ACF-R$^+$: An Asymmetry-sensitive Method for Image-Text Retrieval enhanced by Cross-Modal Fusion and Re-ranking based on Contrastive Learning

Ziyu Gong[a], Yihua Huang[a,*], Chunhua Yu[b], Peng Dai[b], Xing Ge[b], Yiming Shen[b] and Yafei Liu[b]

[a]State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing, 210000, China
[b]State Grid Jiangsu Electric Power Co.,LTD. Material Branch, Nanjing, 210000, China

ARTICLE INFO

Keywords:
Image-text Retrieval
Information Asymmetry
Cross-modal Fusion
Re-ranking Strategy
Contrastive Learning

ABSTRACT

The task of multi-modal retrieval between the image and text modality is to find pertinent information from a designated image or textual corpus. The principal challenge lies in the integration of multi-modal representations and the discernment of fine-grained distinctions among various modalities, with the goal of identifying analogous content and filtering out irrelevant information. However, the existing methods predominantly concentrated on the unification of semantic representations and the alignment of concepts across multiple modalities, while the subtle differences between modalities have not been extensively investigated, which resulted in the issue of information asymmetry. To address this problem, a novel asymmetry-sensitive method based on contrastive learning was proposed in this paper, for achieving unified multi-modal semantic representations while being able to distinguish fine-grained semantic differences by generating corresponding positive and negative samples tailored to different asymmetry types. Then, we introduced a hierarchical cross-modal fusion method to integrate semantic features at both the global and local levels by the multi-modal attention mechanism, for serving to enhance the alignment of conceptual representations across different modalities. On this basis, a re-ranking strategy was proposed by utilizing cross-modal bidirectional retrieval information. This algorithm exploits the K-nearest neighbors of the initial retrieval results to perform a reverse search and combines bidirectional retrieval information, for further improving the performance of image-text retrieval tasks. Extensive experiments conducted on the MSCOCO and Flickr30K datasets, verified the effectiveness of the proposed method.

1. Introduction

The objective of the image-text multi-modal retrieval task is to search for corresponding text or images that express the same or similar contents based on the input image or text. Due to the explosive increase of multimedia data, image-text retrieval is in great demand and plays a crucial role in many downstream research and application fields, such as multi-modal knowledge representation [1, 2], intelligent recommendation [3, 4, 5], and e-commerce product search [6, 7].

To retrieve highly correlated results from the given multimodal queries, existing works on image-text retrieval can be divided into two categories: global feature-based methods [8, 9, 10] and local feature-based methods [11, 12, 13, 14, 15]. The former type of method mainly projects images and sentences as a whole into a unified embedding space for better semantic representation, while the latter type of method focuses on achieving semantic alignment via exploiting cross-modal interactions between visual regions and textual words.

Previous work has mainly focused on learning the unified multi-modal semantic representations to achieve conceptual alignment between visual and textual modalities. Although these efforts played a significant role in promoting multi-modal retrieval tasks and achieved excellent results, image-text multi-modal retrieval still faces the following challenges: existing work does not fully consider the fine-grained differences between modalities, and does not consider the issue of information asymmetry across modalities.

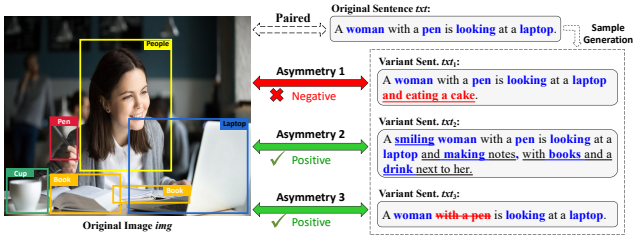


Figure. 1: Examples of different types of fine-grained information asymmetry between images and texts.

In the multi-modal retrieval task, the problem of information asymmetry refers to the situation that different modalities contain different amounts of information when expressing the same or similar concepts. That is, when describing the same scene, one modality may contain more or less information than another modality. This fine-grained difference between modalities would affect the performance of the final image-text multi-modal task. Generally speaking, the pixel-based image modality can provide a more objective description of visual objects, while the natural-language-based text modality often have subjective descriptions of

*Corresponding author
ziyugong@smail.nju.edu.cn (Z. Gong); yhuang@nju.edu.cn (Y. Huang)
ORCID(s): 0009-0008-0525-3089 (Z. Gong); 0000-0003-1806-0936 (Y. Huang)

scenes, and the focus and perspective of the description of the same scene are also different.

Therefore, the information contained in the text modality is often less than that in the image modality, which is a generalization or abstraction of the image information. Based on our statistical results on the image-text retrieval dataset, Flick30K [16], we found that in over 75% of image-text pairs, detectable objects in an image outnumber recognizable notional words in the paired text. This phenomenon may lead to the situation where the corresponding description text for an image may have some content that is the same as the image description, while some content that is different from the image description. This information asymmetry between modalities would further exacerbate the difficulty of the image-text multi-modal retrieval task.

An example is given in Figure 1 to illustrate the problem of information asymmetry. As shown in Figure 1, for a given image-text pair (*img*, *txt*), we divided the problem of information asymmetry into three categories: (1) **Asymmetry-I.** The text modality contains redundant information that is not present in the image. For the variant sentence txt_1, although txt_1 contains most of the objects in the image *img*, such as "people/women", "pen", "laptop", etc., it mistakenly contains information about "cake" that is not present in the image. Therefore, txt_1 should be considered as a negative sample for the image *img*. However, existing models may identify sentence txt_1 as a highly relevant positive sample for image *img* because txt_1 contains most of the information in *img*. This kind of fine-grained difference between modalities is a key challenge faced by multi-modal retrieval. (2) **Asymmetry-II.** The text modality contains richer information than the image modality, and these information are correct. For example, for the variant sentence txt_2, the words "smiling" and "making notes" that are included in txt_2 exceed the scope of objects detected in *img*, but these information are also correct. Therefore, txt_2 should also be considered as a long positive sentence sample. (3) **Asymmetry-III.** Compared to the original image-text pair, the description information in the text modality lacks descriptions for certain parts of the image, but still matches the semantics of the image. For example, for the variant sentence txt_3, although the word "pen" is deleted, the remaining text content still matches the description of *img*. Therefore, although txt_3 contains less information, it is still a short but positive sentence sample for image *img*. However, existing models may identify it as a negative sample or lower its ranking in the ranking list because the proportion of overlapping between txt_3 and *img* with the same concept is low. In summary, if the problem of fine-grained information asymmetry problem between multiple modalities is not solved, it would pose a huge challenge to the image-text multimedia retrieval task.

To solve this problem, we proposed an novel image-text multi-modal retrieval method based on asymmetric sensitivity and cross-modal fusion. Firstly, for each type of asymmetry, we generated corresponding positive and negative text samples to guide the model to learn fine-grained similarities and differences between multiple modalities, thereby learning more high-quality and discriminative multi-modal semantic representations. Then, we proposed a hierarchical cross-modal fusion method based on the multi-head attention mechanism to achieve concept alignment across multi-modalities from both the image-text and region-word levels, capturing the complex relations between the two heterogeneous modalities. This method was named as **ACF**, short for "***A***symmetry-sensitive and ***C***ross-modal ***F***usion".

On this basis, we also noted that, during the inference phrase, the previous retrieval methods separately performed the image-to-text (I2T) and text-to-image (T2I) retrieval tasks and ignored the bidirectional ranking information. This could pose an issue since in the training phase, both tasks are optimized using bidirectional loss functions, leading to disparities between training and inference stages. Therefore, we further proposed a re-ranking-based enhancement strategy for model ACF that comprehensively integrates the ranking information of I2T and T2I in the original image-text similarity matrix, for re-finding the most similar images or text and further boosting the retrieval performance. The re-ranking-enhanced method was named as **ACF-R$^+$**.

The major contributions of our work are summarized as follows:

- An asymmetry-sensitive contrastive learning method was proposed to solve the fine-grained information asymmetry problem between images and texts, where corresponding positives and negatives for each asymmetry type are generated to achieve unified semantic representation for better cross-modality retrieval based on semantic similarity.
- An image-text cross-modal fusion and semantic alignment method was also presented, which is predicated on a cross-modal attention mechanism to facilitate high-quality modality interaction with both local and global features.
- A post-processing re-ranking strategy was further presented to make the retrieval model obtain higher retrieval accuracy without extra training, which made full use of the information in the image-text similarity matrix and leveraged bidirectional retrieval information.
- Extensive experiments conducted on two widely used datasets, MSCOCO and Flickr30K, demonstrated that the proposed asymmetry-sensitive contrastive learning method has superior performance in comparison to the current state-of-the-art baselines.

2. Related work

2.1. Image-text retrieval

Existing research on image-text multi-modal retrieval primarily focused on the fields of unified semantic representation of multi-modal data [17, 18], and cross-modal

semantic interaction [19, 12]. Classified by the model's architecture, the existing models can be divided into two major categories: (1) The dual-encoder paradigm [8, 10, 20, 21]. In this paradigm, the input image and text modality data are encoded separately, and then the semantic representations of the two independent modalities are integrated into a shared semantic representation space, thereby achieving the fusion of the multi-modal semantics. (2) The joint-encoder paradigm [14, 22, 15, 23, 11]. In this paradigm, the input image and text modality data are uniformly encoded by a joint cross-modal encoder, thus realizing the fusion and interaction of heterogeneous multi-modal semantic representations.

