AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval

HONGGUANG ZHU, YUNCHAO WEI, and YAO ZHAO, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, and Peng Cheng Laboratory
CHUNJIE ZHANG and SHUJUAN HUANG, Beijing Jiaotong University and Beijing Key Laboratory of Advanced Information Science and Network Technology

Text-guided image retrieval integrates reference image and text feedback as a multimodal query to search the image corresponding to user intention. Recent approaches employ multi-level matching, multiple accesses, or multiple subnetworks for better performance regardless of the heavy burden of storage and computation in the deployment. Additionally, these models not only rely on expert knowledge to handcraft image-text composing modules but also do inference by the static computational graph. It limits the representation capability and generalization ability of networks in the face of challenges from complex and varied combinations of reference image and text feedback. To break the shackles of the static network concept, we introduce the dynamic router mechanism to achieve data-dependent expert activation and flexible collaboration of multiple experts to explore more implicit multimodal fusion patterns. Specifically, we construct AMC, our *A*daptive *M*ulti-expert *C*ollaborative network, by using the proposed router to activate the different experts with different levels of image-text interaction. Since routers can dynamically adjust the activation of experts for the current samples, AMC can achieve the adaptive fusion mode for the different reference image and text combinations and generate dynamic computational graphs according to varied multimodal queries. Extensive experiments on two benchmark datasets demonstrate that due to benefits from the image-text composing representation produced by an adaptive multi-expert collaboration mechanism, AMC has better retrieval performance and zero-shot generalization ability than the state-of-the-art method while keeping the lightweight model and fast retrieval speed. Moreover, we analyze the visualization of path activation, attention map, and retrieval results to further understand the routing decisions and semantic localization ability of AMC. The codes and pretrained models are available at https://github.com/KevinLight831/AMC.

CCS Concepts: • **Information systems** → **Image search**; **Collaborative search;**

Additional Key Words and Phrases: Text-guided image retrieval, multimodal fusion, mixture-of-experts

This work was supported in part by the National Key Research and Development of China (grant 2018AAA0102100), the National Natural Science Foundation of China (62120106009 and 62072026), and the Beijing Natural Science Foundation (JQ20022).

Authors' addresses: H. Zhu, Y. Wei, and Y. Zhao (corresponding author), Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China, and Peng Cheng Laborotory, Shenzhen 518066, China; emails: hongguang@bjtu.edu.cn, wychao1987@gmail.com, yzhao@bjtu.edu.cn; C. Zhang and S. Huang, Institute of Information Science, Beijing Jiaotong University and Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China; emails: {cjzhang, shujuanhuang}@ bjtu.edu.cn.

1551-6857/2023/05-ART188 $15.00
https://doi.org/10.1145/3584703

ACM Reference format:
Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6, Article 188 (May 2023), 22 pages.
https://doi.org/10.1145/3584703

1 INTRODUCTION

Image retrieval [11, 44] aims to retrieve images that meet the intention of users from the gallery. There are two challenges for image retrieval: the semantic gap and the intention gap. The former has gained great breakthroughs with the development of cross-modal retrieval [6, 11, 30, 41, 61] and large-scale vision-language pretraining [49, 63], but the intention gap has received less attention. Traditional image retrieval only allows users to express their search intent by single modality, such as image or text, but unimodal queries could not describe the user intention accurately. Image is information-rich but ambiguous (i.e., the redundant information makes it difficult for the model to attend to the characteristics that users care about), whereas text is accurate but partial (i.e., the abstract semantics makes it difficult to exhaustively describe the target image). TIRG [57] proposes text-guided image retrieval that utilizes the reference image and text modifier from user feedback as a multimodal query to retrieve the target image. The reference image can be flexibly selected and does not need to strictly match the user intention, whereas the text modifier expresses the concerning characteristics of retrieval intention and mitigates the gap between the reference and target images. This task is mainly applied for interactive and progressive product search [17]. Because of the high similarity between products in the e-commerce scenario, users can perform the personalized and fine-grained product search by entering text that describes the difference between the current recommended products and the target products.

Several works have tried to tackle the problem of text-guided image retrieval from different perspectives. Their efforts can be roughly divided into two methods: (1) learning a better image-text compositor to refine multimodal query, and (2) learning a reliable similarity measure for accurate query-target matching. The first method (1) [31, 57, 69] (Figure 1(a)) integrates input to generate multimodal query representation similar to the target image and transforms the retrieval problem into the spatial nearest neighbor search problem. Benefiting from the uncoupled inference structure, it can achieve good scalability and efficiency through precompute and cache multimodal features when deploying on large-scale databases. The second method [5, 28, 58] (see Figure 1(b)–(d)) tends to have better performance than the first but also brings the additional computational and time burden, of which there are three main approaches: multi-level matching, multiple accesses, and multiple subnetworks. For more details on these approaches, refer to Section 2.1. Given that the preceding first method has the advantages of an intuitive computational framework, separable query-target processing, lower retrieval time cost, and deploy-friendly scalability, it is of practical application value to further improve the performance of that method. Additionally, the same reference image is often associated with different feedback to retrieve the different targets in real life. For example, in the Fashion IQ and Shoes benchmark datasets [18], this type of data statistically accounts for 49.51% and 99.9% of the total data, and the same image is associated with up to 16 and 20 texts in extreme cases. The static models adopt the one-size-fits-all approach to simultaneously optimize multiple composing queries with the same image but different text. It is often suboptimal due to the interference of multi-user intentions (as shown in Figure 2(a)). Inspired by the success of **Mixture-of-Experts (MoE)** methods [37] in solving task-interference issues of multi-task learning, we propose the data-adaptive expert activation and multi-expert collaboration to divide and conquer for multimodal composing queries (as shown in Figure 2(b)). However, different from the traditional MoE methods that use unified multi-layer perceptrons as the experts and sparsely

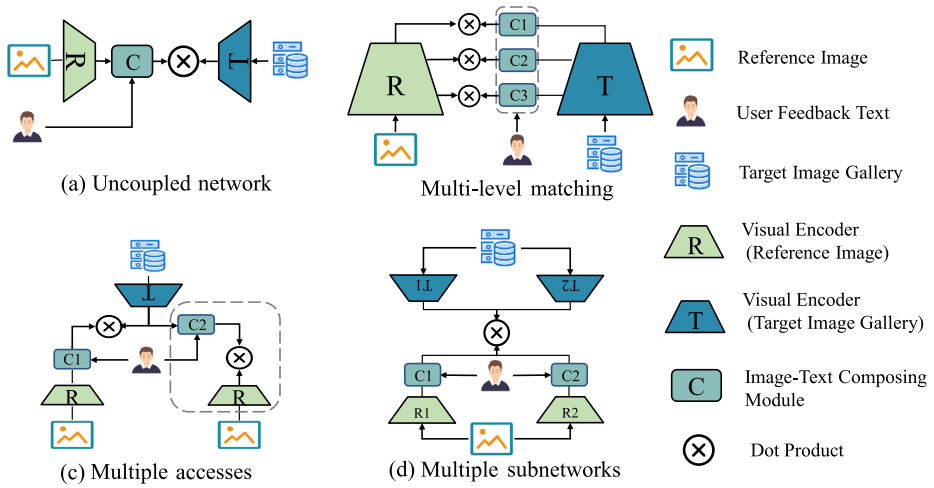


Fig. 1. A sketch of network architectures from four categories on current text-guided image retrieval methods.

