ENCODER: Entity Mining and Modification Relation Binding for Composed Image Retrieval

Zixu Li[1], Zhiwei Chen[1], Haokun Wen[2,3], Zhiheng Fu[1], Yupeng Hu[1*], Weili Guan[2]

[1]School of Software, Shandong University
[2]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[3]School of Data Science, City University of Hong Kong
{Lzx, zivchen, fuzhiheng0215}@mail.sdu.edu.cn, huyupeng@sdu.edu.cn, {whenhaokun, honeyguan}@gmail.com

Abstract

The objective of Composed Image Retrieval (CIR) is to identify a target image that meets the requirement based on a multimodal query (including the reference image and the modification text) provided by the user. Despite the notable success of existing approaches, they fail to adequately address the modification relation between visual entities and modification actions. This limitation is non-trivial due to three challenges: *1) irrelevant factor perturbation*, *2) vague semantic boundaries*, and *3) implicit modification relations*. To address the above challenges, we propose an Entity miNing and modifiCation relatiOn binDing nEtwoRk (ENCODER), which has been designed to mine visual entities and modification actions, and then bind modification relations. Among the various components of the proposed ENCODER, we have initially designed the Latent Factor Filter (LFF) module to filter visual and textual latent factors related to modification semantics based on a threshold gating mechanism. Secondly, we propose Entity-Action Binding (EAB), which comprises modality-shared Learnable Relation Queries (LRQ) that are capable of mining visual entities and modification actions, as well as learning implicit modification relations for entity-action binding. Finally, the Multi-scale Composition module is introduced to achieve multi-scale feature composition, with guidance provided by entity-action binding. Extensive experiments on four benchmark datasets demonstrate the superiority of our proposed method.

1 Introduction

Composed Image Retrieval (Vo et al. 2019) (CIR) serves as an emerging image retrieval paradigm, with the objective of retrieving the target image based on a multimodal query. As illustrated in Figure 1(a), in the CIR task, the multimodal query comprises a reference image and a modification text. The reference image conveys the fundamental retrieval requirements, while the modification text articulates the user's specific modifications to the reference image. CIR has gained considerable attention (Chen et al. 2024; Han et al. 2023b; Yang et al. 2024) in recent years due to its capacity to express intricate retrieval requirements in a flexible manner. It indeed facilitates various applications, including multimodal recommendation (Li et al. 2024; Liu et al. 2024a), moment localization (Wang et al. 2024; Hu et al. 2021a,b, 2023), and dialogue robotics (Chen et al. 2023; Gosselin et al. 2022).

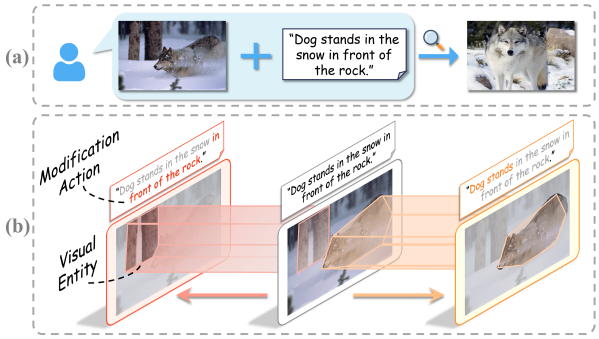

Figure 1: (a) provides an illustrative example of the CIR task. (b) illustrates the phenomenon of modification relation correspondence in the CIR task, whereby the modification text frequently comprises a series of modification actions, each of which is associated with a visual entity in the reference image through a corresponding modification relation. For example, the modification actions "in front of the rock" and "dog stands" correspond to the visual entities "trees" and "dog" in the reference image, respectively.

The key to CIR lies in accurately identifying the modification requirements and locating the corresponding regions in the reference image where these modifications should be made. It is also noteworthy that some pioneering efforts (Yang et al. 2023b; Deng et al. 2021; Zhang et al. 2023) have been made to address the issue of cross-modal semantic alignment of image regions and phrases, with encouraging results. However, due to the inherent text-visual *semantic asymmetry* of the CIR task (Zhao, Song, and Jin 2022), the entities included in the modification actions may not necessarily have a semantic correspondence in the reference image. In such cases, the effectiveness of the aforementioned approach is constrained. As illustrated in Figure 1(b), the conventional semantic alignment methodology is only capable of correlating the term "dog" (highlighted in orange) in the reference image with "dog" in the modification text. However, it is unable to establish a connection with the com-

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

prehensive modification action "dog stands" in the modification text, which is not conducive to the accurate modification of the reference image. Furthermore, the traditional semantic alignment method will also be unsuccessful in the case of "in front of the rock" in the modification text due to the absence of corresponding entities in the reference image. Consequently, in order to construct an effective CIR model, it is essential to consider not only the semantic alignment between visual and textual data but also the binding of modification relations. For ease of expression, in the following, we will refer to the visual entity in the reference image and the modification action in the modification text as "entity" and "action", respectively.

However, implementing modification relation binding between the entity and action is not straightforward due to the following challenges. 1) **Irrelevant factor perturbation**. Before binding the modification relation, it is first necessary to mine the entities and actions. However, not all words in the text and all regions of the image are directly related to the modification requirements, and irrelevant words and visual areas can affect the mining of entities and actions. Therefore, the primary challenge is to identify the visual and textual semantics related to the modification behavior and exclude the irrelevant factors from perturbation. 2) **Vague semantic boundaries**. Taking the modification text in Figure 1(b) as an example, the actions are hidden in various word combinations "the snow", "snow in front", and "in front of the rock", etc., and there is no clear supervisory signal to achieve accurate delineation. Furthermore, The boundaries of the entities are also difficult to recognize due to the irregularity of their shapes. In light of this, how to identify semantic boundaries to mine the entities and actions forms the second challenge. 3) **Implicit modification relations**. As there may be no semantic match between entities and actions, it is challenging to utilize feature similarity alone to measure the modification relation correspondence, and there is a dearth of direct supervisory signals. Consequently, how to identify modification relations and bind the entity with corresponding action is the third challenge.