In addition, for the image-text retrieval task, various objective functions for optimization have been studied to ensure higher similarity score between positive pairs, making it easier to retrieve relevant content and exclude irrelevant content. Eisenschtat and Wolf [24] improved similarity of relevant image-text pairs with the help of metric learning loss. Zheng et al. [25] introduced the instance loss to establish a connection between shared semantics. Most recent approaches employ a hinge-based triplet ranking loss with hard negatives proposed by Huang et al. [26] or a noise contrastive estimation bidirectional loss with hard negatives [27, 28] to bring positive pairs closer and negative pairs farther. Hence, selection strategies for informative samples have been extensively explored. In this line of work, early works [29] randomly chose irrelevant data or synthesize noise data from the original or external datasets as the negatives for model training and semantic representation optimization. In recent years, with the development of generation models [30, 31], an increasing number of studies have begun to utilize the generation language models for the construction and generation of hard negative samples and then these negatives were input to the contrastive-learning-based model, for achieving more discriminative multi-modal semantic representations and improving the quality of multi-modal semantic embeddings. UNITER+DG [32] proposed a denotation graph to generate hard negative text descriptions based on the graph structure relevance. AP-GRL [33] proposed a cross-modal graph with an adaptive pre-optimization and a double-order sampling strategy for unified representation learning. TAGS-DC [34] generated highly difficulty discriminative sentences as the negatives with the masking and refiling strategies for model training.

In summary, generating diverse positive and negative samples is crucial for enhancing the performance of image-text multi-modal retrieval. Therefore, building upon existing work, we addressed the asymmetry problem between modalities by generating different types of positives and negatives with the cross-modal fusion method, for achieving the fine-grained and high-discriminative unified semantic representations. And the method was further enhanced with the proposed re-ranking strategy, by utilizing the bi-directional image-to-text and text-to-image tasks.

2.2. Re-ranking strategy

Re-ranking strategy has been effectively utilized across various unimodal retrieval tasks, including person re-identification [35], object retrieval [36, 37], and text-based image search [38, 39, 40], aiming to enhance retrieval accuracy. The initial ranking list of retrieved candidates can be re-ordered as an additional refinement step across diverse domains. For instance, a k-reciprocal encoding approach for re-ranking re-ID results was introduced by Zhong et al. [41]. Specifically, the k-reciprocal feature is computed for a given image by encoding its k-reciprocal nearest neighbors into a unified vector, which is then utilized for re-ranking based on the Jaccard distance. Leng et al. [42] introduced a bidirectional ranking approach that incorporates contextual similarity between query images into the newly computed similarity.

Re-ranking strategy also plays a crucial role in multi-modal retrieval tasks. Cross-modal image-text retrieval re-ranking enhances model performance by utilizing information from both the source and reverse retrieval processes. For example, Yuan et al. [43] presented a plug-and-play MR re-ranking algorithm, which is aimed at making full use of the information in the similarity matrix and providing secondary optimization for retrieval results. This post-processing algorithm extensively considers the various information in the similarity matrix through a bidirectional retrieval technique.

3. Our methodology

The overview of our proposed method was illustrated in Figure 2. We first introduced the feature extraction method in Section 3.1. Then, Section 3.2 gives a formal definition of information asymmetry and corresponding sample generation methods. In Section 3.3, a hierarchical modality fusion module was proposed to achieve semantic alignment, through a cross-modal attention mechanism. In Section 3.4, we further presented a novel asymmetry-sensitive contrastive learning to optimize the retrieval performance. The re-ranking algorithm in Section 3.5 was utilized to further optimize retrieval performance of image-to-text (I2T) and text-to-image (T2I) retrieval.

3.1. Feature representations across multiple modalities

3.1.1. Visual representation

For the input images, as illustrated in Figure 2, we first used the Faster-RCNN model [44] pretrained on the Visual Genomes dataset [45] to extract K image regions as the raw visual features, denoted as $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_K] \in \mathbb{R}^{K*D_v}$, where D_v is the dimension of the extracted region features. Then, we transformed these feature vectors into the D-dimensional space via a fully-connected (FC) linear projection layer and normalized them. The output region representation of the image is denoted as $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_K] \in \mathbb{R}^{K*D}$. Meanwhile, we followed VSE++ [8] to acquire the global feature with the pretrained ResNet152 model [46] and then also projected the global feature into the D-dimensional space, *i.e.*, $\boldsymbol{G} \in \mathbb{R}^D$.

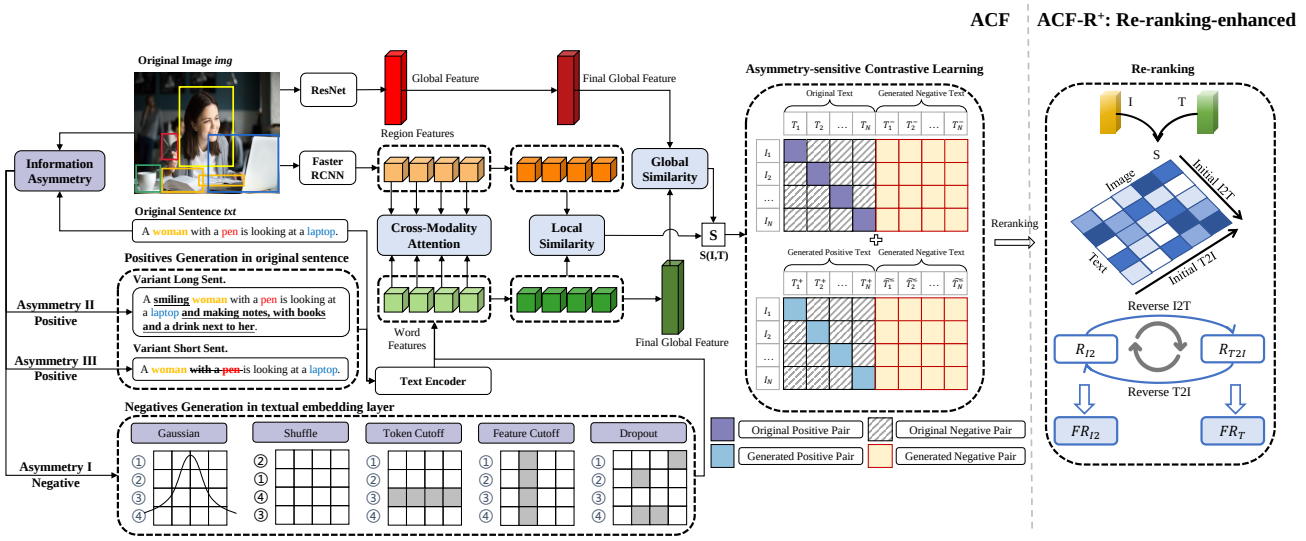


Figure. 2: The overview of our proposed ACF and ACF-R+ model. Based on fine-grained information asymmetry types between images and texts, corresponding positives and negatives are generated, which are provided for the subsequent asymmetry-sensitive contrastive learning with cross-modal fusion. A post-processing re-ranking algorithm during the interference phrase further helps improve retrieval accuracy without additional training.

3.1.2. Textual representation

For the input text, the pretrained BERT [47] model was used as the semantic encoder to obtain word embeddings, denoted as $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L] \in \mathbb{R}^{L*D}$, where L represents the amount of words and D reprsents the dimension of word embeddings.

3.2. Definition of information asymmetry and sample generation

For the same scene, images objectively capture visual information based on pixels, including colors, shapes, textures, and spatial relationships. Textual descriptions can be more subjective in conveying contextual information by words and phrases, but may not convey the same level of fine-grained details as images. Due to their inherent differences, their expressive emphases on describing the same scene differ, leading to information asymmetry.

Definition 1 (Information Asymmetry). Information asymmetry quantifies the differences between the sets of image and text information, indicating their level of dissimilarity. For a paired image-text pair (img, txt), assuming $Inf(img) = A$, $Inf(txt) = B$, information asymmetry (denoted as **IA**) between img and txt could be defined as:

$$\begin{aligned} IA(img, txt) &= A \triangle B = (A - B) \cup (B - A) \\ &= \{x | (x \in A \wedge x \notin B) \vee (x \in B \wedge x \notin A)\}, \end{aligned} \tag{1}$$

where $Inf(\cdot)$ represents the set of information contained in a certain modality and x denotes the constituent element. The set A and set B represent information capacity of an image and a sentence, *i.e.* detectable objects and recognizable words, respectively. Symbol "$\triangle$" is a set operation: symmetric set difference.

Given that $B \subsetneqq A$ for most positive image-text pairs, Equation 1 can usually be simplified to Equation 2.

$$\begin{aligned} IA_{smp}(img, txt) &= A \triangle B = A - B \\ &= \{x | x \in A \wedge x \notin B\} \end{aligned} \tag{2}$$

On this basis, information asymmetry can be further subdivided into three categories:

Definition 2 (Asymmetry-I). The variant text contains more redundant information that does not belong to the corresponding image. Set txt_1 as a variant of txt, and $Inf(txt_1) = Inf(txt) \cup X_1 = B \cup X_1, (X_1 \notin A)$. Then, given that $B \cup X_1 \notin A$, the Asymmetry-I can be formulized as,

$$\begin{aligned} &IA(img, txt_1) = A \triangle (B \cup X_1) \\ &= (A - (B \cup X_1)) \cup ((B \cup X_1) - A) \\ &= (A - B) \cup X_1 = IA_{smp}(img, txt) \cup X_1, \end{aligned} \tag{3}$$

where X_1 denotes the noise, which intensifies information asymmetry. Therefore, txt_1 is used as a negative sample to assist the subsequent model in semantic representation that is sensitive to fine-grained discrimination across multi-modalities. We used four methods to generate diverse negative samples by adding noise information to the textual embedding layers, listed as follows:

- Gaussian Noise: A Gaussian function, represented as $\{\hat{\boldsymbol{h}}_1, \hat{\boldsymbol{h}}_2, ..., \hat{\boldsymbol{h}}_L\} \sim \mathcal{N}(0, \sigma^2)$, was attached to the original textual embedding $\boldsymbol{W}$, where $\hat{\boldsymbol{h}}_i \in \mathbb{R}^D$, σ is the standard variance and L is the length of sentence.
- Token Shuffling: The original textual embedding $\boldsymbol{W}$ is in the form of a $L * D$ matrix, where L is the length

of sentence and D is the dimension of word features. Through random shuffling, the order of tokens in the $L * D$ matrix was altered.