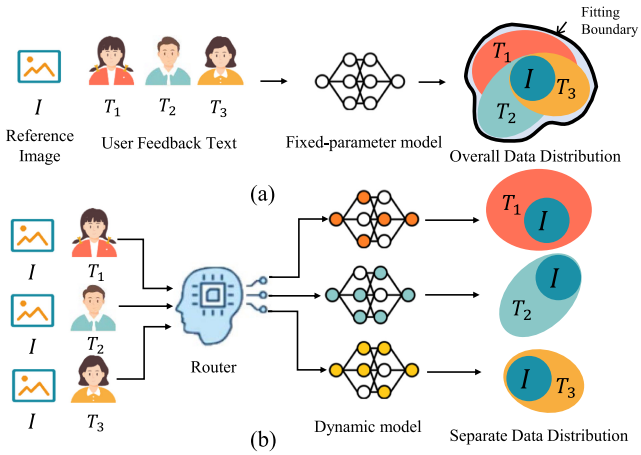


Fig. 2. The combination of reference images and different user feedback texts is a complex distribution. It is suboptimal for a traditional static network (a) with the fixed computational graph to fit overall distribution. Our method (b) can dynamically adjust network nodes to adapt to the current sample distribution and is more flexible and expressive than static networks. For simplicity, the distribution diagram here assumes that the user feedback text retains all content of the original image but just complements with different tendencies.

activate experts to resolve the task-interference issues, we design multiple network nodes with different structures as experts and achieve the unified control of expert activation by the extra router. Compared with the traditional fixed model, the flexibility of our network makes it possible to solve the interference of multi-user intentions and can explore more implicit multimodal fusion patterns. The experiment results demonstrate the effectiveness of our adaptive multi-expert collaboration strategy compared with the overall optimization of the static model and sparse activation of multi-expert paths. More analysis is shown in Section 5.4.

In this article, we aim to revive the traditional uncoupled network architecture (see Figure 1(a)) by adaptive multi-expert collaboration and further improve the performance while keeping the advantages of low-cost deployment and efficient inference. Specifically, we adopt three basic fusion nodes as experts: the **Normalized Identity Node (NIN)** communicates shallow and deep feature representations to prevent network degradation, the **Global Transformation Node (GTN)** leverages text to guide the overall transformation of visual semantic dimensions, and the **Cross-modal Reasoning Node (CRN)** collaboratively reasons image and text feature space to replenish details. Additionally, the proposed composition router can generate the data-dependent path activation value for each node according to the current multimodal query representation and achieve customized fitting for different user tendencies. To regularize the network and supervise router training, we propose **Batch-based Similarity Consistency (BSC)** loss to constrain the representation consistency between queries and targets. Benefiting from the ability to adaptively adjust expert activation and schedule multi-expert collaboration, our model can flexibly adjust the multimodal fusion mechanism according to different queries and achieve significant performance improvement over state-of-the-art methods. Our model also takes into account the characteristics of scalable uncoupled inference architecture, lightweight model size, fast inference speed, better generalization ability, and so on. These characteristics demonstrate the potential of our method for real-world applications. To further facilitate the intuitive understanding, we visualize and analyze the path activation values of routers and give more qualitative results.

Our contributions are as follows:

- Different from the traditional static model of the one-size-fits-all approach, we first propose to combine data-adaptive expert activation and multi-expert collaboration to achieve divide and conquer for different multimodal queries.
- Technically, we propose AMC, an *A*daptive *M*ulti-expert *C*ollaborative network, to achieve the data-adaptive image-text composition for text-guided image retrieval. It contains three fusion nodes with different advantages as independent experts, and a composition router automatically schedules different experts according to the current multimodal queries.
- Extensive experiments on benchmark datasets demonstrate that our method not only achieves noticeable performance improvement and better generalization ability compared with the state-of-the-art method but also cuts down 0.5× on model parameters and accelerates 3× on retrieval speed. It achieves a win-win for performance and application cost and shows promising tendencies in practical application.
- To further understand the dynamic inference process of this new divide-and-conquer paradigm, we provide a detailed statistical and visualization analysis of the activation paths and retrieval results.

2 RELATED WORKS

2.1 Text-guided Image Retrieval

Traditional image retrieval is mostly based on image-image [15, 44, 71] or image-text retrieval [11, 33, 34, 41, 43, 65] whereby users search the target image by inputting image or text. However, it is difficult to perfectly reflect user intention through a unimodal query, especially in e-commerce scenarios with a large similarity among products. To better understand user intention, interactive image retrieval incorporates the feedback of users to navigate satisfactory retrieval results. Such feedback includes relevance [51], attributes [1, 19, 39, 72], sketches [14, 64, 66], spatial layouts [36, 40], or modifier texts [31, 57], of which the text is the most pervasive medium for human-computer interaction and naturally communicates the distinctive requirements of users. In this work, we investigate the composing image retrieval with various text feedbacks, which is also called *text-guided image retrieval*.

The early methods of text-guided image retrieval mainly focus on modifying the attribute of the reference image by the concrete predefined attribute descriptions. For example, AMNet [72] proposes a memory-augmented attribute manipulation network that fuses the store attribution template representations and the reference image representations to retrieve target images. Although previous studies have achieved promising results, they confine the intentions of users to a set of predefined attributes that limit real applications. To this end, TIRG [57] conducts the gating residual mechanism to compose natural language and the reference image for target image retrieval. TIS [69] further manipulates the visual features of the reference image in terms of the text description by the generative adversarial network. MAAF [10] uses the transformer architecture to achieve modality-agnostic attention fusion and studies many combinations of attention fusion modules and textual backbones. CoSMo [31] proposes the content and style modulators, which directly modify the high-level feature of the reference image based on the text modifier. As shown in Figure 1(a), these methods [10, 31, 57, 68, 69] adopt an uncoupled network architecture and can achieve good scalability and efficiency in the application.

Other works pursue accurate query-target matching by complex similarity measure methods, and such methods often have better performance than the former at the cost of computational and storage burdens. VAL [5], as shown in Figure 1(b), fuses the text with the low-level, mid-level, and high-level features of the reference image, respectively, then hierarchically matches with the target image. However, it also causes database storage consumption theoretically three times higher than the uncoupled methods. DCNet [28], as shown in Figure 1(c), proposes the composition and correction network to achieve the cyclic mapping of the query and target image. However, the server needs to iteratively access and process all the combinations of current reference images and images from the target image gallery. Because the target image gallery is often quite large, this amount of calculation is unaffordable in real applications. CLVC-Net [58], as shown in Figure 1(d), owns two subnetworks with independent image and text encoders, which fuses the feature from two subnetworks to represent the final query and target. It is similar to the ensemble approach of heterogeneous networks which can effectively improve performance, but it also results in theoretically double computation and storage consumption compared to the uncoupled methods. Although these methods have made significant progress, they rely on expert experience to handcraft networks and conduct the fixed computation graph to process diverse combinations of text modifiers and reference images. This could limit the modal fusion capability and representation flexibility of the text-guided image retrieval network.

2.2 Mixture-of-Experts

Different from the traditional fixed-parameter network, the MoE models [20] build multiple parallel network branches as experts and selectively activate experts. The outputs of these experts are fused with data-dependent weights by routers. According to the router design, MoE models can be roughly divided into two categories. The first category is soft MoEs [37, 38], whose purpose is to adopt different weights to dynamically rescale the representation from different experts and achieve the judgment from MoE to boost the representation power. For instance, GPSNet [13] introduces the soft-conditional gated path selection network to adaptively learn data-dependent receptive fields and dynamically select semantic context for semantic segmentation. DIME [48] assigns a private router to each expert and constructs the fully connected routing space for each multiple-branch layer. It further expands the representation space while consuming more computation. The second category is hard MoEs [4, 59], whose purpose is to adjust the network structure to allocate appropriate computation for the corresponding samples and improve the inference speed. It is mainly implemented by sparse selective execution and usually needs specific training strategies to guarantee performance. For example, HydraNet [42] replaces the convolutional blocks with