To address the aforementioned issues, we propose an **E**ntity mi**N**ing and modifi**C**ation relati**O**n bin**D**ing n**E**two**R**k (ENCODER), which employs visual and textual semantics to filter latent factors and subsequently complete the mining and binding of visual entities and modification actions. In particular, we initially designed the *Latent Factor Filter (LFF)*, which calculates semantic relevance scores and filters visual and textual latent factors associated with modification semantics in accordance with a threshold gating mechanism. Secondly, we propose *Entity-Action Binding (EAB)*, which is capable of training modality-shared *Learnable Relation Queries (LRQ)* to mine entities and actions, and to learn implicit modification relations in order to achieve entity-action binding. Finally, we present *Multi-scale Composition (MSC)*, which is guided by entity-action binding to facilitate multi-scale feature composition and drive the composed feature closer to the corresponding target image.

In summary, our contributions include:

- In this paper, we present ENCODER, a novel CIR model that represents the first investigation into the modification relation binding between visual entities and modification actions.
- We put forward an LFF module, which facilitates visual and textual latent factor filtering. Furthermore, we propose an EAB module and devise the modality-shared LRQ, which is capable of mining entities and actions, as well as implementing entity-action binding.
- We conduct extensive experiments on four widely used benchmark datasets and demonstrate the superiority of the proposed ENCODER. Moreover, we have released our codes to facilitate other researchers[1].

2 Related Work

Our work is closely related to Composed Image Retrieval (CIR) and Visual Grounding for correspondence mining.

Composed Image Retrieval. Composed Image Retrieval (CIR) is a critical task in multimodal learning, which aims to retrieve the target image based on a reference image and modification text. Existing methods are generally considered to be divided into two categories. Conventionally, the first group of models (Vo et al. 2019; Chen et al. 2021; Gu et al. 2021; Wei et al. 2019; Zhu et al. 2023) utilizes traditional models (e.g. ResNet, LSTM) to extract image and text features, then compose the multimodal query. In contrast, another group of models (Han et al. 2023a; Wen et al. 2023, 2024) utilizes VLP-based models (e.g. CLIP (Radford et al. 2021)) to extract features of the multimodal query and then employ only simple feature alignment and composition strategies to achieve the outstanding performance (Baldrati et al. 2022a). In addition, zero-shot CIR is increasingly attracting interest from researchers (Gu et al. 2023, 2024; Zhang et al. 2024). Notwithstanding the considerable success of these approaches, they fail to consider the relation binding between visual entities and modification actions, which may result in inaccurate modification region localization. However, our model examines the latent factors associated with modification requirements and mines visual entities and modification actions from them, on the basis of which entity-action binding is implemented. The guidance of entity behavioral binding can facilitate the identification of more accurate multimodal feature compositions.

Visual Grounding for correspondence mining. Visual grounding (VG) aims to establish correspondences between visual entities in images and textual descriptions. Some conventional methods rely on text-image pairs and corresponding bounding boxes (Deng et al. 2023; Yang et al. 2023b), while some tend to perform attention mechanism (Deng et al. 2021; Zhang et al. 2023) and utilize pre-trained object detectors to align visual regions and textual phrases (Gupta et al. 2020; Chen et al. 2020). All of these works aim to mine correspondences between visual entities and textual descriptions, being widely used to enable more fine-grained entity-phrase correspondence. These advancements collectively provide a theoretical basis for mining multimodal correspondence for many downstream tasks, like CIR. How-

[1]https://sdu-l.github.io/ENCODER.github.io/

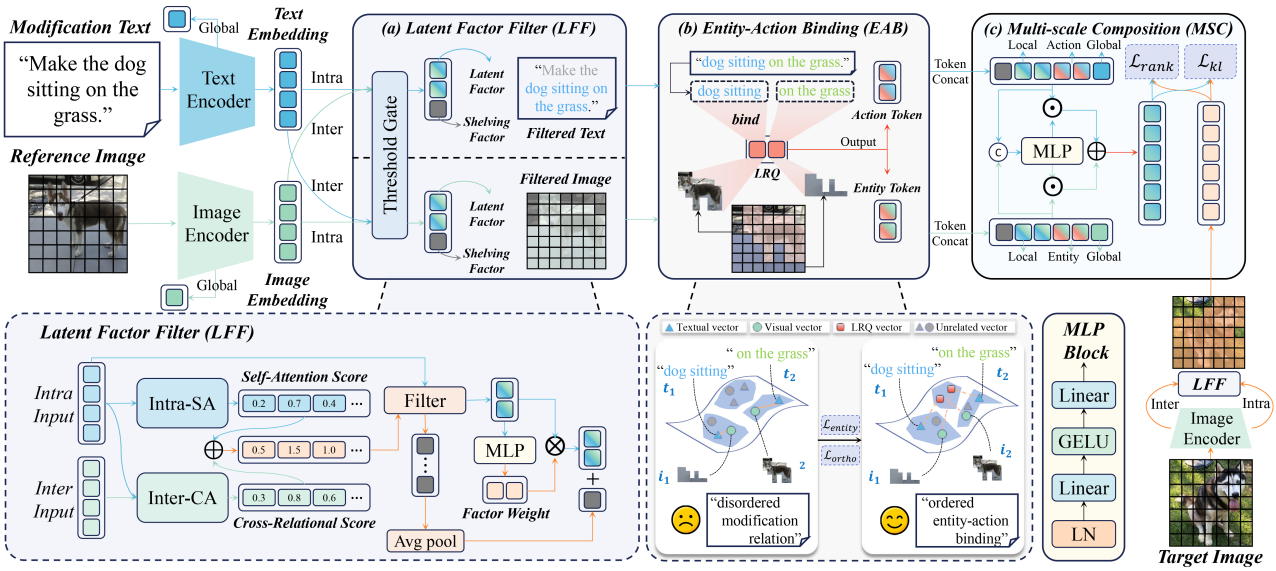

Figure 2: The proposed ENCODER consists of three key modules: (a) Latent Factor Filter, (b) Entity-Action Binding, and (c) Multi-scale Composition.

ever, most of the existing methods only explore the semantic alignment between visual entities and textual descriptions, which may fail in some complex scenarios, e.g., where what needs to be described in Composed Image Retrieval (CIR) is actually the modification relation. With this in mind, we are committed to mining modification relations in CIR that are different from semantic alignment.