- Token Cutoff & Feature Cutoff: We designated the value of zero to either a row (representing a token) or a column (representing a feature) of the matrix $\boldsymbol{W}$.
- Dropout: We selectively eliminated value from $\boldsymbol{W}$ through a probabilistic process, where the likelihood of each value being discarded was determined by a specific probability.

Definition 3 (Asymmetry-II). The variant text encompasses more pertinent information that is directly associated with the corresponding image. Set txt_2 as a variant of txt, and $Inf(txt_2) = Inf(txt) \cup X_2 = B \cup X_2, (X_2 \subset A - B)$. Then, given that $B \cup X_2 \subsetneqq A$, the Asymmetry-II can be formulized as,

$$\begin{aligned} IA(img, txt_2) &= A \bigtriangleup (B \cup X_2) \\ &= A - (B \cup X_2) = (A - B) - X_2 \\ &= IA_{smp}(img, txt) - X_2, \end{aligned} \tag{4}$$

where X_2 is a supplementary information, which has enriched semantics of the original text without involving any incorrect details. Specifically, for each image in the original dataset, there are five corresponding captions sharing similar concepts. To generate a long positive sentence sample, we randomly selected two sentences related to the image content and concatenated them.

Definition 4 (Asymmetry-III). The variant text deletes some of the information that is relevant to the corresponding image. Set txt_3 as a variant of txt, and $Inf(txt_3) = Inf(txt) - X_3 = B - X_3, (X_3 \subset B)$. Then, given that $B - X_3 \subsetneqq A$, the Asymmetry-III can be formulized as,

$$\begin{aligned} IA(img, txt_3) &= A \bigtriangleup (B - X_3) = A - (B - X_3) \\ &= (A - B) \cup X_3 = IA_{smp}(img, txt) \cup X_3. \end{aligned} \tag{5}$$

Despite deleting X_3, txt_3 still conforms to the semantics of the image. Therefore, we truncated the original sentence to produce a shorter yet still positive sentence example, for augmenting the diversity of the positives.

In particular, the generation of negatives occurs on the token embedding layer for achieving significant semantic changes, while the construction of positives is implemented on the original input layer to maintain semantic correctness.

3.3. Cross-modal fusion

To capture intricate correlations between images and texts, a hierarchical modality fusion method has been proposed through a cross-modal attention mechanism, which integrates both local-level and global-level multi-modality features.

3.3.1. Region-word fusion at local level

Considering the bidirectional retrieval requirements, *i.e.*, image-to-text and text-to-image, in the task of image-text retrieval, features $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M] \in \mathbb{R}^{M*D}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N] \in \mathbb{R}^{N*D}$ are defined to represent the query (modality of image or text) and the retrieved results (modality of text or image), respectively, where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $\boldsymbol{y}_i \in \mathbb{R}^D$ represents as the region-level features of visual modality or the word-level features of text modality. In our work, two single symmetrical versions of formula were proposed for bidirectional multi-modality retrieval. Concretely, we set $\boldsymbol{X} := \boldsymbol{V}(M := K)$ and $\boldsymbol{Y} := \boldsymbol{W}(N := L)$ for the task of image-to-text (I2T) retrieval, and $\boldsymbol{X} := \boldsymbol{W}(M := L)$ and $\boldsymbol{Y} := \boldsymbol{V}(N := K)$ for the task of text-to-image (T2I) retrieval. Symbol ":=" is an assignment operator. $\boldsymbol{V} \in \mathbb{R}^{K*D}$, $\boldsymbol{W} \in \mathbb{R}^{L*D}$ can be assigned to $\boldsymbol{X}$ or $\boldsymbol{Y}$ in different multi-modality retrieval subtasks, and $\boldsymbol{V}$ and $\boldsymbol{W}$ are defined in Section 3.1

Instead of simply aggregating the similarity of all possible pairs of regions and words, we used the multi-head cross-modal attention mechanism to allocate weights to visual regions and textual words based on their contribution to the shared semantics, and to capture fine-grained correspondence between two modalities from various representation subspaces, performed as follows,

$$\begin{aligned} \boldsymbol{Y}^* &= MultiHead(\boldsymbol{Query} := \boldsymbol{X}, \boldsymbol{Key} := \boldsymbol{Y}, \boldsymbol{Value} := \boldsymbol{Y}) \\ &= Concat(\boldsymbol{head}_1, ..., \boldsymbol{head}_i, ..., \boldsymbol{head}_H)\boldsymbol{Z}^O, \end{aligned} \tag{6}$$

where $\boldsymbol{Y}^* \in \mathbb{R}^{M*D}$, $\boldsymbol{Query} \in \mathbb{R}^{M*D}$, $\boldsymbol{Key} \in \mathbb{R}^{N*D}$, $\boldsymbol{Value} \in \mathbb{R}^{N*D}$, $\boldsymbol{Z}^O \in \mathbb{R}^{D*D}$, $Concat(\cdot)$ denotes the concatenation operation across multiple attention heads, and H represents the number of attention heads. In our work, $\boldsymbol{head}_i = Att(\boldsymbol{Query}_i := \boldsymbol{X}\boldsymbol{Z}_i{}^X, \boldsymbol{Key}_i := \boldsymbol{Y}\boldsymbol{Z}_i{}^Y, \boldsymbol{Value}_i := \boldsymbol{Y}\boldsymbol{Z}_i{}^Y)$, where Att represents the attention [48] with scaled dot-product, and $\boldsymbol{Z}_i^X, \boldsymbol{Z}_i^Y \in \mathbb{R}^{D*\frac{D}{H}}$.

For the image-to-text retrieval subtask, $\boldsymbol{Y}^*$ represents the text representations $\boldsymbol{W}^* = [\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, ..., \boldsymbol{w}_K^*] \in \mathbb{R}^{K*D}$ for each region with the multi-head attention, while for the text-to-image retrieval subtask, $\boldsymbol{Y}^*$ represents the image representations $\boldsymbol{V}^* = [\boldsymbol{v}_1^*, \boldsymbol{v}_2^*, ..., \boldsymbol{v}_L^*] \in \mathbb{R}^{L*D}$ for each word with the multi-head attention. Here, $\boldsymbol{w}_i^* \in \mathbb{R}^D$ represents the region-attended text representation for the i-th region, and $\boldsymbol{v}_i^* \in \mathbb{R}^D$ represents the word-attended image representation for the i-th word.

3.3.2. Image-text fusion at global level

In order to capture the complex correlations between different modalities, holistic images and sentences were jointly mapped into a common feature space, for ensuring global semantic consistency while also minimizing the heterogeneity between modalities. Concretely, for the text representations $\boldsymbol{W}^*$ obtained at the local-level fusion, we first computed its average value, denoted as $\overline{\boldsymbol{W}}^*$ in Equation 7, and then the averaged vector was projected into a shared vector space through a linear transformation. Finally, we obtained the global representation of text, denoted as $\boldsymbol{W}_g$, in Equation 8.

The formulas of $\overline{\boldsymbol{W}}^*$ and $\boldsymbol{W_g}$ are as follows:

$$\overline{\boldsymbol{W}}^* = \frac{\sum_{i=0}^{K} \boldsymbol{w}_i^*}{K} \tag{7}$$

$$\boldsymbol{W_g} = \boldsymbol{X_w}^T \overline{\boldsymbol{W}}^*, \tag{8}$$

where $\boldsymbol{X_w}$ is a trainable parameter matrix. In the same way, we projected $\overline{\boldsymbol{V}}^*$ (the average vector of $\boldsymbol{V}^*$) and $\boldsymbol{G}$ (the feature embedding of the whole image) into the shared common embedding space as. The corresponding formulas are given as follows,

$$\overline{\boldsymbol{V}}^* = \frac{\sum_{i=0}^{L} \boldsymbol{v}_i^*}{L} \tag{9}$$

$$\boldsymbol{V_{g_1}} = \boldsymbol{X_v}^T \overline{\boldsymbol{V}}^* \tag{10}$$

$$\boldsymbol{V_{g_2}} = \boldsymbol{X_g}^T \boldsymbol{G}, \tag{11}$$

where $\boldsymbol{X_v}, \boldsymbol{X_g}$ are trainable parameter matrices. The final global feature representation of the image is obtained, which is calculated as follows:

$$\boldsymbol{V_g} = Fusion(\boldsymbol{V_{g_1}}, \boldsymbol{V_{g_2}}) \tag{12}$$

3.4. Asymmetry-sensitive contrastive learning for I-T matching

3.4.1. Similarity score

The similarity between image (denoted by I) and text (denoted as T) is measured by the similarity score, which is defined by the summation of the local similarity score and global similarity score. The detailed calculation process is given as follows.