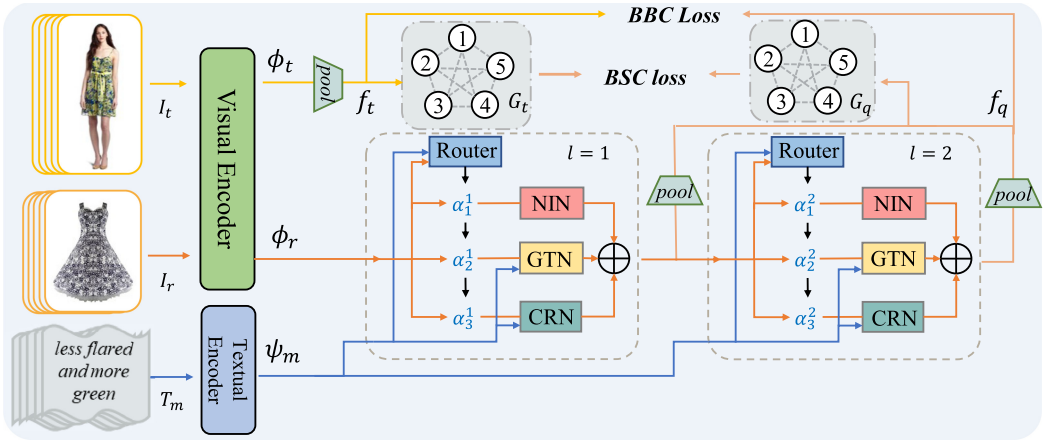


Fig. 3. Illustration of the proposed AMC architecture. The blue and yellow lines represent the processing of textual features and target image features, and the orange line represents the processing of intermediate multimodal representation.

multiple branches and selectively executes them according to the input to achieve the efficient inference of image classification. Uni-Perceive-MoE [73] only activates one expert for each MoE layer to mitigate the task-interference issue in generalist models.

Different from traditional MoE models that use a uniform structure for all experts, we construct multiple multimodal fusion nodes with different structures and use a soft MoE approach to fuse output of nodes according to the path activation value of our composition router. To the best of our knowledge, we are the first to introduce the methodology of MoE to learn the adaptive fusion of diverse multimodal queries on text-guided image retrieval.

3 PROBLEM FORMULATION

Text-guided image retrieval is defined as follows. Given a multimodal composed query of a reference image and its modify text, the system needs to retrieve relevant images from the target image database. Additionally, the reference and target images share the same visual encoder and coding rules. Thus, the essential question is how to use the semantical information from text modifiers to modify the representation of reference images and achieve alignment with target images in the same visual domain. Given a set with N triplets, denoted as $\mathcal{D} = \{(I_r^i, T_m^i, I_t^i)\}_{i=1}^N$, where the I_r and I_t are the reference and target images, and T_m is the text modifiers. Text-guided image retrieval aims to learn a multimodal composition mechanism to represent the multimodal query (I_r^i, T_m^i) and achieve the query that should be as close as possible to the corresponding target image (positive sample) I_t^i and as far away from unmatched target images (negative samples) $\{I_t^{j\neq i} | j = 1, \ldots, N\}$ as possible. Formally, we can formulate the following:

$$\mathcal{H}(\mathcal{F}(I_r^i), \mathcal{G}(T_m^i)) \rightarrow \mathcal{F}(I_t^{j=i} | j = 1, \ldots, N), \tag{1}$$

where $\mathcal{F}$ and $\mathcal{G}$ represent the visual and textual transformation spaces, respectively, whereas $\mathcal{H}$ denotes the multimodal fusion to achieve alignment with the target samples in the semantical representation.

4 METHODOLOGY

As Figure 3 illustrates, first, we respectively extract the high-dimensional feature representations for reference image I_r and user feedback text T_m using the visual encoder and text encoder. Then,

we design our composition layer, which contains a composition router unit and three fusion nodes. Each multimodal query (image + text) will go through the router unit and autonomously choose its own composition path to achieve diverse modifications. The average pooling output will be used as the final query representation f_q for the target image retrieval. The target image I_t is also processed by the same visual encoder and pooled as the target feature f_t.

4.1 Visual and Textual Representation

For fair comparison with most prior works [5, 28, 31, 58], we choose a shared CNN as the visual backbone to capture the local features of reference and target images. We denote the features as follows:

$$\phi_r = CNN(I_r), \quad \phi_t = CNN(I_t), \tag{2}$$

where $\phi_r, \phi_t \in \mathbb{R}^{H\times W\times D}$ refer to the intermediate representation of reference and target images. Additionally, since the retrieval task requires a compressed vector representation to facilitate storage, the local features of target images ϕ_t will be averagely pooled to represent the target ground truth representation f_t. Note that retrieval finally needs the compressed image representation for storage, and we do not pursue the spatial alignment of the feature map to accomplish the image generation task. Thus, for simplicity, we denote the reference representation as $\phi_r \in \mathbb{R}^{K\times D}, K = H \times W$. To represent the semantics of text modifier T_m, we first split words with the tokenization and obtain the corresponding word vectors on the word embedding. Following other works [5, 16, 54, 62], we use the text encoder followed by max-pooling to obtain the final text representation $\psi_m \in \mathbb{R}^D$.

4.2 Adaptive Multi-expert Collaborative Network

Different from the previous networks that process all samples with fixed paths and parameters, we propose our AMC with self-adjusting path activation and accomplish data-dependent semantic fusion for reference image and text. It mainly contains three different experts and a router unit. To emphasize the structural nature of the different experts, we refer to the different experts as nodes in the following. The router unit controls the activation degree of each path and commands each basic node to divide the labor and cooperate according to the current multimodal query representation and modified text representation ψ_m. These basic nodes can construct the modality interaction from different views and extend the representation capability of the model. Operations on these nodes can be formulated as follows:

$$P_i^l = \begin{cases} \mathcal{F}_i^l(X^{l-1}), & i = 1 \\ \mathcal{F}_i^l(X^{l-1}, \psi_m), & i = 2 \text{ or } 3 \end{cases} \tag{3}$$

where $\mathcal{F}_i^l(\cdot)$ represents the processing function of the i-th node in the l-th composition layer. $X^{l-1} \in \mathbb{R}^{K\times D}$ denotes the multimodal query feature of the $(l-1)$-th composition layer, and $P_i^l \in \mathbb{R}^{K\times D}$ denotes the output intermediate representation of the i-th node in the l-th composition layer. Because the same type of nodes in different composition layers own the same architecture but different learned parameters, we omit the subscript index of the composition layer for simplicity in the following.

4.2.1 Normalized Identity Node. On the one hand, some reference images may not need excessive interaction with the text. On the other hand, stacking too deep cross-modal fusion units may lead to network degradation and gradient vanishing. Inspired by He et al. [21], we construct the NIN to retain and normalize output of the previous layer and facilitate the optimization of gradient

flow. It can be formulated as

$$\mathcal{F}_0(X) = \mathcal{N}(X), \tag{4}$$

where $\mathcal{N}$ is the layer normalization [2, 60] without learning parameters. At the same time, other basic nodes will also use it to normalize the output of the unit for stable training.

4.2.2 Global Transformation Node. To align the text with the current query in latent representation space, and selectively suppress or highlight the representation from each semantic dimension, we propose a GTN to modify visual representations based on text. Specifically, the text representation ψ_m is mapped to generate the scaling and shifting vectors as follows:

$$\gamma = W_\gamma \psi_m + b_\gamma, \quad \beta = W_\beta \psi_m + b_\beta, \tag{5}$$

where $W_\gamma, W_\beta \in \mathbb{R}^{D\times D}$ are the learnable affine matrices. Then, the current query representation X is modulated by the following affine transformation operation, which can be formulated as follows:

$$\mathcal{F}_1(X) = \mathcal{N}(\gamma \odot (X) + \beta), \tag{6}$$

where $\odot$ denotes the element-wise product. This is different from AdaIN [24] and BatchNorm [25]. The former is used between stages of style transfer networks [45] and needs to normalize the input to remove the style of original image. The latter use the fixed parameters γ and β to all samples in inference. However, we adopt text-guided modulation on high-level representation and do not need to normalize the input. Additionally, γ and β are changing vectors with the input text. GTN is the unilateral modification of current query representation based on the feedback text, and the purpose is to achieve unified coarse-grained semantic adjustment to all visual local features. It is a passive operation of the text domain for adapting to the image domain.