3 ENCODER

As the primary innovation, our model (ENCODER) employs multimodal query semantics for latent factors filtering and implements cross-modal entity-action binding, thereby facilitating learning more accurate multimodal composed features. As illustrated in Figure 2, ENCODER consists of three critical modules: (a) *Latent Factor Filter (LFF)*, which calculates the cross-modal semantic relevance and filters both visual and textual latent factors associated with modification semantics, mitigating the perturbation of possible irrelevant factors. (b) *Entity-Action Binding (EAB)*, which aims to investigate the semantic boundaries of visual and textual latent factors so as to aggregate them into visual entities and modification actions, respectively, while learning the implicit modification relations between them, thereby achieving entity-action binding. (c) *Multi-scale Composition (MSC)*, which is guided by the entity-action binding relation to construct a multimodal query feature at multiple scales, with the objective of promoting the composed feature to the target image feature. In this section, we first formulate the CIR task and then elaborate on each module of ENCODER.

3.1 Problem Formulation

The Composed Image Retrieval (CIR) task aims to retrieve the target image that fulfills the multimodal query. Let $\mathcal{T} = \{(x_r, t_m, x_t)_n\}_{n=1}^{N}$ denote a set of N triplets, where x_r, t_m and x_t refer to the reference image, modification text and target image, respectively. Fundamentally, our goal is to optimize a metric space where the embedding of the multimodal query (x_r, t_m) should be as close as possible to the corresponding target image x_t, which is formulated as follows,

$$\mathcal{G}(x_r, t_m) \rightarrow \mathcal{G}(x_t), \tag{1}$$

where $\mathcal{G}$ denotes the to-be-optimized embedding function for both the multimodal query and the target image.

3.2 Latent Factor Filter (LFF)

To improve the accuracy of subsequent entity-action binding, we devise LFF to filter latent factors related to modification semantics in the reference image and modification text via inter-modal and intra-modal interaction, thereby mitigating irrelevant factor interference and facilitating subsequent aggregation for visual entity and modification action.

Specifically, we first utilize CLIP, which is proven effective for CIR tasks (Baldrati et al. 2022b), to extract global feature $\mathbf{E}_r^g \in \mathbb{R}^D$ and local feature $\mathbf{E}_r^l \in \mathbb{R}^{C \times D}$ of the reference image x_r, formulated as,

$$\mathbf{E}_r^g = \Phi_{\mathbb{I}}^g(x_r), \mathbf{E}_r^l = \mathrm{FC}_{\mathbb{I}}(\Phi_{\mathbb{I}}^l(x_r)), \tag{2}$$

where $\Phi_{\mathbb{I}}^g$, $\Phi_{\mathbb{I}}^l$ denote the last layer and penultimate layer of the CLIP image encoder, respectively. D is the CLIP global embedding dimension and $\mathrm{FC}_{\mathbb{I}}$ aligns local embedding dimension to D. Similarly, we obtain the global and local features of the target image ($\mathbf{E}_t^g \in \mathbb{R}^D$, $\mathbf{E}_t^l \in \mathbb{R}^{C \times D}$) and modification text ($\mathbf{E}_m^g \in \mathbb{R}^D$, $\mathbf{E}_m^l \in \mathbb{R}^{S \times D}$), where C, S represent the channel number of image and text, respectively.

Afterwards, to fully exploit cross-modal semantic associations, we compute cross-relational scores between each visual factor and textual factor via inter-modal cross-attention

and intra-modal self-attention. However, excessive attention to inter-modal interactions may lead to neglect of intra-modal information. To address this, we also compute the intra-modal self-attention score. Taking the reference image as an example, we regard the local feature $\mathbf{E}_r^l$ of the reference image as intra input, with the local feature of the modification text $\mathbf{E}_m^l$ as inter input. The intra input undergoes intra-modal self-attention (Intra-SA) to obtain Intra Factor Score $\mathbf{w}_s$, while Inter Factor Score $\mathbf{w}_c$ is calculated via inter-modal cross-attention (inter input as Query, intra input as Key and Value), formulated as,

$$\begin{cases} \mathbf{w}_s = \text{Intra-SA}(\mathbf{E}_r^l), \\ \mathbf{w}_c = \text{Inter-CA}(Q = \mathbf{E}_m^l, \{K, V\} = \mathbf{E}_r^l), \end{cases} \tag{3}$$

where $\mathbf{w}_s, \mathbf{w}_c \in \mathbb{R}^C$. Subsequently, to remove irrelevant factors with weak relevance to the modification semantic, we implement a threshold gating mechanism, which combines self-attention score $\mathbf{w}_s$ with the cross-relational score $\mathbf{w}_c$ to obtain $\mathbf{E}_r^{s\prime}$. Factors exceeding the specified threshold σ are retained as latent factors for subsequent entity-action binding. Meanwhile, we retain factors with combined scores below σ since they may appear as preserved regions in the target image, denoted as the shelving factor $\breve{\mathbf{E}}_r^s$ as follows,

$$\mathbf{E}_r^{s\prime} = \{\mathbf{E}_{ri}^l | (\mathbf{w}_s + \mathbf{w}_c)_i > \sigma\},\ \breve{\mathbf{E}}_r^s = \{\mathbf{E}_{ri}^l | (\mathbf{w}_s + \mathbf{w}_c)_i < \sigma\}. \tag{4}$$