The local similarity score is predominantly represented by the average of the sum of word representations from the textual modality and region representations from the image modality, with the following formula for its computation:

$$S_{local}(I,T) = u_1 \cdot \frac{\sum_{i=1}^{K} R(\boldsymbol{v}_i, \boldsymbol{w}_i^*)}{K} + (1-u_1) \cdot \frac{\sum_{j=1}^{L} R(\boldsymbol{w}_j, \boldsymbol{v}_j^*)}{L}, \tag{13}$$

where $\boldsymbol{w}_i^*$ represents the word embedding in $\boldsymbol{W}^*$ attended to the image region, $\boldsymbol{v}_j^*$ represents the region embedding in $\boldsymbol{V}^*$ attended to the words, and $R(\boldsymbol{x}, \boldsymbol{y})$ represents the cosine similarity between different modalities, denoted as $R(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$. u_1 is a hyper-parameter and is set to 0.5. K represents the number of regions from an image and L represents the number of words in the text. It should be noted that during the retrieval process of image-to-text and text-to-image, formula $\frac{\sum_{i=1}^{K} R(\boldsymbol{v}_i, \boldsymbol{w}_i^*)}{K}$ and formula $\frac{\sum_{j=1}^{L} R(\boldsymbol{w}_j, \boldsymbol{v}_j^*)}{L}$ are calculated separately, and the results are then averaged by summation.

For the global similarity score involving the entire image (denoted by I) and the whole text (denoted by T), it is calculated as follows,

$$S_{global}(I,T) = R(\boldsymbol{V_g}, \boldsymbol{W_g}) \tag{14}$$

Finally, utilizing the above defined local and global similarity scores, the ultimate similarity score between image I and text T is calculated as follows:

$$S(I,T) = u_2 \cdot S_{local}(I,T) + (1-u_2) \cdot S_{global}(I,T), \tag{15}$$

where u_2 serves as a hyper-parameter.

3.4.2. Image-text matching based on asymmetry-sensitive contrastive learning

In this paper, a novel method based on sensitivity to information asymmetry between modalities was proposed. By generating positive and negative samples for each type of asymmetric information defined in Section 3.2, it enhances the perception of information asymmetry and provides more discriminative semantic representations for multi-modal retrieval tasks. The details of the method are given as follows.

Image-Text matching for Asymmetry-I. In the training phase of the model, a batch comprises N image-text pairs from the training set. In accordance with the concept of Asymmetry-I and the negative sample generation method defined in Section 3.2, N noise-added negative sentences are generated. Consequently, in the case of Asymmetry-I, for every positive image-text (I,T) pair, we retrieved $N-1$ in-batch negative image examples (denoted as $\{\widehat{I}_n\}_{n=1}^{N-1}$), $N-1$ in-batch negative sentence examples (denoted as $\{\widehat{T}_n\}_{n=1}^{N-1}$), and N generated negative sentence examples (denoted as $\{T_n^-\}_{n=1}^{N}$). These negative examples contain noise data unrelated to image I and can serve as negative samples for training and optimizing the multi-modal semantic representations. Based on the generation of negative examples, the loss function for image-text matching in the case of Asymmetry-I is expressed as follows:

$$L_{ACF_I}(I,T) = \frac{e^{S(I,T)/\tau}}{e^{S(I,T)/\tau} + \sum_{n=1}^{N-1} e^{S(\widehat{I}_n,T)/\tau}} + \frac{e^{S(I,T)/\tau}}{e^{S(I,T)/\tau} + \sum_{n=1}^{N-1} e^{S(I,\widehat{T}_n)/\tau} + \sum_{n=1}^{N} \alpha_n \cdot e^{S(I,T_n^-)/\tau}}, \tag{16}$$

where α_n is the hyper-parameter. If the similarity score of the image-text pair surpasses the positive pair, α_n is set to 0. Otherwise, α_n is set to 1.

Image-Text matching for Asymmetry-II and Asymmetry-III. Similarly, in the case of Asymmetry-II and Asymmetry-III, based on the data generation methods defined in Asymmetry-II and III, we generated more detailed or concise sentences for each positive image-text pair (I,T) as the generated positive sentence examples, and denote the generated positive sentence in the text modality as T^+. In this case, for (I,T), (I,T^+) is the newly generated positive sample pair, while for $N-1$ image-text pairs in the same batch except for (I,T), (I,T^+) can be used as the newly generated negative image-text pair. Similar to Asymmetry-I, we can also add noisy information to the newly generated positive text sentence and convert it into negatives. Therefore, for each positive image-text pair (I,T^+), during the model training process, there also exists $N-1$ in-batch negative images (denoted as $\{\widehat{I}_n\}_{n=1}^{N-1}$),

$N-1$ in-batch negative sentences (denoted as $\{\widehat{T_n^+}\}_{n=1}^{N-1}$), and N generated negative sentences (denoted as $\{\widetilde{T_n^-}\}_{n=1}^{N}$). The role of T^+ is similar to that of T in Equation 16. Therefore, for Asymmetry-II and III, the objective function can be defined as:

$$L_{ACF_{II\&III}}(I,T^+) = \frac{e^{S(I,T^+)/\tau}}{e^{S(I,T^+)/\tau} + \sum_{n=1}^{N-1} e^{S(\widehat{I}_n,T^+)/\tau}} + \frac{e^{S(I,T^+)/\tau}}{e^{S(I,T^+)/\tau} + \sum_{n=1}^{N-1} e^{S(I,\widehat{T_n^+})/\tau} + \sum_{n=1}^{N} \alpha_n \cdot e^{S(I,\widetilde{T_n^-})/\tau}} \tag{17}$$

Joint Training Objectives. By generating corresponding positive and negative pairs for different types of asymmetry, the joint training objective of our proposed multi-modal retrieval method, ACF, based on sensitivity to inter-modality asymmetry is formalized as follows:

$$L_{ACF}(I,T) = \lambda L_{ACF_I}(I,T) + (1-\lambda) L_{ACF_{II\&III}}(I,T^+), \tag{18}$$

where λ serves as the hyper-parameter. The above methods and strategies constitute model ACF.

3.5. Re-ranking strategy

Most existing methods separately conducted image-to-text (I2T) and text-to-image (T2I) tasks in the retrieval models by unidirectionally exploiting the learned image-text similarity matrix to search for retrieval candidates based on either an input image or text query. These previous retrieval schemes overlooked the valuable bidirectional retrieval information that exists between images and text, leading to suboptimal performance in cross-modal retrieval tasks. To address this limitation, we further focused on developing a re-ranking strategy that leverages bidirectional retrieval information without extra training and refines the retrieval results obtained from the initial query, to enhance the performance of image-to-text (I2T) and text-to-image (T2I) retrieval tasks simultaneously.

3.5.1. I2T re-ranking

As shown in Figure 2, we first conducted the initial image-to-text (I2T) retrieval based on a given query image I and defined the initial ranking list of the top-$\hat{K}$ candidate text as $\overrightarrow{R_{I2T}}(I,\hat{K})$.

$$\overrightarrow{R_{I2T}}(I,\hat{K}) = \{T_1,...,T_j,...,T_{\hat{K}}\}, \tag{19}$$

where T_j is the text ranked at position j in the initial retrieval results. Then for each candidate text T_j, we performed reverse T2I retrieval and obtained $\hat{N}$-nearest neighbour images of T_j, which could be defined as follows,

$$\overleftarrow{R_{T2I}}(T_j,\hat{N}) = \{I_1,...,I_z,...,I_{\hat{N}}\}, \tag{20}$$

where I_z represents the image ranked at position z in the reverse retrieval results.

For paired images and text, they can be retrieved from each other by I2T or T2I retrieval forwardly and reversely. To comprehensively utilize two-way retrieval results, we further determined the index position of query image I in the reverse search results for each candidate text T_j.

$$p(T_j) = z, \quad if \;\; I_z = I, I_z \in \overleftarrow{R_{T2I}}(T_j,\hat{N}) \tag{21}$$

Next, we established a position set P of image I, encompassing index positions for all candidate texts within the initial $\hat{K}$-nearest neighbors.

$$P(I,\hat{K}) = \{p(T_1),...,p(T_j),...,p(T_{\hat{K}})\} \tag{22}$$

For two initial retrieval results of image I, denoted as $\{T_p,T_q\}(p<q)$, if the index position of query image I in the reverse search results for candidate text T_q is higher than that for T_p, there is a higher probability that T_q is the matching text for image I and the final ranking list should be refined as $\{T_q,T_p\}$. Therefore, we reordered the position set $P(I,\hat{K})$ to further optimize the sorting of candidate texts in $\overrightarrow{R_{I2T}}(I,\hat{K})$. $FR_{I2T}(I,\hat{K})$ refers to the improved list of retrieved text for the query image I.

$$FR_{I2T}(I,\hat{K}) = Ranking(P(I,\hat{K})) \tag{23}$$

3.5.2. T2I re-ranking

Similarly, we first performed the text-to-image (T2I) retrieval using a provided query text T and established the initial ranking list of the top-$\hat{K}$ candidate images as $\overrightarrow{R_{T2I}}(T,\hat{K})$.

$$\overrightarrow{R_{T2I}}(T,\hat{K}) = \{I_1,...,I_j,...,I_{\hat{K}}\}, \tag{24}$$

where I_j denotes the image positioned at rank j in the initial retrieval outcomes.

