Learnable Pillar-based Re-ranking for Image-Text Retrieval

Leigang Qu
leigangqu@gmail.com
National University of Singapore

Meng Liu
mengliu.sdu@gmail.com
Shandong Jianzhu University

Wenjie Wang∗
wenjiewang96@gmail.com
National University of Singapore

Zhedong Zheng
zdzheng@nus.edu.sg
National University of Singapore

Liqiang Nie
nieliqiang@gmail.com
Harbin Institute of Technology (Shenzhen)

Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore

ABSTRACT

Image-text retrieval aims to bridge the modality gap and retrieve cross-modal content based on semantic similarities. Prior work usually focuses on the pairwise relations (i.e., whether a data sample matches another) but ignores the higher-order neighbor relations (i.e., a matching structure among multiple data samples). Re-ranking, a popular post-processing practice, has revealed the superiority of capturing neighbor relations in single-modality retrieval tasks. However, it is ineffective to directly extend existing re-ranking algorithms to image-text retrieval. In this paper, we analyze the reason from four perspectives, i.e., generalization, flexibility, sparsity, and asymmetry, and propose a novel learnable pillar-based re-ranking paradigm. Concretely, we first select top-ranked intra- and inter-modal neighbors as pillars, and then reconstruct data samples with the neighbor relations between them and the pillars. In this way, each sample can be mapped into a multimodal pillar space only using similarities, ensuring generalization. After that, we design a neighbor-aware graph reasoning module to flexibly exploit the relations and excavate the sparse positive items within a neighborhood. We also present a structure alignment constraint to promote cross-modal content based on semantic similarities. Prior work usually focuses on the pairwise matching relations between an image and a text, as shown in Figure 1(a). They completely ignore the higher-order neighbor relations among multiple images and texts illustrated in Figure 1(b), and therefore suffer from suboptimal retrieval performance. As a prevailing post-processing technology to explore neighbor relations, re-ranking has revealed remarkable effectiveness in conventional single-modality retrieval tasks, such as image retrieval [16, 46], document retrieval [27, 29], and recommendation [41]. As shown by the blue line in Figure 2, the retrieval performance of image retrieval can be largely improved via query expansion re-ranking (i.e., aQE [35]). This motivates us to investigate re-ranking to exploit neighbor relations in the context of ITR.

In general, existing re-ranking approaches can be divided into three main categories: Query Expansion (QE) [8, 35], Diffusion [16], and Neighbor-based methods [48]. However, they are not suitable for cross-modal ITR due to the following challenges: 1) Architecture Generalization. Some re-ranking methods strongly rely on specific model architectures. For example, QE can be only applied to two-tower frameworks since it requires intermediate features for expansion [35], and thus can not well serve the one-tower

CCS CONCEPTS

• Information systems → Novelty in information retrieval; Multimedia and multimodal retrieval.

KEYWORDS

Image-Text Matching; Cross-Modal Retrieval; Re-ranking

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION

Image-text retrieval (ITR), as a fundamental vision-language task, has attracted substantial attention over the past decade [5, 7, 34]. It aims to bridge the heterogeneous modality gap and achieve semantic matching by bidirectional retrieval: 1) image-to-text retrieval, searching for correct descriptive texts given an image, and 2) text-to-image retrieval, finding the relevant images for a given text. By pursuing cross-modal understanding and reasoning, ITR has become essential to other visual-language applications such as visual question answering [1] and video moment retrieval [23].

Existing approaches to ITR fall into two categories: The first category, i.e., two-tower framework [7, 19], separately maps the visual and textual samples into a joint embedding space to calculate the cross-modal similarity. In contrast, the second category, i.e., one-tower framework [5, 34], models fine-grained interactions by the cross-attention mechanism and directly outputs the similarity. Despite impressive progress, both categories only consider the pairwise matching relations between an image and a text, as shown in Figure 1(a). They completely ignore the higher-order neighbor relations among multiple images and texts illustrated in Figure 1(b), and therefore suffer from suboptimal retrieval performance. As a prevailing post-processing technology to explore neighbor relations, re-ranking has revealed remarkable effectiveness in conventional single-modality retrieval tasks, such as image retrieval [16, 46], document retrieval [27, 29], and recommendation [41]. As shown by the blue line in Figure 2, the retrieval performance of image retrieval can be largely improved via query expansion re-ranking (i.e., aQE [35]). This motivates us to investigate re-ranking to exploit neighbor relations in the context of ITR.

In general, existing re-ranking approaches can be divided into three main categories: Query Expansion (QE) [8, 35], Diffusion [16], and Neighbor-based methods [48]. However, they are not suitable for cross-modal ITR due to the following challenges: 1) Architecture Generalization. Some re-ranking methods strongly rely on specific model architectures. For example, QE can be only applied to two-tower frameworks since it requires intermediate features for expansion [35], and thus can not well serve the one-tower
framework. 2) **Representation Flexibility.** Existing re-ranking approaches [16, 35, 48] are designed with rigid or restricted rules from prior knowledge, making them difficult to automatically explore neighbor relations in complex multimodal data. Specifically, QE aggregates neighbors to enrich the query in a limited unidirectional way, the diffusion mechanism performs ranking on fixed manifolds, and neighbor-based methods [32, 48] heavily depend on certain pre-defined neighbor relations. 3) **Sparse Relevance.** The average ratio of positive items over the benchmark datasets of ITR (e.g., MS-COCO (5K) [20]) is much smaller than those of image retrieval (e.g., RParis6K [31]), specifically 0.02% vs. 2.3%. This makes re-ranking on ITR more challenging since noisy negative information are widely propagated among items when capturing the higher-order neighbor relations. And 4) **Modality Asymmetry.** The two asymmetry retrieval directions of ITR result in different similarity distributions and heterogeneous neighbors, yet existing re-ranking methods are designed for single-modality retrieval, thus failing to consider the asymmetry problem. As shown in Figure 2, we can observe the great performance degradation of αQE (the red line) as the aggregated neighbors increase, even though we choose a two-tower backbone to fit it. We argue that it is caused by flexibility, sparsity, and asymmetry challenges.