4.2.3 Cross-modal Reasoning Node. Although GTN can directly modify the multimodal query but lacks some flexibility and overall understanding. Therefore, we introduce the CRN to unify the two domain features into the same latent space and further perform the holistic refinement. It is the fine-grained adjustment of query features by jointly interacting with feedback text. Specifically, we first concatenate current query features and textual features in dimension and then project them to different spaces (Q, K, and V) after dimensionality reduction representation:

$$X_c = W([X; \psi_m]), \tag{7}$$

$$Q = X_c W_Q, \quad K = X_c W_K, \quad V = X_c W_V, \tag{8}$$

where $W \in \mathbb{R}^{2D\times D}$ and $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$. The semicolon (;) denotes the concatenate operation across the feature dimension.

Then, the multi-headed attention mechanism [56] captures the dependencies of different spatial elements of the feature map under multiple subspaces. With the help of the weighted combination of spatial elements and following the feed-forward network (FFN), CRN can purposefully enhance the corresponding semantic dimensions. It can be formulated as follows:

$$MultiHead(X) = [head_1; \ldots; head_h], \tag{9}$$

$$head_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \tag{10}$$

$$\mathcal{F}_2(X) = \mathcal{N}(FFN(MultiHead(X)) + MultiHead(X)), \tag{11}$$

where h denotes the number of heads. Q_i denotes the i-th head divided from Q along the feature dimension. FFN denotes the two-layer multi-layer perceptron with ReLU activation function.

4.2.4 Composition Router. As the commander of our multi-expert collaborative network, the composition router conducts the scheduling of three nodes of the current layer for further modification and completion. Based on the current intermediate multimodal representation and textual features, the router unit performs different degrees of activation for the paths of three nodes and aggregates them into the next layer of input. The path activation values of the router unit for each layer can be represented as follows:

$$\alpha^l = \mathcal{R}^l(X^{l-1}, \psi_m), \tag{12}$$

where $\mathcal{R}^l(\cdot,\cdot)$ represents the router function in the l-th composition layer. The output feature of l-th composition layer can be formulated as follows:

$$X^l = \begin{cases} \phi_r, & l = 1 \\ \sum_{i=1}^{3} \alpha_i^l P_i^l, & l > 1 \end{cases} \tag{13}$$

where α_i^l and P_i^l respectively represent the activation value and output intermediate feature of the i-th node in the l-th composition layer. The router for each composition layer can be expressed in the following functional form concerning input $X \in \mathbb{R}^{K\times D}$and $\psi_m \in \mathbb{R}^D$:

$$r = [X^l; \psi_m] \in \mathbb{R}^{K\times 2D}, \tag{14}$$

$$\Upsilon = \mathcal{N}\left(W_1\left(\frac{1}{K}\sum_{i=1}^{K} r_i\right)\right), \tag{15}$$

$$\mathcal{R}^l(x, y) = \sigma(W_2(\xi(\Upsilon))), \tag{16}$$

where σ denotes the sigmoid function and ξ denotes the ReLU activation function. $W_1 \in \mathbb{R}^{2D\times\frac{D}{2}}$ and $W_2 \in \mathbb{R}^{\frac{D}{2}\times 3}$.

4.3 Loss Function

To push the composed query representation close to its target representation and away from other unmatched images in our training latent space, following other works [28, 31, 57], we adopt **Batch-based Classification (BBC)** loss. It is also called *InfoNCE loss* [55] in the self-supervised learning fields, and early works [10, 31, 57] show that it is more discriminative and has faster convergence than the triplet loss in this task:

$$\mathcal{L}_{BBC} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{exp\{\kappa(f_q^i, f_t^i)/\tau\}}{\sum_{j=1}^{B} exp\{\kappa(f_q^i, f_t^j)/\tau\}}, \tag{17}$$

where B is the size of the mini-batch and τ is a learnable temperature parameter. κ can be the arbitrary similarity kernel that is implemented by the cosine similarity distance in our experiments.

Moreover, to regularize the network and supervise router training, we propose BSC loss to constrain the consistency of the internal relationship of the intermediate representation and corresponding target representation. The reason we do not directly input the corresponding target representation to the router and learn the supervised constraint is that it will entangle the query and target processing branches like Figure 1(c) and increase the retrieval time complexity from the original $O(N)$ to $O(N^2)$. Specifically, we first obtain the pooling vectors $\bar{X^l} \in \mathbb{R}^{1\times D}$ from the output of each composition layer and then concatenate and normalize them to unit vector v_q of which

L is the maximum number of composition layers:

$$\bar{X^l} = \frac{1}{K}\sum_{i=1}^{K}(X^l), \tag{18}$$

$$v_q = \left\|[\bar{X^1}; ...; \bar{X^L}]\right\|_2 \in \mathbb{R}^{1\times(L*D)}. \tag{19}$$

Then, we construct the query graph $G_q(v_q^i, v_q^j) \in \mathbb{R}^{B\times B}$ by calculating the affinity edge of each sample pair (v_q^i, v_q^j) in the mini-batch. The target graph $G_t(f_t^i, f_t^j)$ is constructed similarly. We formulate the regularization as follows:

$$G_q(v_q^i, v_q^j) = (v_q^i)^T v_q^j, \quad G_t(f_t^i, f_t^j) = (f_t^i)^T f_t^j, \tag{20}$$

$$\mathcal{L}_{BSC} = \left\|G_q(v_q^i, v_q^j) - G_t(f_t^i, f_t^j)\right\|_2^2. \tag{21}$$

Finally, the total objective function is the weighted sum of the preceding BBC and BSC losses:

$$\mathcal{L} = \mathcal{L}_{BBC} + \lambda\mathcal{L}_{BSC}, \tag{22}$$

where λ is the balance hyperparameter.

5 EXPERIMENTS

5.1 Datasets

To demonstrate the effectiveness and practicability of the proposed method on the text-guided image retrieval task, we evaluate our model on two widely used real-world benchmark datasets: Fashion IQ [18] and Shoes [3]. *Fashion IQ* [18] is a natural language based interactive fashion product retrieval dataset introduced by the Fashion IQ challenges at the ICCV 2019 and CVPR 2020 workshops. There are 77,684 fashion images of three categories: Dress, Shirt, and Tops & Tees (Toptee) from Amazon.com. The average text length is 10.69 words. Since the ground truth is not publicly available, only the training set of 18,000 triplets and the validation set of 6,016 triplets are available. We follow the same evaluation protocol as that of Guo et al. [18], using the same training split and evaluating on the validation set. *Shoes* is a dataset crawled from Like.com and has 10,000 images of shoes for training and 4,658 images for test. The average text length is 5.32 words. All text pieces of these datasets are annotated by humans as the real user feedback of the e-commerce scenario.

5.2 Experimental Settings

Implementation Details. For a fair comparison with most of similar methods [5, 28, 31, 58], we use ResNet50 [21] pertaining to ImageNet [8] as our visual backbone and only choose the feature map of the last stage as the high-level visual representation of the reference and target images. For textual representation, as in other works [5, 28], our textual encoder is composed of an embedding layer initialized by 300-dimension GloVe word embedding [46] and LSTM [23] with 1024 hidden units. For training, the Adam optimizer [29] with weight decay factor 1e-6 is used to train the model with 50 epochs. The learning rate is set as 1e-4 and decays by factor 10 every 25 epochs. Consistent with prior works [5, 28, 31, 58], the mini-batch size is 32 and the dimensions of shared embedding are 1024. λ is set to 1.0, and the number of heads h is 8. All experiments are implemented in PyTorch with one RTX 3090 GPU, and we fixed the random seeds to ensure reproducibility.

Evaluation Metric. Following the convention of other works [5, 31], we use the standard evaluation protocol for each task and measure the performance with **Recall@K (R@K)**, which is defined as the percentage of queries for which the ground truth item is ranked among the top-K retrieved items. Especially, we report the average value of R@K as the overall metric to compare with other methods.