Subsequently, for every candidate image I_j, we conducted reverse image-to-text (I2T) retrieval, acquiring the $\hat{N}$-nearest neighbor text of I_j, as outlined below:

$$\overleftarrow{R_{I2T}}(I_j,\hat{N}) = \{T_1,...,T_z,...,T_{\hat{N}}\} \tag{25}$$

Slightly different from I2T re-ranking, the retrieval results for text T are inherently linked to the retrieval results of other semantically related texts, as each image typically has multiple associated texts. Therefore, we found K'-nearest neighbour text set of the query text T, which is defined as follows:

$$G(T,K') = \{T_1,...,T_{K'}\} \tag{26}$$

Following a process similar to the re-ranking procedure in I2T re-ranking, the refined results are derived by conducting I2T retrieval for each image in $\overrightarrow{R_{T2I}}(T,\hat{K})$. The specific steps involved are outlined as follows:

$$p(I_j) = z \quad if \;\; T \in G(T,K'), T_z \in \overleftarrow{R_{I2T}}(I_j,\hat{N}) \tag{27}$$

$$P(T, \hat{K}) = \{p(I_1), ..., p(I_j), ..., p(I_{\hat{K}})\} \quad (28)$$

$$FR_{T2I}(T, \hat{K}) = Ranking(P(T, \hat{K})), \quad (29)$$

where $FR_{T2I}(T, \hat{K})$ is an updated list of retrieval images for querying text T. Finally, the re-ranking-enhanced method was named as ACF-R$^+$.

4. Experiments and analyses

4.1. Dataset

In this section, to evaluate the performance of our method, extensive experiments were conducted on two widely-used public datasets: MSCOCO [49] and Flickr30K [16]. The detailed description of the datasets are as follows.

- MSCOCO: This dataset contains 123,287 images, where each image has 5 corresponding description sentences with human annotations. Following previous work [8, 50], we split 113,287 images for training, 5,000 images for validation and 5,000 images for testing. We used MSCOCO (5K) for the evaluation, *i.e.*, directly testing on the full 5K images.

- Flickr30K: This dataset contains 31,783 images, each accompanied by five sentences describing its content. We split 29,783 images and their corresponding sentences into the training set, 1,000 images into the validation set, and 1,000 images into the test set.

4.2. Experimental settings

4.2.1. Evaluation protocols

We used R@K (K=1,5,10) as the indicators to evaluate the performance of our method, which are standard metrics in the retrieval task and measure the percentage of the ground truth results being retrieved included in the top-K results. The higher R@K means the better performance for the retrieval task.

4.2.2. Implementation details

Our proposed model was trained and evaluated on one NVIDIA A100 GPU with the PyTorch library. We used Adam [51] as the optimizer for the model and set its batch size to 64 with 20 epochs for training. The dimension of the shared representations of the image and text are both set to 1024, denoted as D, and each image is divided into 36 regions, denoted as K=36, as the local features of the image. The dimension of the representations for each region of the image is set to 2048, denoted as D_v. For the dataset MSCOCO, we set the learning rate of the optimizer to 5e-4, declined by 10% every 10 epochs. The hyper-parameters τ and u_2 were set to 0.05 and 0.8, respectively. For the dataset Flickr30K, we set the learning rate of the optimizer to 2e-4, declined by 10% every 10 epochs. The hyper-parameters τ and u_2 were set to 0.01 and 0.6, respectively. Both the global hyper-parameters u_1 and λ were set to 0.5. We conducted the experiments for 5 times and reported the average values.

4.3. Results of the image-text retrieval task and analyses

According to the results listed in Table 1, the performance of our proposed method surpassed the existing state-of-the-art baselines, such as VSE++ [8], SCAN [50], DIME [52], DSRAN [10], TAGS-DC [34], UNITER+DG [32], SOHO [55], ViSTA [56], COTS [57], LightningDoT [58], LexLIP [18], on both two public datasets for the multi-modality retrieval task, which verified the effectiveness of our method. Here, "I2T" refers to the process of retrieving text descriptions with a given image, while "T2I" refers to the retrieval of images with the given text description.

For dataset MSCOCO (5K), compared to the baseline method LightningDoT [58], our method outperformed LightningDoT on subtask I2T by 30.2%, 11.8%, and 6.3% on the three metrics of R@1, R@5, and R@10, respectively. On another subtask T2I, our method also achieved a consistent significant improvement compared to LightningDoT. Compared to the SOTA baseline, LexLIP [18], our method also obtained a large improvement of 24.6%, 8.7%, 4.6% in terms of R@1, R@5, R@10, respectively, for the I2T subtask, and achieved the improvements of 13.5%, 8.2%, 5.6% in terms of R@1, R@5, R@10, respectively, for the T2I subtask. Similarly, for dataset Flickr30K (1K), on the subtaks of I2T and T2T, our method can surpass existing strong baseline models such as LexLIP [18], LightningDoT [58], and COTS [57] on all three metrics of R@1, R@5, and R@10. Especially on the R@1 metric, our method can significantly surpass LexLIP by 7.7%, LightningDoT by 12.6%, and COTS by 8.5% for subtask I2T. For subtask T2I, our method also achieved consistent performance improvement. In summary, by utilizing generated positives and negatives according to our defined fine-grained information asymmetry types, the proposed ACF method showed the superior performance for the image-text retrieval task and the experimental results also validated the effectiveness of our proposed method.

4.4. Ablation experiments and results

Effect of different generated data samples. To investigate the impact of generated positives and negatives on the image-text retrieval task, we generated three different types of sample data and designed three variant models based on them, according to the different types of information asymmetry patterns. These three different types of variant models are respectively referred to as: (1) "-*w/o* Positives": This variant model only generates negative sentences for model training based on Asymmetry-I. (2) "-*w/o* Negatives": This variant model only generates positive sentences for model training based on Asymmetry-II and III. (3) "-*w/o* Positives & Negatives": This variant model does not consider the problem of information asymmetry between modalities, therefore, this method does not generate any positives or negatives for additional model training and optimization.

According to the experimental results from line 1 to 4 in Table 2, we found that after removing the generated positives (annotated as"-*w/o* Positives"), the generated negatives (denoted as "-*w/o* Negatives"), and all generated data (denoted

Table 1
Main experimental results of image-text retrieval on the datasets of MSCOCO (5K test set) and Flickr30K (1K test set).

Models	MSCOCO Test (5K)						Flickr30K Test (1K)					
	I2T Retrieval			T2I Retrieval			I2T Retrieval			T2I Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [50]	50.4	82.2	90.0	38.6	69.3	80.4	67.4	90.3	95.8	48.6	77.7	85.2
DIME [52]	59.3	85.4	91.9	43.1	73.0	83.1	81.0	95.9	98.4	63.6	88.1	93.0
VSE++ [8]	41.3	71.1	81.2	30.3	59.4	72.4	52.9	80.5	87.2	39.6	70.1	79.5
ViLT [53]	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
CLIP [54]	62.2	86.5	93.2	47.2	75.0	83.9	87.3	97.6	98.7	70.6	90.4	94.4
COOKIE [30]	61.7	86.7	92.3	46.6	75.2	84.1	84.7	96.9	98.3	68.3	91.1	95.2
DSRAN [10]	57.9	85.3	92.0	41.7	72.7	82.8	80.5	95.5	97.9	59.2	86.0	91.9
TAGS-DC [34]	67.8	89.6	94.2	53.3	80.0	88.0	90.6	98.8	99.1	77.3	94.3	97.3
SOHO [55]	66.4	88.2	93.8	50.6	78.0	86.7	86.5	98.1	99.3	72.5	92.7	96.1
Unicoder-VL [31]	62.3	87.1	92.8	46.7	76.0	85.3	86.2	86.3	99.0	71.5	90.9	94.9
UNITER+DG [11]	51.4	78.7	87.0	39.1	68.0	78.3	78.2	93.0	95.9	66.4	88.2	92.2
ViSTA [56]	68.9	90.1	95.4	52.6	79.6	87.6	89.5	98.4	99.6	75.8	94.2	96.9
COTS [57]	69.0	90.4	94.9	52.4	79.0	86.9	90.6	98.7	99.7	76.5	93.9	96.6
LightningDoT [58]	64.6	87.6	93.5	50.3	78.7	87.5	86.5	97.5	98.9	72.6	93.1	96.1
LexLIP [18]	70.2	90.7	95.2	53.2	79.1	86.7	91.4	99.2	99.7	78.4	94.6	97.1
ACF (ours)	**94.8**	**99.4**	**99.8**	**66.7**	**87.3**	**92.3**	**99.1**	**99.8**	**100.0**	**83.0**	**95.4**	**97.6**

Table 2
Ablation Experiments on dataset MSCOCO and Flickr30K.

Model Variants	MSCOCO (5K)				Flickr30K (1K)			
	I2T Retrieval		T2I Retrieval		I2T Retrieval		T2I Retrieval	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
ACF (Complete Model)	**94.8**	99.4	**66.7**	**87.3**	**99.1**	**99.8**	**83.0**	**95.4**
-*w/o* Positives	93.1	99.1	65.1	86.0	99.0	99.7	79.6	92.7
-*w/o* Negatives	94.8	99.3	61.2	83.3	98.8	99.6	77.2	91.9
-*w/o* Positives & Negatives	93.1	98.8	59.1	81.6	98.8	99.4	76.4	91.5
-*w/o* CMF	39.7	71.2	31.0	58.7	77.5	95.9	47.0	71.3
-*w* triplet loss	89.4	**99.8**	58.6	79.8	98.3	99.2	73.3	87.9

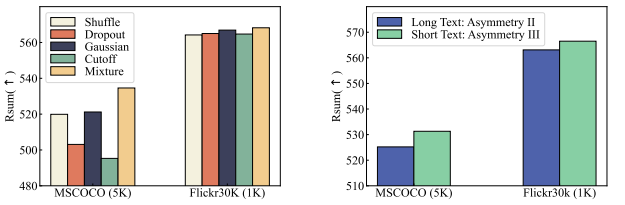
as "-*w/o* Positives & Negatives") from the complete model, the performance of these three variant models decreased on both datasets. For example, for the dataset MSCOCO, after removing Positives, Negatives, and Positives & Negatives, the model performance on the R@1 metric for subtask T2I decreased from 66.7% to 65.1%, 61.2%, and 59.1%, respectively. Similarly, for the dataset Flickr30K, after removing the three types of generated sample data, the performance of its corresponding variant models decreased from 83.0% to 79.6%, 77.2%, and 76.4% in the metric of R@1 for the subtask T2I, compared to the complete model, *i.e.*, ACF. This ablation experiment indicates that generating positive and negative sample data according to the proposed asymmetric types is beneficial for improving multi-modal retrieval performance.