Furthermore, to mitigate rigid boundaries of latent factors, which could result in some shelving factors being misclassified as latent factors, we employ an MLP followed by $\mathrm{Softmax}$. This adaptive weight sensing of latent factors enhances the weights of factors more relevant to modification semantics and yields reference latent factor $\mathbf{E}_r^{s\prime}$ as follows,

$$\mathbf{w}_f = \text{Softmax}(\text{MLP}(\mathbf{E}_r^{s\prime})^\top),\ \mathbf{E}_r^s = \mathbf{w}_f \cdot \mathbf{E}_r^{s\prime}, \tag{5}$$

where $\mathbf{w}_f \in \mathbb{R}^{P\times K}$ represents the factor weight, P denotes the number of latent factors to be learnt, $\mathbf{E}_r^s \in \mathbb{R}^{P\times D}$. In addition, to preserve the semantic information of the shelving factors without diverting focus from the latent factors, we employ average pooling on the shelving factors. Analogously, we utilize the local feature $\mathbf{E}_m^l$ of the modification text as intra input, and that $\mathbf{E}_r^l$ of the reference image as inter input for **LFF**, yielding modification latent factors $\mathbf{E}_m^s \in \mathbb{R}^{P\times D}$. Notably, to mitigate potential interference from modification text while maintaining its semantic independence, we resort to the local feature of the target image $\mathbf{E}_t^l$ as both the inter and intra inputs and solely apply self-attention to mine the target latent factors $\mathbf{E}_t^s \in \mathbb{R}^{P\times D}$.

3.3 Entity-Action Binding (EAB)

To mine visual entities and modification actions and relational bind them, we design **EAB** module. Through training modality-shared *Learnable Relation Queries (LRQ)*, EAB first mines semantic relations in the reference image and modification text to probe the semantic boundary and aggregate their latent factors into visual entities and modification actions, respectively. Afterwards, LRQ learns the implicit modification relation between visual entities and modification actions to assist entity-action binding as a medium.

Explicit Entity-Action Mining. Firstly, we initialize the learnable relation queries $\mathbf{R} \in \mathbb{R}^{E\times D}$, abbreviated as LRQ, which essentially serves as a set number (smaller than the number of latent factors) of learnable query embeddings, where E is the number of queries, and D is CLIP's embedding dimension. Then we utilize LRQ to capture semantic relations among latent factors, which serves as a medium to assist the aggregation of visual entities and modification actions. Considering the process of visual entity mining, the queries of LRQ first interact with each other, calculating self-attention weight $\mathbf{e}_s$. Meanwhile, LRQ $\mathbf{R}$ can additionally interact with reference latent factors $\mathbf{E}_r^s$, and yield the cross-attention weight $\mathbf{e}_c$. In formal terms, we have,

$$\mathbf{e}_s = \text{MLP}(\mathbf{R}\mathbf{R}^\top/\sqrt{D}),\ \mathbf{e}_c = \text{MLP}(\mathbf{R}\mathbf{E}_r^{s\top}/\sqrt{D}), \tag{6}$$

where $\mathbf{e}_s \in \mathbb{R}^{E\times E}, \mathbf{e}_c \in \mathbb{R}^{E\times P}$. Afterwards, based on $\mathbf{e}_c, \mathbf{e}_s$, the adaptive attention weights of the LRQ for the visual latent factors are obtained via $\mathbf{e}_r = \text{Softmax}(\mathbf{e}_s \cdot \mathbf{e}_c) \odot \mathbf{e}_c \in \mathbb{R}^{E\times P}$. Finally, to model the visual entities, we assign the above attention weight $\mathbf{e}_r$ to the reference latent factors and employ MLP to adaptively learn the entity weights, thereby mapping the latent factors to entity tokens, where each token matches the corresponding query of LRQ. In addition, to pave the way for binding visual entities and modification actions, we also employ LRQ for the textual modality, which is shared parameters with visual modality, and similarly obtain action tokens from modification latent factors as follows,

$$\mathbf{E}_r^e = \text{MLP}(\mathbf{e}_r \cdot \mathbf{E}_r^s),\ \mathbf{E}_m^e = \text{MLP}(\mathbf{e}_m \cdot \mathbf{E}_m^s), \tag{7}$$

where $\mathbf{E}_r^e, \mathbf{E}_m^e \in \mathbb{R}^{E\times D}$ are entity tokens and action tokens.

Implicit Modification Binding. Benefiting from the previous modality-shared LRQ, the process of obtaining entity tokens and action tokens aligns the semantic channels of both the visual entities and modification actions. Therefore, we utilize LRQ as a medium for binding entity-action implicit modification relations. However, there may be semantic overlap in the modification relations corresponding to different queries since LRQ contains multiple pairs of modification relations. To ensure semantic independence and make queries in LRQ concentrate on their own independent modification relation (i.e., entity-action pairs), we design a binding orthogonal regularization, formulated as follows,

$$\mathcal{L}_{ortho} = \|\mathbf{R}^\top\mathbf{R} - \mathbf{I}\|_F^2, \tag{8}$$

where $\mathbf{I} \in \mathbb{R}^{E\times E}$ and $\|\cdot\|_F^2$ is the Frobenius norm of matrix. So far, we have obtained a plausible LRQ medium where each query implies an independent implicit modification relation. In order to bind the corresponding entity-action, we separately compute the similarity distributions of both visual entities and modification actions with the LRQ and then promote them to be consistent. Specifically, let $\mathbf{b}_i^r = [b_{i1}^r, ..., b_{iE}^r]$ denote the similarity distribution of the i-th token of entity tokens with the query set of LRQ, where the similarity b_{ij}^r is obtained as follows,

$$b_{ij}^r = \frac{\exp\{\mathrm{s}(\mathbf{E}_{ri}, \mathbf{R}_j)/\tau\}}{\sum_{e=1}^E \exp\{\mathrm{s}(\mathbf{E}_{ri}, \mathbf{R}_e)/\tau\}}, \tag{9}$$

where $\mathrm{s}(\cdot)$ denotes the cosine similarity function, $\mathbf{E}_{ri}, \mathbf{R}_j$ represent the i-th token of entity tokens and the j-th query of

LRQ, respectively. Analogously, we obtain similarity distributions between each token in action tokens and each query of LRQ, denoted as $\mathbf{b}_i^m = [b_{i1}^m, ..., b_{iE}^m]$. Then we utilize KL divergence to converge the similarity distributions, formulated as follows,

$$\mathcal{L}_{bind} = \frac{1}{E}\sum_{i=1}^{E} D_{KL}\left(\mathbf{b}_i^m \| \mathbf{b}_i^r\right) = \frac{1}{E}\sum_{i=1}^{E}\sum_{j=1}^{E} b_{ij}^t \log \frac{b_{ij}^m}{b_{ij}^r}. \quad (10)$$

3.4 Multi-scale Composition (MSC)

Finally, to promote the multi-scale semantics perception of multimodal composed features, we integrate the multi-scale features from the reference image and the modification text to obtain the composed feature and push the composed feature close to the target image feature.