5.3 Model Comparison

To verify the effectiveness of our method in text-guided image retrieval in the fashion domain, we compare our method with the following representative and recent methods on benchmark datasets. Additionally, we briefly classify and illustrate these methods according to their commonalities:

- *Multi-level visual representation*: The VAL [5], MCR [70], SAC [26], MAAF [10], and DATIR [16] methods use multi-level feature fusion or hierarchical matching to obtain more accurate similarity scores. However, multi-level feature fusion will incur additional calculation and hierarchical matching will bring the multiplied consumption of database storage with the multi-level intermediate features.
- *Combination of text-to-image (t2i) retrieval and image-to-image (i2i) retrieval*: The DCNet [28] and ARTEMIS [7] methods split the text-guided image retrieval into the combination of two retrieval tasks and need to twice access to all images in the database. However, it causes extra consumption of calculation in the real deployment.
- *Auxiliary information from pretraining model or dataset*: The CIRPLANT method [35] uses OSCAR [32], which is pretrained in the ImageNet-domain dataset as the visual and textual backbones, and SAC [26] uses the pretrained BERT [9] as the textual backbone. CurlingNet [67] uses the Fashion200k and Fashion-Gen datasets [50] to pretrain the visual encoder.
- *Two-step modification*: The CoSMo [31] method modulates the content and style of reference images based on the text. SAC [26] uses two-step textual feature fusion to achieve localization and modification of the representations of reference images.
- *Additional constraints*: JPM [62] proposes alignment constraints between visual and textual domains, and further MCR [70] hopes to predict words of text by visual information by the image caption task. Furthermore, DATIR [16] considers maximizing the mutual information between the text representation and its semantically consistent visual representation.
- *Multiple subnetwork method*: CLVC-Net [58] contains two mutually enhancement subnetworks: local-wise and global-wise. Each subnetwork has independent image-text composition units and vision-language encoders. In inference, the final calculated similarity score is relatively more accurate by the joint judgment from the two subnetworks, but its storage consumption and computation are also doubled.

We also compare with the Fashion IQ 2019/2020 Workshop (Fashion IQ -W) winners [28, 54, 67] following the standard experimental settings. We additionally note that some methods [5, 70] adapt the different training strategies for different benchmarks because of the difference in data distribution (i.e., product category and text length), but we conduct the same training settings for all datasets to emphasize the validity of the adaptive multi-expert collaborative methodology.

5.3.1 Results on Fashion IQ . Table 1 presents our results on the Fashion IQ dataset. It can be noted that our model significantly outperforms the existing methods in all metrics. In the overall average R@K (K = 10,50) metric, our method achieves superior performance to the state-of-the-art methods and outperforms the most competitive method with an obvious margin (i.e., by about 2.7%). Additionally, without bells and whistles, our AMC keeps the uncoupled architecture and does not resort to multi-level matching [5, 16], multiple accesses [7, 28], multiple subnetworks [58], and bloated pretrained encoders [26, 35]. Evidently, it proves that performance improvement does not necessarily require complex similarity computation at the cost of inference efficiency, but it can be achieved as well by just transforming the static composition module into the data-adaptive multi-expert collaborative fusion. Moreover, it is worth noting that the performance gains from this new dynamic modeling ideology are more cost-effective, and we further illustrate it by comparing the model parameters, inference speed, and generalization ability in a comprehensive manner in Section 5.3.3.

Table 1. Quantitative Comparison on the Fashion IQ Dataset

Method	Fashion IQ								
	Dress		Toptee		Shirt		Average		Average
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
FiLM [47] [AAAI 2018]	14.23	33.34	17.30	37.68	15.04	34.09	15.52	35.04	25.28
TIRG [57] [CVPR 2019]	14.87	34.66	19.08	39.62	18.26	37.89	17.40	37.39	27.40
Relationship [52] [NIPS 2017]	15.44	38.08	21.10	44.77	18.33	38.63	18.29	40.49	29.39
CIRPLANT [35] [ICCV 2021]	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53	30.20
VAL [5] [CVPR 2020]	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04	33.82
CurlingNet [67] [Fashion IQ -W 2019]	24.44	47.69	25.19	49.66	18.59	40.57	22.74	45.97	34.36
DATIR [16] [MM 2021]	21.90	43.80	27.20	51.60	21.90	43.70	23.70	46.40	35.05
JPM [62] [MM 2021]	21.38	45.15	27.78	51.70	22.81	45.18	23.99	47.34	35.66
MAAF [10] [Arxiv 2020]	23.80	48.60	27.90	53.60	21.30	44.20	24.30	48.80	36.60
ARTEMIS [7] [ICLR 2022]	25.68	51.25	21.57	44.13	28.59	55.06	25.25	50.08	37.68
RTIC [54] [Fashion IQ -W 2020]	28.21	51.41	28.00	55.58	21.30	44.80	25.83	50.59	38.22
MCR [70] [MM 2021]	26.20	51.20	29.70	56.40	22.40	46.00	26.10	51.20	38.65
CoSMo [31] [CVPR 2021]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31	39.45
DCNet [28] [AAAI 2021]	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89	40.83
SAC *w/ BERT* [26] [WACV 2022]	26.52	51.01	32.70	61.23	28.02	51.86	29.08	54.70	41.89
CLVC-Net* [58] [SIGIR 2021]	29.85	56.47	33.50	64.00	28.75	54.76	30.70	58.41	44.55
Our AMC	**31.73**	**59.25**	**36.21**	**66.60**	**30.67**	**59.08**	**32.87**	**61.64**	**47.25**

The star ($\star$) denotes that ensemble results were obtained from two subnetworks.

Table 2. Quantitative Comparison on the Shoes Dataset

Method	Shoes			
	R@1	R@10	R@50	Average
FiLM [47] [AAAI 2018]	10.19	38.39	68.30	38.96
TIRG [57] [CVPR 2019]	12.60	45.45	69.39	42.48
Relationship [52] [NIPS 2017]	12.31	45.10	71.45	42.95
VAL [5] [CVPR 2020]	16.49	49.10	73.53	46.37
CoSMo [31] [CVPR 2021]	16.72	48.36	75.64	46.91
MAAF [10] [Arxiv 2020]	16.45	49.95	76.36	47.58
DATIR [16] [MM 2021]	17.20	51.10	75.60	47.97
MCR [70] [MM 2021]	17.85	50.95	77.24	48.68
SAC *w/ BERT* [26] [WACV 2022]	18.50	51.73	77.28	49.17
ARTEMIS [7] [ICLR 2022]	18.72	53.11	79.31	50.38
DCNet [28] [AAAI 2021]	–	53.82	79.33	–
CLVC-Net* [58] [SIGIR 2021]	17.64	54.39	**79.47**	50.50
Our AMC	**19.99**	**56.89**	79.27	**52.05**

The star ($\star$) denotes that ensemble results were obtained from two models.

5.3.2 Results on Shoes. Table 2 shows the comparison of our method and other state-of-the-art methods on the Shoes dataset. The result further validates the effectiveness of our methods. Compared with other methods, our approach achieves considerable performance improvement and surpasses the previous best method with margins of 1.55% on average R@K (K = 1,10,50). Additionally, we noticed that our method does not significantly exceed other methods in R@50, which we attribute to the fact that the short average text length (5.32 words) in the Shoes dataset may not accurately represent the distinction of user intent and can easily lead to overfitting. However, we can still observe the significant performance superiority when K is relatively small (R@1, R@10).

