, G. 2021. RTIC: Residual Learning for Text and Image Composition using Graph Convolutional Network. ArXiv.2104.03015.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *ICCV*, 3027–3035.

Tautkute, I.; and Trzcinski, T. 2021. I Want This Product but Different : Multimodal Retrieval with Synthetic Query Expansion. In *CVPR*.

Tian, Y.; Newsam, S.; and Boakye, K. 2022. Image Search with Text Feedback by Additive Attention Compositional Learning. ArXiv.2203.03809.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. ArXiv.1807.03748.

Vo, N.; Lu, J.; Chen, S.; Murphy, K.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - An Empirical Odyssey. In *CVPR*.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *CVPR*, 10938–10947.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *ACM SIGIR*.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*.

Yang, Y.; Wang, L.; Xie, D.; Deng, C.; and Tao, D. 2021a. Multi-Sentence Auxiliary Adversarial Networks for Fine-Grained Text-to-Image Synthesis. *IEEE Transactions on Image Processing*, 30: 2798–2809.

Yang, Y.; Wang, M.; Zhou, W.; and Li, H. 2021b. Cross-modal Joint Prediction and Alignment for Composed Query Image Retrieval. In *ACM MM*.

Zhang, F.; Xu, M.; Mao, Q.; and Xu, C. 2020. Joint Attribute Manipulation and Modality Alignment Learning for Composing Text and Image to Image Retrieval. In *ACM MM*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *CVPR*.

Zhao, Y.; Song, Y.; and Jin, Q. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. In *ACM SIGIR*.