Specifically, we first concatenate the global feature, local feature, latent factors, and shelving factors, which correspond to each element of a triplet. For the reference image, the multi-scale factors are denoted as $\mathbf{E}_r = [\breve{\mathbf{E}}_r^s, \mathbf{E}_r^s, \mathbf{E}_r^e, \mathbf{E}_r^g] \in \mathbb{R}^{Q\times D} (Q{=}1{+}P{+}E{+}1)$. Similarly, we obtain multi-scale factors of modification text $\mathbf{E}_m = [\breve{\mathbf{E}}_m^s, \mathbf{E}_m^s, \mathbf{E}_m^e, \mathbf{E}_m^g] \in \mathbb{R}^{Q\times D}$, and multi-scale factors of target image $\mathbf{E}_t = [\breve{\mathbf{E}}_t^s, \mathbf{E}_t^s, \mathbf{E}_t^g] \in \mathbb{R}^{Q'\times D}$ ($Q' = 1 + P + 1$). We then employ MLP to perform multi-scale interactions on $\mathbf{E}_r, \mathbf{E}_m$ to learn the respective modification weights of the reference image and modification text, formulated as follows,

$$\mathbf{W} = \mathrm{MLP}([\mathbf{E}_r, \mathbf{E}_m]), \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{Q\times 2D}$ denotes the modification weight. Then we utilize chunk operation to split it into $\mathbf{W}_r, \mathbf{W}_m \in \mathbb{R}^{Q\times D}$ and aggregate them to multi-scale factors of the reference image and modification text, respectively. Finally, the aggregated factors are summed to obtain the final multimodal composed feature $\mathbf{E}_c \in \mathbb{R}^{Q\times D}$, formulated as follows,

$$\mathbf{E}_c = \mathbf{W}_r \mathbf{E}_r + \mathbf{W}_m \mathbf{E}_m. \quad (12)$$

Moreover, we employ the batch-based classification loss commonly utilized in the CIR task to push the composed features close to the target image feature as follows,

$$\mathcal{L}_{rank} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{ti}\right)/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{tj}\right)/\tau\right\}}\right\}, \quad (13)$$

where as $\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{ti}$ indicate the average pooled $\mathbf{E}_c, \mathbf{E}_t$ of the i-th triplet, respectively. B represents batch size, and $\mathrm{s}(\cdot)$ denotes the cosine similarity. To optimize composed feature space, we promote consistency in the similarity distribution between each composed feature and all target images in the batch. Specifically, let $\mathbf{s}_i^c = [s_{i1}^c, ..., s_{iB}^c]$ represent the similarity distribution of the i-th composed feature, where the similarity s_{ij}^c with the j-th target image is computed as,

$$s_{ij}^c = \frac{\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{tj}\right)/\tau\right\}}{\sum_{b=1}^{B}\exp\left\{\mathrm{s}\left(\bar{\mathbf{E}}_{ci}, \bar{\mathbf{E}}_{tb}\right)/\tau\right\}}. \quad (14)$$

Similarly, we can obtain the similarity distribution between the i-th target image feature and the other target image features in a batch, denoted as $\mathbf{s}_i^t = [s_{i1}^t, ..., s_{iB}^t]$. Subsequently, we use KL divergence to converge these two similarity distributions, formulated as follows,

$$\mathcal{L}_{kl} = \frac{1}{B}\sum_{i=1}^{B} D_{KL}\left(\mathbf{s}_i^t \| \mathbf{s}_i^c\right) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} s_{ij}^t \log \frac{s_{ij}^t}{s_{ij}^c}. \quad (15)$$

Finally, integrating all modules above, we parameterize the final objective function as follows,

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \left(\mathcal{L}_{rank} + \kappa\mathcal{L}_{kl} + \gamma\mathcal{L}_{bind} + \mu\mathcal{L}_{ortho}\right), \quad (16)$$

where $\mathbf{\Theta}^*$ is the to-be-optimized parameter for ENCODER and κ, γ, μ are the trade-off hyper-parameters.

4 Experiments

This section delves into our comprehensive experiments of ENCODER and the corresponding analyses.

4.1 Experimental Settings

Datasets. Following previous works, we chose four benchmark datasets for evaluation, including three fashion-domain datasets, FashionIQ (Wu et al. 2021), Shoes (Guo et al. 2018), Fashion200K (Han et al. 2017) and an open-domain dataset CIRR (Liu et al. 2021b).

Implementation Details. ENCODER is built upon the pre-trained CLIP (Radford et al. 2021) (ViT-B/32 version). We trained ENCODER using the AdamW optimizer with the initial learning rate of 5e-5, while the batch size is set to 128 and the learning rate for CLIP is 1e-6. Empirically, we maintained a consistent embedding dimension D of 512 throughout the network. We set the latent factor number P to 4 and the query number E of LRQ to 3. We also adopt the temperature factor τ to 0.1 for Eqn.(9,13,14). Through a comprehensive grid search, we set $\kappa = 0.8, \gamma = 0.5$, and $\mu = 0.5$ for all four datasets. All experiments were conducted on a single NVIDIA Tesla T4 GPU with 16GB memory and trained 10 epochs.

Evaluation. We implemented widely accepted evaluation standards, with Recall@k (short for R@k) serving as the key indicator. For Shoes and Fashion200K, we calculated R@k ($k = 1, 10, 50$) and their mean value. FashionIQ evaluation used R@10, R@50, and their category-wise averages. CIRR assessment included R@k (k=1, 5, 10, 50), R_{subset}@k (k= 1, 2, 3), and the average of R@5 and R_{subset}@1.