Furthermore, we conducted a more in-depth analysis of the impact of sample data generation methods corresponding to different asymmetric types on the performance of the multi-modal retrieval task. As defined in Section 3.2, we proposed five negative sample generation methods for Asymmetry-I, including "Dropout", "Shuffle", "Cutoff", "Gaussian", and "Mixture". Among them, "Mixture" refers to the mixed use of the first four data generation methods. For Asymmetry-II, we proposed a long positive sentence generation method to provide more semantic content for the text description. For Asymmetriy-III, we proposed a short positive sentence sample generation method to remove some semantic information from the text.

In order to evaluate the performance of the model under different data generation strategies, we used a comprehensive evaluation metric Rsum, which means that Rsum is the sum of R@1, R@5, and R@10 in I2T and T2I subtasks. The higher the Rsum, the better the overall performance of the model.

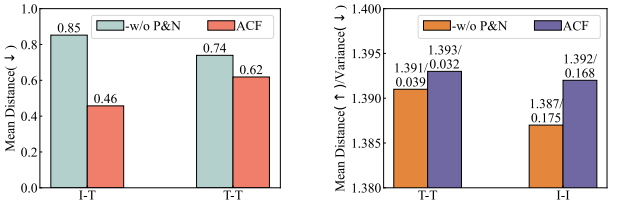
According to the experimental results in Figure 3, it can be observed that: (1) For Asymmetry-I, the negative sample generation method based on "Mixture" achieved the highest Rsum value, exceeding the Rsum values compared to the "Shuffle", "Dropout", "Gaussian", and "Cutoff" strategies. The reason for this is that the hybrid negative sample generation method, denoted as "Mixture", can provide the model with more diverse negative samples, thereby helping the model to distinguish the differences between different modalities at a fine-grained level and learn more robust multi-modal semantic representations. (2) For Asymmetry-II and Asymmetry-III, compared to generating long positive sentences, short positive sentences can more effectively help the model to improve the performance of multi-modal retrieval. The possible reason can be that the short sentence samples can reduce the overlap and redundancy of the same semantics, allowing the model to pay more attention to the differences between images and texts. At the same time, a training set composed of multiple short positive sentences can enrich the semantic information of texts from multiple perspectives, thereby improving the performance of the image-text retrieval task.

Effect of cross-modal fusion. As shown in Table 2, after removing the cross-modal fusion algorithm from the ACF



(a) Negatives: Asymmetry-I (b) Positives: Asymmetry-II and III

Figure. 3: Comparison of multi-modal retrieval performance using diverse data sample generation methods corresponding to different asymmetric types. (a) The impact of different negative sample generation strategies on model performance in the case of Asymmetry-I. (b) The impact of different positive sample generation methods on model performance in the case of Asymmetry-II and Asymmetry-III. The arrow symbol (↑) indicates that the higher the indicator, the better the performance of the model.



(a) Alignment: Positives get close. (b) Uniformity: Negatives get apart.

Figure. 4: Statistics of the mean distance between positive pairs (a) and negative pairs (b) on dataset MSCOCO. The upper arrow (↑) means the higher the value, the better the performance. The down arrow (↓) means the lower the value, the better the performance. -w/o P&N is short for $\text{ACF}_{\text{-w/o Positives \& Negatives}}$

.

method (denoted by "*-w/o* CMF"), we found a significant decrease in the performance of the method on all metrics in both datasets. For example, after removing CMF, for the MSCOCO dataset, the method performance on the R@5 metric in subtask I2T decreased from 99.4% to 71.2%. Similarly, for the Flickr30K dataset, the method performance on the R@5 metric in subtask T2I decreased from 95.4% to 71.3%. These experimental results all indicated that the cross-modal fusion method based on the multi-head attention mechanism plays a crucial role in the multi-modal retrieval task, for helping to achieve concepts alignment and fusion between different modalities.

Effect of the loss function. In order to verify the effectiveness and superiority of the loss function proposed in Equations 16 and 17, we replaced our proposed loss function with the existing triplet loss, denoted as "*-w* triplet loss", in the ablation experiment. Based on the experimental results in Table 2, we found that the performance of the method replaced with triplet loss decreased on a large portion of the metrics in both datasets. For example, on the Flickr30K dataset, compared to our method, the method with triplet loss showed a decrease of the performance, from 83.0% to 73.3%, in the R@1 metric for the T2I task. This ablation experiment once again validated the effectiveness of our proposed multi-modal retrieval method based on asymmetric sensitivity and contrastive learning.

4.5. Experiments on alignment and uniformity

For the image-text multi-modal retrieval task based on contrastive learning, alignment and uniformity are two key metrics for assessing the semantic representations of multi-modal data. Therefore, the following experiments evaluated these metrics separately.

4.5.1. Evaluation for alignment

For the alignment indicator, it tends to pull in the distance between positive sample pairs. In the image-text retrieval task, there are two main types of positive pairs: (1) inter-modality image-text positive pairs. (2) intra-modality text-text positive pairs.

In this work, we used Euclidean distance to measure the distance between positive pairs for inter-modality and intra-modality, respectively, which can indicate the performance of alignment. We conducted this experiment on the MSCOCO dataset, and the experimental results are shown in Figure 4(a). We found that the complete model, ACF, can achieve better alignment performance than the variant model, $\text{ACF}_{\text{-w/o Positives \& Negatives}}$. Specifically, for both positive image-text pairs (I-T) and positive text-text pairs (I-I), the distance between positive pairs obtained by the ACF model is closer than that obtained by the variant model $\text{ACF}_{\text{-w/o Positives \& Negatives}}$. This experiment shows that generating positive and negative pairs for different asymmetry types is helpful in improving the quality of the multi-modal semantic representations and achieving better cross-modality semantic alignment.

We also provided two visualization examples from MSCOCO in low-dimensional space by T-SNE [59] algorithm to prove better alignment and the results are illustrated in Figure 5. We found that distances between positive image-text (I-T) pairs and positive text-text (T-T) pairs learned by our ACF method (denoted with red color) are closer than those learned from model $\text{ACF}_{\text{-w/o Positives \& Negatives}}$ (denoted with blue color). It is consistent with our conclusions in Figure 4(a), which once again proves that our method can learn robust semantics for better alignment performance.

4.5.2. Evaluation for uniformity

In addition, we also evaluated the performance of another key attribute of multi-modal semantic representation quality based on contrastive learning, namely uniformity. This indicator tends to evenly distribute the learned feature representation vectors on a hypersphere. For this purpose, we calculated the average Euclidean distance between all images and all texts with different semantics to measure the uniformity, respectively. The experimental results are shown in Figure 4(b).

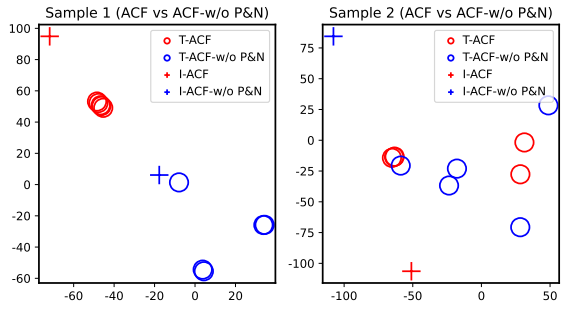


Figure. 5: The visualization of alignment: positive pairs of two samples in low-dimensional space. "+": image representations, "o": text representations. Red color and blue color indicate representations learned from model ACF and model ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$, respectively. P&N is short for Positives & Negatives

.

According to the experimental results, it can be seen that: (1) For all texts, the average semantic distance between the semantic representations learned by the complete model, ACF, is larger than that of the variant model, ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$, (1.393 of ACF v.s. 1.391 of ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$), and the variance of the average semantic distance learned by ACF is smaller, (0.032 of ACF v.s. 0.039 of ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$). (2) For all images, the complete model ACF, also learns images with larger average distances and smaller average variances between semantic representations, compared to the variant method, ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$. The experimental results indicated that by generating corresponding positive and negative pairs for different asymmetric types, our proposed ACF method can distribute the learned multi-modal semantic representation embeddings more evenly in the high-dimensional hyper-sphere, demonstrating the better uniformity performance.