Table 3. More Comparison Experiments with CLVC-Net on the Fashion IQ Dataset

Metric	CLVC-Net* [58]	Our AMC	Comparison
Model Params. (M)	98.59	52.59	×**0.53**
Retrieval Time (s)	21.06	6.98	×**3.01**
Dress → Dress	43.67	45.49	↑ 1.82
Dress → Shirt	18.89	22.00	↑ 3.11
Dress → Toptee	30.67	34.75	↑ 4.08
Toptee → Toptee	48.75	51.41	↑ 2.66
Toptee → Dress	25.58	30.52	↑ 4.94
Toptee → Shirt	33.53	36.61	↑ 3.08
Shirt → Shirt	41.76	44.88	↑ 3.12
Shirt → Dress	16.24	18.67	↑ 2.43
Shirt → Toptee	33.83	37.56	↑ 3.73

"Dress → Shirt" denotes the overall average results of the model trained in the Dress subset but tested in the Shirt subset. Others are similarly marked.

Furthermore, it can be seen that these methods have different performance rankings on different datasets, which implies that their methods cannot work flexibly with various image-text queries and suffer from limited generalization ability. However, our method achieves superior performance on these datasets. It further proves the superiority and generalization of our AMC, which can adaptively fuse the reference image and user intent in an elastic way.

5.3.3 More Results. In Table 3, we further compare our method with the public pretrained model of the previous best method CLVC-Net [58] in terms of model size, retrieval speed, and zero-shot generalization capability. We estimate the model size and retrieval speed by averaging measurements on three test subsets of Fashion IQ . The retrieval speed is measured by the inference time of the test set (the mini-batch size is 32), and we exclude the influence of data loading time. The results show that our model parameters are roughly half of the comparison method and our retrieval speed is three times faster. This demonstrates that our method is scalable, efficient, and promising for real-world applications. Moreover, we compare the zero-shot generalization capability by transfer testing across subsets, and the results of Table 3 show that our method outperforms the comparison method in both homogenous representation capability and heterogenous transfer capability for three subsets. This demonstrates that our adaptive multi-expert collaboration method has both better performance and better generalization than the previous static methods.

5.4 Ablation Studies

We conduct ablation studies to investigate the impact of different configurations for our AMC model. To facilitate comparison and analysis, all ablation studies adopt the same experimental settings. The different experimental variables are as follows:

- $l = 1$ *or* 3: To explore the effect of the number of composition layers, we compare the results of the different numbers of layers.
- *Triplet loss*: A few previous works [5, 16] use Triplet loss instead of the popular BBC loss, and we compare the results of different losses and keep the default setting with previous works [5, 16], which set the hyperparameter margin to 0.2.
- *W/o BSC loss*: To verify the effect of our BSC loss, we remove the BSC loss by setting $\lambda = 0$.
- *W/o NIN/GTN/CRN*: To investigate the importance of different nodes, we respectively remove each node from our network.

Table 4. Ablation Study on the Fashion IQ and Shoes Datasets

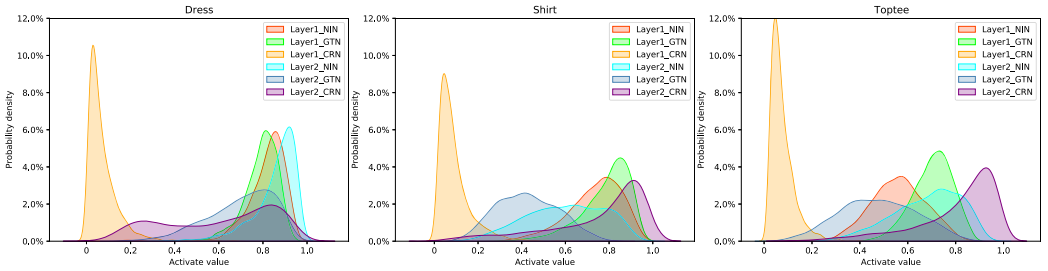
Method	Fashion IQ (Avg)		Shoes		
	R@10	R@50	R@1	R@10	R@50
Layer = 1	31.82	60.07	18.57	54.73	78.30
Layer = 3	32.57	61.05	**20.31**	55.98	78.53
Triplet loss	29.09	57.86	20.30	55.09	78.33
W/o BSC loss	31.91	60.99	20.19	55.46	78.67
W/o NIN	31.90	60.32	18.46	52.74	78.13
W/o GTN	30.25	58.35	18.72	55.30	78.95
W/o CRN	32.31	60.77	17.67	54.90	78.16
W/o Router	30.82	59.78	18.32	53.02	77.68
Random Router	30.15	58.84	15.00	50.87	75.43
Hard Router	25.76	51.86	12.50	44.56	71.46
Our AMC	**32.87**	**61.64**	19.99	**56.89**	**79.27**

The results of Fashion IQ are the average result of three subsets.

- *W/o Router*: We remove the Router units from our network, and it is equivalent to directly connecting all nodes without adaptive activation for experts.
- *Random Router*: We replace the path activation value of the routers with a uniform distribution ($R^l(x, y) \sim U(0, 1)$). This will activate each expert with random intensity.
- *Hard Router*: Like Herrmann et al. [22], we use Gumbel-Softmax [27] to discretize the path activation value to one-hot encoding while ensuring the network can be optimized by back-propagation. This will lead to the sparse activation of experts.

We can gain the following observations and analysis from the results in Table 4:

(1) *Number of layers*: The appropriately deep composition layers will improve performance but slightly degrade when the network is too deep. We suppose that the deeper composition layers also make the optimization more difficult. Considering the tradeoff between performance and model parameters, we finally set the number of composition layers to 2 in our experiment.
(2) *Different loss*: The results found that the performance of Triplet loss is slightly lower than the BBC loss in the Shoes dataset and significantly lower in the Fashion IQ dataset. We suppose that the hyperparameter margin of different datasets may have room for optimization.
(3) *Regularization loss*: Compared with our model, *w/o BSC loss* has decreased performance, and it demonstrates the effectiveness of our BSC loss to regularize network training through inter-sample relationships
(4) *Fusion nodes*: First, we find that removing any one of the three nodes will lead to performance degradation. Second, we find that GTN may be relatively important for the Fashion IQ dataset, whereas NIN may be relatively important for the Shoes dataset. We speculate that this is due to differences in the datasets, and the Shoes dataset may be more prone to overfitting. However, this illustrates the flexibility and generalization of our network, which can be adapted to different datasets by the collaboration and adjustment of multiple experts.
(5) *Router settings*: The comparison with *w/o Router* shows that our AMC is more effective than the static full-connected nodes version. The reason is that our soft router can activate paths adaptively based on the current input data. Compared with *Random Router*, the better performance indicates that our router learns from data how to selectively activate the node path. The performance comparison with *Hard Router* shows that pure pursuit of path streamlining



(a) Dress (b) Shirt (c) Toptee

Fig. 4. Statistical visualization of the path activation values. NIN of the first composition layer is denoted as "Layer1_NIN" and similarly for the others.

will lead to a significant performance degradation. Multi-expert collaboration is more effective than single-expert activation for each composition layer.

6 VISUALIZATION

6.1 Path Activation Values

In the Fashion IQ test set, we perform statistical analysis for each node activation value of our two-layer method AMC. Specifically, we perform inference with AMC for test subsets and calculate the probability density curve of node activation values. The results are shown in Figure 4, and we have the following observation and analysis from these results:

(1) For three test subsets, CRN owns overall smaller activation values over NIN and GTN in the first layer. We speculate that the multimodal queries are more inclined to transform for the whole semantic dimension by GTN at the primary interaction stage, and the model does not tend to perform complex CRN inference to supplement the representation. In the end-to-end optimization of deep learning networks, it is common for the network learning process to first prefer simple overall transformations and then complex local adjustment. Similar phenomena and more analysis can be found in the work of Geirhos et al. [12].
(2) In the second layer, except for the Dress subset, we find that the activation value of CRN is generally higher than that of GTN. We guess that the multimodal queries are mainly modified details by CRN at this stage. The role of GTN and CRN can be analogous to the low-frequency and high-frequency components of an image, which respectively represent the whole and the details of a multimodal query.
(3) NIN is relatively active at all layers of all subsets, which indicates that the proper shortcut is useful to optimize complex interactions of dynamic paths. It also bridges the shallow reference image features and deep multimodal intermediate features to selectively change some attributes of interest to the user while keeping other attributes of the reference image as consistent as possible.
(4) The distribution of path activation values for the three subsets are different from each other. This demonstrates that our approach can adaptively adjust path activation values according to different multimodal composing queries.