4.2 Performance Comparison

We conducted an extensive comparative analysis of ENCODER against two categories of CIR baselines: conventional model-based baselines (TIRG (Vo et al. 2019), CLVC-Net (Wen et al. 2021), etc.) and CLIP-based baselines (CLIP4CIR (Baldrati et al. 2022a), SSN (Yang et al. 2024), etc.). Through the results in Table 1, Table 2 and Table 3, we have the following three observations. 1) VLP-based models generally perform better than conventional models, which confirms the powerful feature extraction capability of the VLP model and its effectiveness on the CIR task. 2) ENCODER consistently surpasses all baselines on FashionIQ, Shoes, Fashion200K, and CIRR. Notably, ENCODER achieves 19.8% improvements over the

Method	FashionIQ								Shoes			
	Dresses		Shirts		Tops&Tees		Avg					
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@1	R@10	R@50	Avg
TIRG (Vo et al. 2019)	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	12.60	45.45	69.39	42.48
CIRPLANT (Liu et al. 2021a)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	-	-	-	-
CLVC-Net (Wen et al. 2021)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	17.64	54.39	79.47	50.50
FashionVLP (Goenka et al. 2022)	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	-	49.08	77.32	-
ARTEMIS (Delmas et al. 2022)	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	18.72	53.11	79.31	50.38
EER (Zhang et al. 2022)	30.02	55.44	25.32	49.87	33.20	60.34	29.51	55.22	20.05	56.02	79.94	52.00
CRN (Yang et al. 2023a)	32.67	59.30	30.27	56.97	37.74	65.94	33.56	60.74	18.92	54.55	80.04	51.17
ComqueryFormer (Xu et al. 2023)	33.86	61.08	35.57	62.19	42.07	69.30	37.17	64.19	-	-	-	-
MGUR (Chen et al. 2024)	32.61	61.34	33.23	62.55	41.40	72.51	35.75	65.47	18.41	53.63	79.84	50.63
SyncMask (Song et al. 2024)	33.76	61.23	35.82	62.12	44.82	72.06	38.13	65.14	-	-	-	-
LF-CLIP (Baldrati et al. 2022b)	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	-	-	-	-
CLIP4CIR (Baldrati et al. 2022a)	33.81	59.40	39.99	60.45	41.41	65.37	38.40	61.74	-	-	-	-
Prog. Lrn. (Zhao, Song, and Jin 2022)	38.18	64.50	48.63	71.54	52.32	76.90	46.38	70.98	22.88	58.83	84.16	55.29
FashionSAP (Han et al. 2023b)	33.71	60.43	41.91	70.93	33.17	61.33	36.26	64.23	-	-	-	-
FAME-ViL (Han et al. 2023a)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	-	-	-	-
BLIP4CIR (Liu et al. 2024b)	40.65	66.34	40.38	64.13	46.86	69.91	42.63	66.79	-	-	-	-
BLIP4CIR+Bi (Liu et al. 2024b)	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	-	-	-	-
SSN (Yang et al. 2024)	34.36	60.78	38.13	61.83	44.26	69.05	38.92	63.89	-	-	-	-
ENCODER(Ours)	**51.51**	**76.95**	**54.86**	**74.93**	**62.01**	**80.88**	**56.13**	**77.59**	**26.97**	**65.59**	**86.48**	**59.68**

Table 1: Performance comparison on FashionIQ and Shoes relative to R@k(%). The overall best results are in bold, while the best results over baselines are underlined.

Method	R@k				R$_{subset}$@k			(R@5+R$_{subset}$@1)/2
	k=1	k=5	k=10	k=50	k=1	k=2	k=3	
TIRG (Vo et al. 2019)	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
CIRPLANT (Liu et al. 2021a)	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS (Delmas et al. 2022)	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
ComqueryFormer (Xu et al. 2023)	25.76	61.76	75.90	95.13	51.86	76.26	89.25	56.81
LF-CLIP (Baldrati et al. 2022b)	33.59	65.35	77.35	95.21	62.39	81.81	92.02	63.87
CLIP4CIR (Baldrati et al. 2022a)	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR (Liu et al. 2024b)	40.17	71.81	83.18	95.69	72.34	88.70	95.23	72.08
BLIP4CIR+Bi (Liu et al. 2024b)	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
SSN (Yang et al. 2024)	43.91	77.25	86.48	97.45	71.76	88.63	95.54	74.51
ENCODER(Ours)	**46.10**	**77.98**	**87.16**	**97.64**	**76.92**	**90.41**	**95.95**	**77.45**

Table 2: Performance comparison on CIRR with respect to R@k(%) and R$_{subset}$@k(%). The overall best results are in bold, while the best results over baselines are underlined.

best baseline for R@10 on FashionIQ-Avg, 17.9% for R@1 on Shoes, 6.3% for R$_{subset}$@1 on CIRR, and 6.8% for R@1 on Fashion200K, respectively. These demonstrate ENCODER's effectiveness and generalization ability in both fashion-domain and open-domain CIR datasets. 3) While SSN and BLIP4CIR show sub-optimal performance on open-domain dataset CIRR for some metrics, they struggle to match previous SOTA's capabilities on FashionIQ. This may be due to their lack of the ability to easily adapt to a specific data domain. In contrast, ENCODER achieves the optimal in both open-domain and fashion-specific scenarios, highlighting its broader multimodal understanding.

4.3 Ablation Study

To illuminate the pivotal role of each module in ENCODER, we conducted an in-depth comparison with its derivatives: **Derivative (a):** When evaluating LFF independently, only EAB and MSC are employed (**C#1**). Additionally, we employ abbreviations that only utilize LFF on reference images or modification texts (**C#2, C#3**). **Derivative (b):** When assessing EAB in isolation, only LFF and MSC are utilized

Method	R@1	R@10	R@50	Avg
TIRG (Vo et al. 2019)	14.10	42.50	63.80	40.13
CLVC-Net (Wen et al. 2021)	22.60	53.00	72.20	49.27
FashionVLP (Goenka et al. 2022)	-	49.90	70.50	-
EER (Zhang et al. 2022)	-	50.88	70.60	-
CRN (Yang et al. 2023a)	-	53.50	74.50	-
ComqueryFormer (Xu et al. 2023)	-	52.20	72.20	-
MGUR (Chen et al. 2024)	21.80	52.10	70.20	48.03
ENCODER(Ours)	**24.14**	**55.98**	**74.82**	**51.65**

Table 3: Performance comparison on Fashion200K with respect to R@k(%). The overall best results are in bold, while the best results over baselines are underlined.