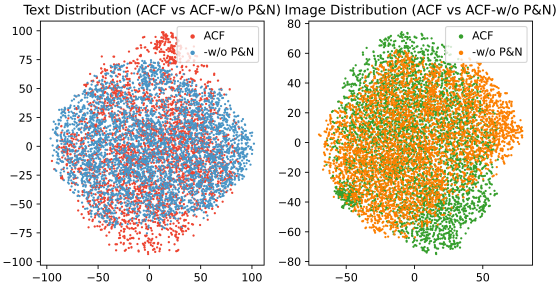
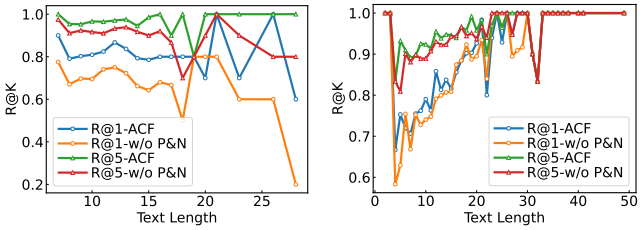


Figure. 6: The visualization of uniformity: the distribution of visual and textual representations in low-dimensional space under model ACF and ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$. Each point denotes a text or an image. Different colors indicate representations learned from different models. P&N is short for Positives & Negatives.

We also sampled 5,000 sentences and 5,000 images from MSCOCO, and visualized their distribution by projecting high-dimensional feature embeddings into low-dimensional space by T-SNE [59] algorithm. As can be seen from Figure 6, textual and visual representations learned from our model ACF are more widely dispersed in the two-dimensional plain than those learned from model ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$, which is consistent with our conclusions in Figire 4(b).



(a) Retrieval in dataset MSCOCO. (b) Retrieval in dataset Flickr30K.

Figure. 7: The influences of the length of the text queries on the text-to-image retrieval task.

4.6. Influences of the length of the text queries

To analyze and evaluate the impact of the length of the query text on the performance of the text-to-image retrieval task (T2I), we conducted in-depth experiments on the MSCOCO and Flickr30K datasets. For each dataset, we sampled 1000 sentences and measured the retrieval performance corresponding to query texts of different lengths. The experimental results are presented in Figure 7.

Based on the experimental results, we observed that: (1) Regardless of the length of the query text, the performance of the complete model ACF is consistently superior to that of the variant model, ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$, in terms of the retrieval performance metrics R@1 and R@5, reaffirming the effectiveness and superiority of our proposed asymmetric sensitivity method. (2) When the length of the query text is less than 10 words, our proposed ACF method still achieved outstanding retrieval performance, demonstrating that our method is equally sensitive and effective for short text in the multi-modal retrieval task. One possible explanation is that, according to Asymmetry-III, our model generates short sentences as positive samples during the model training phase, enhancing the ability of the model to represent short text sentences.

4.7. Effect of the post-processing re-ranking

After completing the training of the image-text retrieval model and obtaining the initialized retrieval ranking list, we further performed a re-ranking-based enhancement strategy on the retrieval results. Our proposed re-ranking enhancement strategy is executed during the inference stage and is orthogonal to the training process of the multi-modal retrieval model. It aims to further improve the ranking of relevant images or texts based on the preliminary ranking results of the retrieval model, thereby enhancing the performance of image-text retrieval.

According to the retrieval results shown in Table 3, it can be observed that: (1) When conducting the re-ranking

Table 3
The impact of post-processing re-ranking on MSCOCO and Flick30K. ACF-R$^+$ represents the re-ranking-enhanced method.

Model Variants (ours)	MSCOCO Test (5K)						Flickr30K Test (1K)					
	I2T Retrieval			T2I Retrieval			I2T Retrieval			T2I Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ACF	94.8	99.4	**99.8**	66.7	87.3	**92.3**	99.1	99.8	**100.0**	83.0	95.4	**97.6**
ACF-R$^+$ $_{\text{-}w \text{ Re-ranking}}$	**96.7**	**99.6**	**99.8**	**90.8**	**92.3**	**92.3**	**99.3**	**100.0**	**100.0**	**93.2**	**95.7**	**97.6**

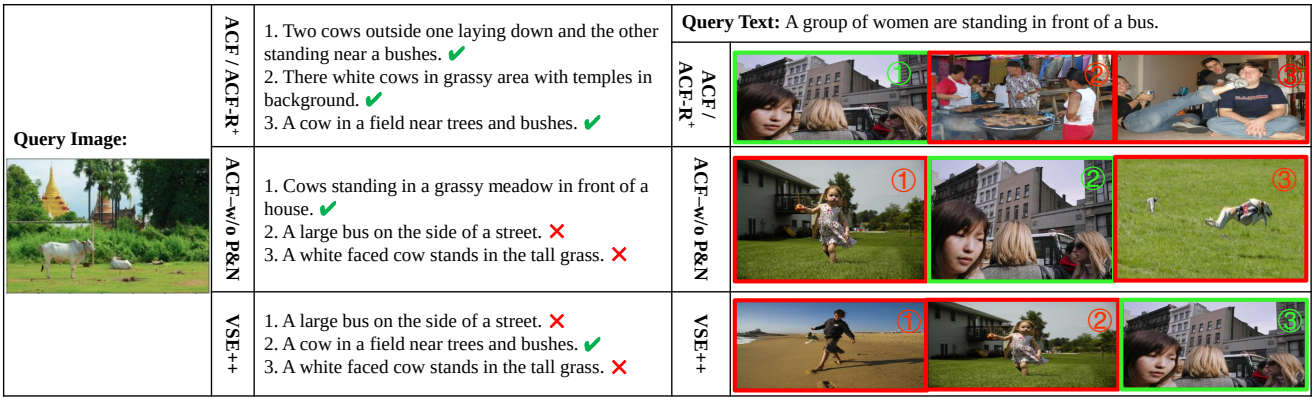


Figure. 8: Top-3 I2T and T2I retrieval results on Flickr30K. Green indicates the image-text pairs that are matched, while read indicates the image-text pairs that are not matched.

enhancement strategy, the retrieval performance of the enhanced method, denoted as ACF-R$^+$, experienced an improvement compared to the original version, *i.e.*, ACF, on the R@1, R@5, and R@10 metrics in both two datasets. For example, for the MSCOCO dataset, after executing the re-ranking strategy, the R@1 performance of ACF improved from 94.8% to 96.7% for the image-to-text (I2T) subtask. Similarly, on the Flickr30K data set, after performing the re-ranking strategy, the R@1 performance of ACF improved significantly from 83.0% to 93.18% for the subtask text-to-image (T2I). (2) After conducting the re-ranking enhancement strategy, the performance of ACF-R$^+$ significantly improved on the text-to-image(T2I) subtask, especially in terms of R@1 and R@5. The main reason is that the original ACF model has already obtained relatively high retrieval accuracy in the image-to-text(I2T) subtask, thereby when performing text-to-image(T2I) retrieval in the enhanced model ACF-R+, reverse high-precision I2T retrieval helps to change the relative position of the initial T2I retrieval results. In summary, this re-ranking experiment verified that our proposed re-ranking strategy is helpful for further improving the performance of the image-text retrieval task and also validated the effectiveness of the proposed re-ranking strategy.

4.8. Case study

To qualitatively validate the effectiveness of the proposed method ACF and ACF-R$^+$, we visualized several image-text retrieval samples in Figure 8. The left column of Figure 8 is the image-to-text samples. Given an image, our model ACF and the enhanced model ACF-R$^+$ can both retrieve three corresponding correct sentences, while variant model ACF$_{\text{-}w/o \text{ Positives \& Negatives}}$ and VSE++ only retrieved one relevant correct sentence. Similarly, as shown in the right column of Figure 8, for a given text, both ACF and ACF-R$^+$ can achieved the best search result for the text-to-image retrieval, with the most relevant image ranked first. Although the rest of results do not exactly match the sentence, they are still semantically similar to the image. For example, the second and third image both contain "a group of people" rather than "one person", which are more related to text semantics. According to the case study, both the performance on image-to-text and text-to-image retrieval tasks demonstrated the effectiveness and superiority of our proposed method ACF and its re-ranking-enhanced version, *i.e.*, ACF-R$^+$.

5. Conclusion

In this paper, we presented a novel asymmetry-sensitive method for multi-modal retrieval based on contrastive learning. Concretely, in order to address fine-grained information asymmetry issues, we generated corresponding positives and negatives for each asymmetry type, which are fully utilized in the optimization of contrastive learning. Our method can augment the sensitivity to the subtle differences between multiple modalities, thereby facilitating the generation of more discriminative multi-modal semantic representations. This enhancement is instrumental in bolstering the effect of subsequent image-text retrieval tasks. Moreover, from both local and global perspectives, a hierarchical cross-modal fusion module was proposed to capture sophisticated correspondence between visual and semantic modalities through

the multi-modal attention mechanism. Additionally, a re-ranking strategy without additional training was introduced to enhance retrieval performance, which incorporates bidirectional retrieval information and maximizes utilization of information within the image-text similarity matrix. Experimental results performed on two public datasets have demonstrated the effectiveness and superiority of our proposed method compared to the state-of-the-art baselines.

CRediT authorship contribution statement

Ziyu Gong: Conceptualization of this study, Methodology, Software, Writing - original draft. **Yihua Huang:** Writing - review and editing. **Chunhua Yu:** Resources, Funding acquisition. **Peng Dai:** Resources, Funding acquisition. **Xing Ge:** Investigation, Funding acquisition. **Yiming Shen:** Data curation, Funding acquisition. **Yafei Liu:** Data curation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Science and Technology Project of State Grid Corporation of China: Research and Application of Key Technologies of Electronic Data Storage and Verification in Material Supply Chain With Multi-Credible Sources (No. 5700-202418240A-1-1-ZN). We would like to express our gratitude to the reviewers for their comments and suggestions on this paper, and thank Chengcheng Mai (ph.D., Nanjing Normal University, School of Computer Science and Electronic Information/School of Artificial Intelligence) for his help and advice. This work has previously been accepted by ICME 2024. On this basis, we further proposed a re-ranking based optimization method for improving the multi-modal retrieval performance. The extended experimental results have proved the effectiveness of the augmented strategy.