In Figure 5, we show the path activation values (the third column in Figure 5) for some retrieval examples from the Fashion IQ dataset. These path activation values are arranged in the order of the first and second composition layers from left to right for the first and second layers, and nodes NIN, GTN, and CRN from top to bottom. To show more intuitively, we respectively mark the high activation values higher than 0.5 in red and the low activation values less than 0.5 in blue.

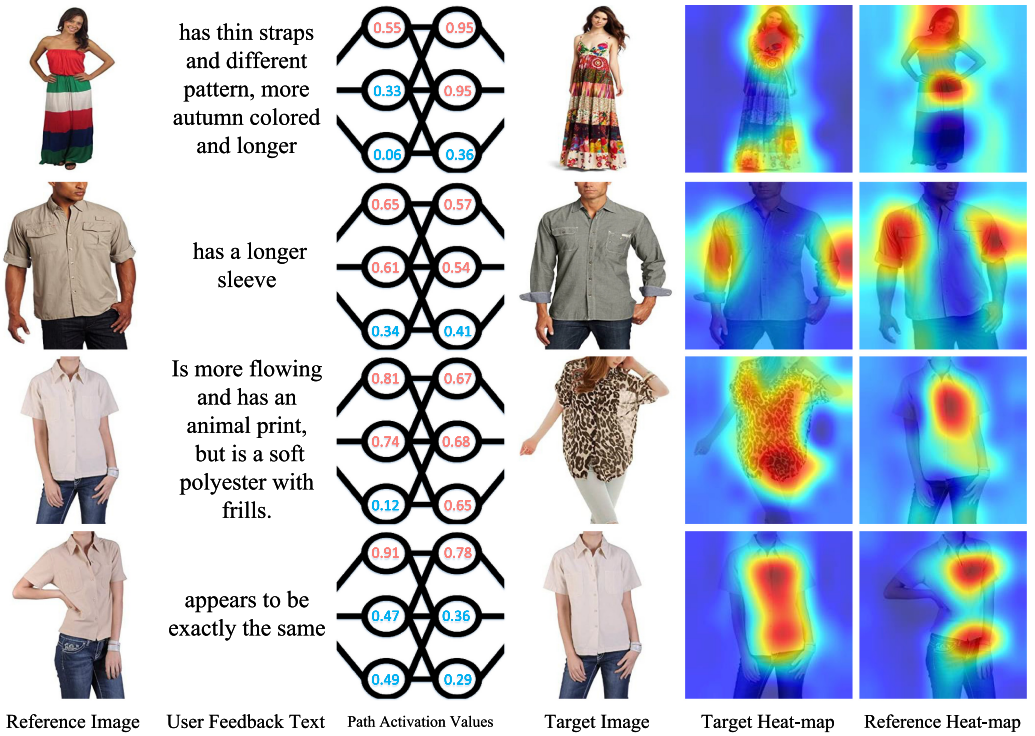


Fig. 5. Visualization examples from the Fashion IQ [18] dataset. We give the path activation values of the current sample in the order of nodes in the network architecture graph (see Figure 3). We also calculate the most similar dimension of the multimodal query representation and the target image representation, and we use the Grad-CAM algorithm [53] to locate the region of interest of this dimension in the reference image and the target image.

First, we can notice that each node is in different activation states for different multimodal queries. Second, we notice that the input text of the fourth row "appears to be exactly the same"—that is, the target image and the reference image own the same product but have different views. The only node with high activation is NIN, which mainly contains the identity connection with layer normalization. This shows that the path policy of our AMC for the current sample is to keep the features of the reference image as unchanged as possible. This policy is also consistent with the user intent of the current sample—that is, searching images that are consistent with the product in the reference image. To a certain extent, these results illustrate that our network indeed adaptively adjusts the activation of different nodes according to the current samples.

6.2 Semantic Dimension Localization

We also present the heat map of reference and target images obtained with the Grad-CAM algorithm [53] like in the work of Delmas et al. [7]. Specially, we first calculate the similarity score of multimodal query representation and target image representation and get the dimension d_i that contributes the most to the similarity score. After the backpropagation of d_i, we reweigh the feature maps of the last convolutional layer of the visual encoder with the pooled gradients of the same layer. Finally, the visual heat map for the last convolutional layer is used as the visual semantic localization tool to perceive the region of interest corresponding to the most similar dimension.

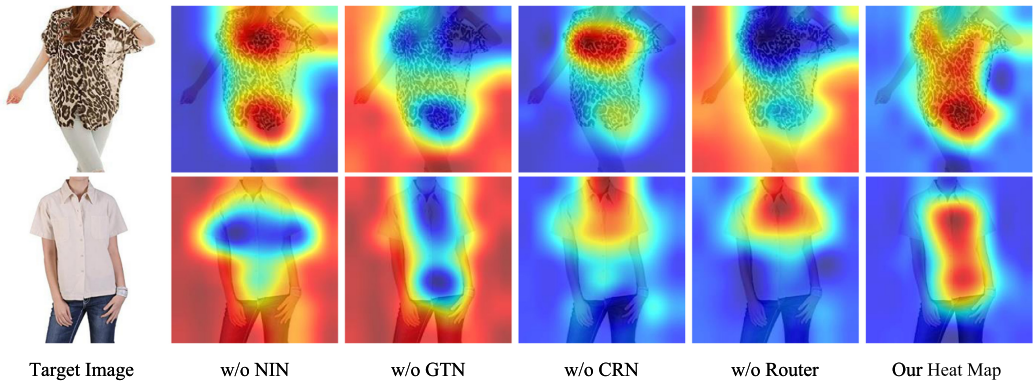


Fig. 6. The attention map visualization of the ablation study. For the examples shown in lines 3 and 4 in Figure 5, we visualized the region of interest to the target image for each trained model that removes the experts or router in the ablation study experiment.

We interpolate the heat map obtained by backpropagating d_i along the target visual branch and then linearly fuse it with the target image to obtain the "Target Heat Map" (the fifth column in Figure 5). Similarly, we interpolate the heat map obtained by backpropagating d_i along the image-text composition branch and then linearly fuse it with the reference image to obtain the "Reference Heat Map" (the sixth column in Figure 5).

By comparing the samples in Figure 5, it can be seen that the reference image and the target image have consistent semantic activation regions, and these regions are only related to a high-level semantic concept and not related to the absolute position of the images (i.e., the sleeve area of the two arms in the second row of Figure 5). Even if the text indicates the same product, it is still positioned in the area of the product rather than the human body, such as in the fourth row of Figure 5. Additionally, these semantic activation regions are highly matched with the complementary semantics of the user feedback text (i.e., the activated regions of Target Heat Map and "animal print" of text in the third row of Figure 5). This shows that our method clearly distinguishes the semantic concept of the animal print in the dress. As well, in the first row of Figure 5, we find that our method pays attention to the "thin straps" and "autumn color" regions of the target images. These observations illustrate that our AMC indeed focuses on semantic regions of reference and target images corresponding to the semantics of user feedback text.

6.3 Ablation Study Visualization

Because the three experts cooperate within each layer and stack across layers, it is difficult to get the visualization of the target image by testing each expert separately—that is, the expert NIN mainly communicates with the intermediate features of different layers and does not explicitly participate in the fusion with text features. Thus, corresponding to the setting of the ablation study in Table 4, we visualize the attention map of each model that can be removed by the expert or router. Additionally, we specifically selected examples with different texts corresponding to different target images. They are the examples shown in the third and fourth lines of Figure 5.