(**C#4**). We also adopt abbreviations that link EAB exclusively to the reference image or modification text (**C#5, C#6**). **Derivative (c):** To gain deep insights into the effectiveness of both LFF and EAB, we remove both LFF and EAB (**C#7**). Finally, we explore MSC's impact on composing multimodal queries by solely removing MSC (**C#8**). **Derivative (d):** To validate the correctness of ENCODER's optimization process, we design abbreviations for our proposed loss functions by separately removing them in Eqn.(16) (**C#9, C#10, C#11**).

C#	Deriv.	LFF	EAB	MSC	FashionIQ-Avg		Shoes	CIRR
					R@10	R@50	Avg	Avg
1	(a)	Ref-Img	✓	✓	50.08	73.96	52.62	75.90
2			✓	✓	55.08	76.57	58.32	76.92
3		Mod-Text	✓	✓	52.21	74.72	54.01	76.15
4	(b)	✓		✓	49.06	73.25	53.03	75.74
5		✓	Ref-Img	✓	54.32	76.62	56.65	75.81
6		✓	Mod-Text	✓	53.18	75.98	55.80	76.96
7	(c)			✓	48.58	72.73	53.01	75.40
8		✓	✓		47.54	72.70	48.45	71.36
9	(d)	w/o $\mathcal{L}_{ortho}$			47.82	72.83	52.98	77.11
10		w/o $\mathcal{L}_{bind}$			48.50	73.21	52.88	76.46
11		w/o $\mathcal{L}_{kl}$			47.64	72.93	52.95	74.31
		ENCODER			**56.13**	**77.59**	**59.68**	**77.45**

Table 4: Ablation Studies of ENCODER with different components and various settings on FashionIQ, Shoes, and CIRR. Note that C# is the number of different configurations. And the "Ref-Img" and "Mod-Text" refer to the reference image and modification text, respectively.

From Table 4, we obtain the following observations. 1) w/o MSC (**C#8**) shows the worst performance among all variants, which reveals the importance of MSC in precisely composing multi-scale semantic information. 2) Removing either LFF or EAB (**C#1, C#4**) leads to a performance drop. Moreover, the model without neither LFF nor EAB (**C#7**) shows the worst performance among the first seven variants. This is reasonable since both LLF and EAB play crucial roles in ENCODER. 3) Applying LFF or EAB only to reference images or modification text (**C#2-C#3, C#5-C#6**) results in decreased performance, but it is still better than removing both LFF and EAB (**C#7**). This suggests that both components need to be applied to learn the entity-action relations comprehensively. 4) Both **C#9** and **C#10** perform worse than ENCODER, which proves the effectiveness of entity-action binding for semantic independence. 5) For $\mathcal{L}_{kl}$ in Eqn.(16) (**C#11**), removing it leads to poor performance. This may be because it requires guidance from the target image distribution for multimodal query learning.

To investigate the sensitivity of ENCODER to the latent factor number P and query number E of LRQ, we present results on FashionIQ, as shown in Figure 3. We observed that as the values of P or E increase, ENCODER's performance gradually reaches the optimal and then drops. This is reasonable because the model requires a certain number of factors or LRQs to capture the focus of the entity-action pairs. However, too large values may lead to irrelevant factor perturbations and chaotic entity-action pairs.

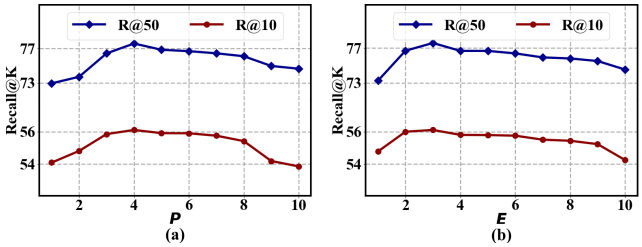

Figure 3: Sensitivity to (a) Latent Factor Number P and (b) Query Number E of LRQ on the FashionIQ dataset.

4.4 Case Study

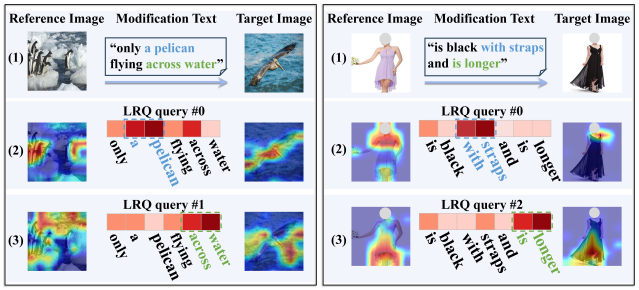

Figure 4: Attention visualization for LRQ on (a) CIRR and (b) FashionIQ datasets.

Figure 4 exhibits the attention distribution of the LRQ over the multimodal query and the target image on CIRR and FashionIQ datasets. For CIRR in Figure 4(a), it can be clearly observed that different LRQ queries focus on different entities and actions. For example, query #0 focused on the "penguin" in the reference image and the phrase "a pelican" in the modification text. The two themselves are semantically distant, however, LRQ succeeds in modeling the entity-action correspondence in modification relation. Also, query #1 successfully linked "ice" to "water". For FashionIQ in Figure 4(b), query #0 successfully acquired a focus on "straps" and the phrase "with straps", while query #2 gained attention to a different region that represents the "is longer" claim to "dress". The above results demonstrate that LRQ can focus on both the entity and the corresponding action.

5 Conclusion

In this paper, we designed ENCODER to understand the correspondence between visual entities and modification actions in the CIR task. In summary, our model comprises two main strategies. 1) Create a latent factor filtering module to mine multimodal semantics related to modification action. 2) Design the modality-shared LRQ that can be employed to mine entities and actions, and then learn the implicit modification relations that bind entity-action pairs. These strategies helped us to better compose multimodal features at multi-scales. Extensive experiments on four benchmark datasets demonstrated the superiority of our ENCODER.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, No.:62276155, No.:U24A20328, No.:62476071, No.:62206156, No.:62206157, and No.:624B2047; in part by the Natural Science Foundation of Shandong Province, No.:ZR2021MF040 and No.:ZR2022QF047; in part by the China National University Student Innovation & Entrepreneurship Development Program, No.:202410422071.