References

[1] Y. Huang, J. Tang, Z. Chen, R. Zhang, X. Zhang, W. Chen, Z. Zhao, T. Lv, Z. Hu, W. Zhang, Structure-clip: Enhance multi-modal language representations with structure knowledge, CoRR (2023). doi:10.48550/arXiv.2305.06152.

[2] M. Agosti, S. Marchesin, G. Silvello, Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval, ACM Trans. Inf. Syst. 38 (2020) 38:1–38:48.

[3] J. Shi, V. Chaurasiya, Y. Liu, S. Vij, Y. Wu, S. Kanduri, N. Shah, P. Yu, N. Srivastava, L. Shi, G. Venkataraman, J. Yu, Embedding based retrieval in friend recommendation, in: SIGIR, 2023.

[4] C. Ruan, A. Stewart, H. Li, R. Ye, D. Vengerov, H. Wang, Dynamic embedding-based retrieval for personalized item recommendations at instacart, in: WWW '23 Companion, 2023.

[5] H. Cheng, S. Wang, W. Lu, W. Zhang, M. Zhou, K. Lu, H. Liao, Explainable recommendation with personalized review retrieval and aspect learning, in: ACL, 2023.

[6] S. Li, F. Lv, T. Jin, G. Lin, K. Yang, X. Zeng, X. Wu, Q. Ma, Embedding-based product retrieval in taobao search, in: SIGKDD, 2021.

[7] W. Zhu, X. Lv, B. Yang, Y. Zhang, X. Yong, L. Xu, Y. Feng, H. Zhang, Q. Da, A. Zeng, R. Chen, Cross-lingual product retrieval in e-commerce search, in: PAKDD, 2022.

[8] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: improving visual-semantic embeddings with hard negatives, in: British Machine Vision Conference, 2018.

[9] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: CVPR, 2019.

[10] K. Wen, X. Gu, Q. Cheng, Learning dual semantic relations with graph attention for image-text matching, IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 2866–2879. URL: https://doi.org/10.1109/TCSVT.2020.3030656. doi:10.1109/TCSVT.2020.3030656.

[11] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, UNITER: universal image-text representation learning, in: ECCV, 2020.

[12] J. Li, R. R. Selvaraju, A. Gotmare, S. R. Joty, C. Xiong, S. C. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, in: NeurIPS, 2021.

[13] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang, Visualbert: A simple and performant baseline for vision and language, CoRR (2019). URL: http://arxiv.org/abs/1908.03557.

[14] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, in: ECCV, 2020.

[15] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: NeurIPS, 2019.

[16] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: ICCV, 2015.

[17] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, S. Savarese, Ulip: Learning unified representation of language, image and point cloud for 3d understanding, in: CVPR, 2023.

[18] Z. luo, P. Zhao, C. Xu, X. Geng, T. Shen, C. Tao, J. Ma, Q. lin, D. Jiang, Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text retrieval, in: CVPR, 2023.

[19] X. Wang, L. Li, Z. Li, X. Wang, X. Zhu, C. Wang, J. Huang, Y. Xiao, Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models, in: WSDM, 2023.

[20] J. Chen, H. Hu, H. Wu, Y. Jiang, C. Wang, Learning the best pooling strategy for visual semantic embedding, in: CVPR, 2021.

[21] G. Moro, S. Salvatori, G. Frisoni, Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval, Neurocomputing 538 (2023). URL: https://doi.org/10.1016/j.neucom.2023.03.057.

[22] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: CVPR, 2021.

[23] Z. Gan, Y. Chen, L. Li, C. Zhu, Y. Cheng, J. Liu, Large-scale adversarial training for vision-and-language representation learning, in: NeurIPS, 2020.

[24] A. Eisenschtat, L. Wolf, Linking image and text with 2-way nets, in: CVPR, 2017.

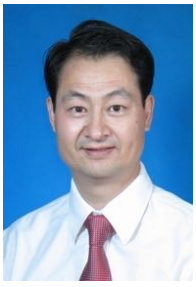
[25] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multim. Comput. Commun. Appl. 16 (2020) 51:1–51:23. URL: https://doi.org/10.1145/3383184. doi:10.1145/3383184.

[26] Y. Huang, Q. Wu, W. Wang, L. Wang, Image and sentence matching via semantic concepts and order learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 636–650.

[27] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence

and Statistics, 2010.
[28] Y. Li, D. Wu, Y. Zhu, A multiple positives enhanced NCE loss for image-text retrieval, in: MMM, 2022.
[29] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: CVPR, 2015.
[30] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, J. Shao, Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation, in: ICCV, 2021.
[31] G. Li, N. Duan, Y. Fang, D. Jiang, M. Zhou, Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining, in: AAAI, 2020.
[32] B. Zhang, H. Hu, V. Jain, E. Ie, F. Sha, Learning to represent image and text with denotation graph, in: EMNLP, 2020.
[33] Q. Cheng, Q. Guo, X. Gu, Adversarial pre-optimized graph representation learning with double-order sampling for cross-modal retrieval, Expert Systems with Applications 231 (2023) 120731. URL: https://doi.org/10.1016/j.eswa.2023.120731.
[34] Z. Fan, Z. Wei, Z. Li, S. Wang, X. Huang, J. Fan, Negative sample is negative in its own way: Tailoring negative sentences for image-text retrieval, in: NAACL, 2022.
[35] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, R. Hu, Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing, IEEE Trans. Multim. 18 (2016) 2553–2566.
[36] D. Qin, S. Gammeter, L. Bossard, T. Quack, L. V. Gool, Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors, in: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011, IEEE Computer Society, 2011, pp. 777–784.
[37] X. Shen, Z. Lin, J. Brandt, S. Avidan, Y. Wu, Object retrieval and localization with spatially-constrained similarity measure and k-nn reranking, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, IEEE Computer Society, 2012, pp. 3013–3020.
[38] W. H. Hsu, L. S. Kennedy, S. Chang, Video search reranking via information bottleneck principle, in: Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, October 23-27, 2006, ACM, 2006, pp. 35–44.
[39] L. Yang, A. Hanjalic, Supervised reranking for web image search, in: Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010, ACM, 2010, pp. 183–192.
[40] L. Yang, A. Hanjalic, Prototype-based image search reranking, IEEE Trans. Multim. 14 (2012) 871–882.
[41] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 3652–3661.
[42] Q. Leng, R. Hu, C. Liang, Y. Wang, J. Chen, Person re-identification with content and context re-ranking, Multim. Tools Appl. 74 (2015) 6989–7014.
[43] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, X. Sun, Remote sensing cross-modal text-image retrieval based on global and local information, IEEE Trans. Geosci. Remote. Sens. 60 (2022) 1–16.
[44] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. URL: https://doi.org/10.1109/TPAMI.2016.2577031. doi:10.1109/TPAMI.2016.2577031.
[45] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (2017) 32–73. doi:10.1007/s11263-016-0981-7.
[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.
[47] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019.
[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017.
[49] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014.
[50] K. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: ECCV, 2018.
[51] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.
[52] L. Qu, M. Liu, J. Wu, Z. Gao, L. Nie, Dynamic modality interaction modeling for image-text retrieval, in: SIGIR, 2021.
[53] H. Lu, N. Fei, Y. Huo, Y. Gao, Z. Lu, J.-R. Wen, Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval, in: CVPR, 2022.
[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, 2021.
[55] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, J. Fu, Seeing out of the box: End-to-end pre-training for vision-language representation learning, in: CVPR, 2021.
[56] M. Cheng, Y. Sun, L. Wang, X. Zhu, K. Yao, J. Chen, G. Song, J. Han, J. Liu, E. Ding, J. Wang, Vista: Vision and scene text aggregation for cross-modal retrieval, in: CVPR, 2022.
[57] H. Lu, N. Fei, Y. Huo, Y. Gao, Z. Lu, J. Wen, COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval, in: CVPR, 2022.
[58] S. Sun, Y. Chen, L. Li, S. Wang, Y. Fang, J. Liu, Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval, in: NAACL-HLT, 2021.
[59] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Ziyu Gong received the bachelor's degree from the School of Informatics, Xiamen University, Fujian, China, in 2022. She is currently working toward the Ph.D. degree with the Department of Computer Science and Technology, Nanjing University, Nanjing, China. Her research interests include multi-modal retrieval, text mining, and large language model. She has published several research papers in ICME 2024, ICMR 2024.



Yihua Huang received the PhD degree from Nanjing University (NJU). He is currently a professor of the Department of Computer Science and Technology and State Key Laboratory for Novel Software Technology at Nanjing University. His research interests include data mining, machine learning algorithms & systems for big data, and parallel computing. He has published over 60 works in leading academic journals and conferences and served as the deputy director of the Big Data Expert Committee of the China Computer Society.

Chunhua Yu received the master degree from the School of Economics and Management, Southeast University, China, in 2012. His research interests include supply chain management, sustainable operations management, corporate social responsibility.

Peng Dai received the master degree from the School of Automation, Nanjing University of Science and Technology, China, in 2014. His research interests include smart grid, supply chain management, and artificial intelligence in electricity.

Xing Ge received the master degree from the Department of Electrical Engineering, North China Electric Power University, China, in 2018. Her research interests include deep learning and new energy grid connection security.

Yiming Shen received the master degree from the College of Computer Science, Chongqing university, China, in 2020. Her research interests include deep learning, image recognition, and big data retrieval.

Yafei Liu received the master degree from the School of Electrical Engineering, Southeast University, China, in 2019. Her research interests include data mining, improved benders decomposition, and supply chain management.