The ablation study visualization is shown in Figure 6, and we can see that removing any expert will lead to a big change in the area of attention. According to our analysis of the previous experiments, GTN controls the overall transformation. In Figure 6, we also can observe that the attention map of w/o GTN cannot directly locate the body of the garment. Additionally, the feedback text of the second example requires that the target image is exactly the same as the reference image,

Fig. 7. Quantitative examples of the Fashion IQ and Shoes datasets. We show the top-6 retrieved results. The blue and red boxes indicate the reference image and target image, respectively.

and it will result in maintaining the same features as possible. Thus, in the second example, we can observe that w/o NIN also cannot focus on the body of the garment. In addition, the router dynamically controls the activation of each expert, and w/o Router is equivalent to the fixed parameter model. It is can be observed that the attention areas of the fixed parameter model are less comprehensive than our dynamic model. Moreover, in the first example, the fixed parameter model also does not pay obvious attention to the garment itself. This means that our dynamic activation mechanism is more effective than the fixed model when all experts stay the same.

6.4 Quantitative Results

Figure 7 illustrates some quantitative results obtained by our AMC on the Fashion IQ [18] and Shoes [3] datasets. We report the top-6 retrieved images and use the blue and red boxes to mark the reference and target images. We can see that a big challenge for this task is the strong similarity between product images. For the reference image and corresponding user feedback text, the top-6 retrieved images have similar semantical information. Additionally, we note that our method does compose image and text semantics to retrieve target images, and the learned text complementary semantics contain color, style, texture, graphics, and so on (i.e., "animal print," "red stripes," "black and white flowers," and "ankle straps"). Compared with the rigid rules of the evaluation metrics that accurately retrieve only the correct retrieval item, searching for more candidates that match the user intent for further selection is a more realistic approach. Figure 7 illustrates that our method can also retrieve other similar product images when retrieving the target image.

7 CONCLUSION

Breaking with the conventional thinking of static modeling for the text-guided image retrieval task, we first proposed Adaptive Multi-expert Collaborative network (AMC) to achieve data-dependent path activation and data-adaptive multimodal fusion. Concretely, we designed three expert nodes with different advantages to fuse image-text representation and a composition router to dynamically coordinate these expert nodes. Attributed to this adaptive multi-expert cooperation mechanism, our model is more flexible than traditional static models in facing the challenges of various complex image-text combinations. Extensive experiments demonstrated that our approach not only achieves superior retrieval performance but also has excellent generalization capabilities. Additionally, our methods achieve a better tradeoff on performance and model parameters while ensuring retrieval efficiency. These results are expected to inspire more works to explore both effective and efficient text-guided image retrieval from the data-adaptive network adjustment perspective.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks), and the Ascend AI Processor used for this research. The codes using MindSpore (https://www.mindspore.cn) will also be released at https://github.com/KevinLight831/AMC.

REFERENCES

[1] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. 2018. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7708–7717.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] 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*. 663–676.

[4] Shaofeng Cai, Yao Shu, and Wei Wang. 2021. Dynamic routing networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3588–3597.

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

[6] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4 (March 2022), Article 95, 23 pages.

[7] Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, and Diane Larlus. 2021. ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity. In *Proceedings of the International Conference on Learning Representations*.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 248–255.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] 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* (2020).

[11] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[13] Qichuan Geng, Hong Zhang, Xiaojuan Qi, Gao Huang, Ruigang Yang, and Zhong Zhou. 2021. Gated path selection network for semantic segmentation. *IEEE Transactions on Image Processing* 30 (2021), 2436–2449.

[14] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. 2019. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1171–1180.

[15] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *Proceedings of the European Conference on Computer Vision*. 241–257.
[16] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. 2021. Image search with text feedback by deep hierarchical attention mutual information maximization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4600–4609.
[17] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS'18)*. 1–11.
[18] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019. The Fashion IQ dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794* (2019).
[19] 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*. 1463–1471.
[20] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2022. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2022), 7436–7456.
[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.
[22] Charles Herrmann, Richard Strong Bowen, and Ramin Zabih. 2020. Channel selection using Gumbel Softmax. In *Proceedings of the European Conference on Computer Vision*. 241–257.
[23] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[24] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
[25] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*. 448–456.
[26] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. 2022. SAC: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4021–4030.
[27] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144* (2016).
[28] 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 (AAAI'21)*. 1–9.
[29] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*.
[31] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 802–812.
[32] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*. 121–137.
[33] Zheng Li, Caili Guo, Xin Wang, Zerun Feng, Jenq-Neng Hwang, and Zhongtian Du. 2022. Unified loss of pair similarity optimization for vision-language retrieval. *arXiv preprint arXiv:2209.13869* (2022).
[34] Ruoyu Liu, Yao Zhao, Shikui Wei, Liang Zheng, and Yi Yang. 2019. Modality-invariant image-text embedding for image-sentence matching. *ACM Transactions on Multimedia Computing, Communications and Applications* 15, 1 (Feb. 2019), Article 27, 19 pages.
[35] 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/CVF International Conference on Computer Vision*. 2125–2134.
[36] Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, and Yaochen Li. 2020. Spatial-content image search in complex scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2503–2511.
[37] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1930–1939.
[38] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1930–1939.
[39] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. 2020. Fine-grained fashion similarity learning by attribute-specific embedding network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11741–11748.

[40] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. 2017. Spatial-semantic image search by visual feature synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4718–4727.

[41] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 4 (Nov. 2021), Article 128, 23 pages.

[42] Ravi Teja Mullapudi, William R. Mark, Noam Shazeer, and Kayvon Fatahalian. 2018. HydraNets: Specialized dynamic architectures for efficient inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8080–8089.

[43] Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. 59–68.

[44] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*. 3456–3465.

[45] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2337–2346.

[46] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*.

[47] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[48] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1113.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[50] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-Gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317* (2018).

[51] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8, 5 (1998), 644–655.

[52] Adam Santoro, David Raposo, David G. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, 4967–4976.

[53] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[54] Minchul Shin, Yoonjae Cho, and Seongwuk Hong. 2020. Fashion-IQ 2020 challenge 2nd place team's solution. *arXiv e-prints arXiv:2007.06404* (2020).

[55] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints arXiv:1807.03748* (2018).

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS'17)*. 1–11.

[57] 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/CVF Conference on Computer Vision and Pattern Recognition*. 6439–6448.

[58] 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*. 1369–1378.

[59] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. 2018. BlockDrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8817–8826.

[60] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the International Conference on Machine Learning*. 10524–10533.

[61] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2022. Interactive re-ranking via object entropy-guided question answering for cross-modal image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 3 (March 2022), Article 68, 17 pages.

[62] 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*. 3303–3311.

[63] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. FILIP: Fine-grained interactive language-image pre-training. In *Proceedings of the International Conference on Learning Representations*.

[64] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. 2018. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 300–317.

[65] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, and Alexander G. Hauptmann. 2018. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Transactions on Multimedia* 21 (2018), 1276–1288.

[66] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. 2016. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 799–807.

[67] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. 2020. CurlingNet: Compositional learning between images and text for Fashion IQ data. *arXiv e-prints arXiv:2003.12299* (2020).

[68] Feifei Zhang, Mingliang Xu, and Changsheng Xu. 2021. Geometry sensitive cross-modal reasoning for composed query based image retrieval. *IEEE Transactions on Image Processing* 31 (2021), 1000–1011.

[69] Feifei Zhang, Mingliang Xu, and Changsheng Xu. 2022. Tell, imagine, and search: End-to-end learning for composing text and image to image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 2 (March 2022), Article 59, 23 pages.

[70] Gangjian Zhang, Shikui Wei, Huaxin Pang, and Yao Zhao. 2021. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5353–5362.

[71] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2014. Attribute-augmented semantic hierarchy: Towards a unified framework for content-based image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 11, 1s (Oct. 2014), Article 21, 21 pages.

[72] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1520–1528.

[73] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. 2022. Uni-perceiver-MoE: Learning sparse generalist models with conditional MoEs. *arXiv preprint arXiv:2206.04674* (2022).

Received 17 August 2022; revised 5 January 2023; accepted 12 February 2023