References

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022a. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4959–4968.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022b. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21466–21474.

Chen, X.; Song, X.; Peng, G.; Feng, S.; and Nie, L. 2021. Adversarial-enhanced hybrid graph network for user identity linkage. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*, 1084–1093.

Chen, X.; Song, X.; Wei, Y.; Nie, L.; and Chua, T.-S. 2023. Dual Semantic Knowledge Composed Multimodal Dialog Systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1518–1527. ACM.

Chen, Y.; Zheng, Z.; Ji, W.; Qu, L.; and Chua, T.-S. 2024. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. In *International Conference on Learning Representations*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. arXiv:2203.08101.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.

Deng, J.; Yang, Z.; Liu, D.; Chen, T.; Zhou, W.; Zhang, Y.; Li, H.; and Ouyang, W. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE transactions on pattern analysis and machine intelligence*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14105–14115.

Gosselin, F.; Kchir, S.; Acher, G.; Keith, F.; Lebec, O.; Louison, C.; Luvison, B.; de Chamisso, F. M.; Meden, B.; Morelli, M.; Perochon, B.; Rabarisoa, J.; Vienne, C.; and Ameyugo, G. 2022. Robot Companion, an intelligent interactive robot coworker for the Industry 5.0. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 8918–8925. IEEE.

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 *Proceedings of the ACM International Conference on Multimedia*, 4600–4609.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2023. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*.

Gu, G.; Chun, S.; Kim, W.; Kang, Y.; and Yun, S. 2024. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13225–13234.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31.

Gupta, T.; Vahdat, A.; Chechik, G.; Yang, X.; Kautz, J.; and Hoiem, D. 2020. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, 752–768. Springer.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, 1463–1471.

Han, X.; Zhu, X.; Yu, L.; Zhang, L.; Song, Y.-Z.; and Xiang, T. 2023a. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2669–2680.

Han, Y.; Zhang, L.; Chen, Q.; Chen, Z.; Li, Z.; Yang, J.; and Cao, Z. 2023b. Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15028–15038.

Hu, Y.; Liu, M.; Su, X.; Gao, Z.; and Nie, L. 2021a. Video moment localization via deep cross-modal hashing. *IEEE Transactions on Image Processing*, 30: 4667–4677.

Hu, Y.; Nie, L.; Liu, M.; Wang, K.; Wang, Y.; and Hua, X.-S. 2021b. Coarse-to-fine semantic alignment for cross-modal moment localization. *IEEE Transactions on Image Processing*, 30: 5933–5943.

Hu, Y.; Wang, K.; Liu, M.; Tang, H.; and Nie, L. 2023. Semantic collaborative learning for cross-modal moment localization. *ACM Transactions on Information Systems*, 42(2): 1–26.

Li, Z.; Liu, F.; Wei, Y.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2024. Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. In *Proceedings of the ACM International Conference on Multimedia*, 9660–9669. ACM.

Liu, F.; Liu, Y.; Chen, H.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2024a. Understanding Before Recommendation: Semantic Aspect-Aware Review Exploitation via Large Language Models. *ACM Transactions on Information Systems*, 1–26.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021a. 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.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021b. 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.

Liu, Z.; Sun, W.; Hong, Y.; Teney, D.; and Gould, S. 2024b. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5753–5762.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Song, C. H.; Hwang, T.; Yoon, J.; Choi, S.; and Gu, Y. H. 2024. SyncMask: Synchronized Attentional Masking for Fashion-centric Vision-Language Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13948–13957.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 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.

Wang, K.; Liu, H.; Jie, L.; Li, Z.; Hu, Y.; and Nie, L. 2024. Explicit Granularity and Implicit Scale Correspondence Learning for Point-Supervised Video Moment Localization. In *Proceedings of the ACM International Conference on Multimedia*, 9214–9223.

Wei, Y.; Wang, X.; Guan, W.; Nie, L.; Lin, Z.; and Chen, B. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing*, 29: 1–14.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1369–1378.

Wen, H.; Song, X.; Yin, J.; Wu, J.; Guan, W.; and Nie, L. 2024. Self-Training Boosted Multi-Factor Matching Network for Composed Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 3665–3678.

Wen, H.; Zhang, X.; Song, X.; Wei, Y.; and Nie, L. 2023. Target-guided composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 915–923.

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 *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11307–11317.

Xu, Y.; Bin, Y.; Wei, J.; Yang, Y.; Wang, G.; and Shen, H. T. 2023. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, 25: 8346–8357.

Yang, Q.; Ye, M.; Cai, Z.; Su, K.; and Du, B. 2023a. Composed image retrieval via cross relation network with hierarchical aggregation transformer. *IEEE Transactions on Image Processing*.

Yang, X.; Liu, D.; Zhang, H.; Luo, Y.; Wang, C.; and Zhang, J. 2024. Decomposing Semantic Shifts for Composed Image Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6576–6584.

Yang, Z.; Kafle, K.; Dernoncourt, F.; and Ordonez, V. 2023b. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19165–19174.

Zhang, G.; Wei, S.; Pang, H.; Qiu, S.; and Zhao, Y. 2022. Composed image retrieval via explicit erasure and replenishment with semantic alignment. *IEEE Transactions on Image Processing*, 31: 5976–5988.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Zhang, Z.; Wei, Z.; Huang, Z.; Niu, R.; and Wang, P. 2023. One for all: One-stage referring expression comprehension with dynamic reasoning. *Neurocomputing*, 518: 523–532.

Zhao, Y.; Song, Y.; and Jin, Q. 2022. Progressive learning for image retrieval with hybrid-modality queries. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1012–1021.

Zhu, H.; Wei, Y.; Zhao, Y.; Zhang, C.; and Huang, S. 2023. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6): 1–22.