To handle the above challenges, we present a Learnable Pillar-based Re-Ranking (LeaPRR) framework for ITR. 1) First, to well fit in with one-tower and two-tower frameworks, we abandon any intermediate content features and only take the final similarities which can be calculated by any framework as the input to our re-ranking model. For the sake of informative entity (i.e., queries and items) representation, we define top-ranked multimodal neighbors as pillars and then reconstruct entities with the similarities between pillars and these entities, as shown in Figure 1(c). In this way, each entity can be mapped into a pillar space. Different from the conventional modality-independent content space, the pillar space is constructed with multimodal pillars to alleviate the asymmetry problem. 2) Moreover, as for the flexibility and sparsity challenges, we propose a neighbor-aware graph reasoning module based on the neighborhood centered with a given query, to flexibly explore complicated neighbor relations, including the query-neighbor and neighbor-neighbor ones. 3) To further deal with the high sparsity of positive items, we exert local and global ranking constraints to reinforce the discriminative power of refined pillar features, which can adaptively absorb valuable information and resist noise from neighbors. 4) Finally, we present a structure alignment constraint by transforming contextual structures from one modality to another, further combating the asymmetry issue.

To sum up, our main contributions are as follows:

- We discuss the limitations of existing re-ranking approaches in the ITR task and present four challenges, i.e., generalization, flexibility, sparsity, and asymmetry.
- To deal with the four challenges, we reformulate re-ranking in the multi-modality field and propose a learnable pillar-based re-ranking paradigm for the first time, which is plug-and-play and applicable to all existing ITR backbones.
- Extensive experiments based on different base ITR backbones on two datasets, i.e., Flickr30K [42] and MS-COCO [20], validate the effectiveness and superiority of our model. The codes and settings are released at https://github.com/LgQu/LeaPRR.

## 2 RELATED WORK

### 2.1 Image-Text Retrieval

Current ITR models can be divided into two categories, i.e., **two-tower framework** and **one-tower framework**, in accordance with the manner for modality interaction. As for the two-tower framework [7, 10, 11, 33, 47], images and texts are independently mapped into a joint feature space in which the semantic similarities are calculated via cosine function or Euclidean distance. For example, Frome et al. [11] represented visual and textual instances in a common modality-agnostic space to assess semantic relevance. To further improve the discriminative power, Zheng et al. [47] argued that the commonly used ranking loss is not effective confronted with large-scale multimodality data and presented a new instance loss to exploit the intra-modal data distribution in an end-to-end manner. With the popularity of bottom-up attention [1], recent research attentions have been shifted to capturing and modeling relations between entities (e.g., objects or words) via intra-modal interaction. As a representative model, VSRN [19] reasons across regions in the visual modality by using GCN and Gated Recurrent Unit network for short-long term relation modeling. Qu et al. [33] discussed the modality asymmetry issue and proposed a multi-view summarization architecture. Chen et al. [7] discovered that simple
global pooling functions can outperform the complicated models and proposed a generalized pooling operator. These two-tower models ensure satisfactory retrieval efficiency as features in each modality can be extracted in parallel and indexed in an offline way, but suffer from poor accuracy due to the coarse-grained matching.

In contrast, the one-tower framework [6, 9, 17, 34] embraces fine-grained cross-modal interactions to achieve more thorough matching between fragments (e.g., objects and words). As the pioneering work, SCAN [17] infers some latent region-word alignments by means of cross-modal interaction. Thereafter, much effort has been dedicated to mining cross-modal matching relations. For instance, Chen et al. [6] incorporated the iteration strategy into SCAN to perform multi-step cross-modal relation reasoning. Liu et al. [21] tailored a graph-based network to explore fine-grained correspondence from node level and structure level. Diao et al. [9] inferred multi-level semantic correlations in a similarity graph and filtered noisy alignments via the attention mechanism. Differently, Qu et al. [34] argued possible optimal patterns may not be explored by existing modality interaction patterns and developed a unified dynamic modality interaction modeling network for automatic routing. Despite the superior retrieval accuracy, this line of work has the problem of heavy computational overhead.

To sum up, most existing works focus on representing data samples in joint embedding space or intensively excavating fine-grained cross-modal alignments, but they ignore the efficient and practical post-processing procedure, i.e., re-ranking.

3.2 Re-ranking

As an effective and practical technology, re-ranking has attracted a flurry of interest in traditional single-modality retrieval tasks, such as image retrieval [3, 8, 12, 18, 28, 44, 48] and document retrieval [4, 24–26, 43, 49]. A rich line of studies has explored Query Expansion by using the nearest neighbors of a query to generate a new enriched one. For instance, average query expansion (AQE) [8] was first proposed to aggregate local features by mean-pooling the top-k images in the original ranking list. Afterward, Gordo et al. [12] and Radenovic et al. [35] argued that the mean-pooling would induce much noise information and presented AQEwD and αQE, respectively. Another thread of work is Diffusion-based re-ranking strategy, which depends on a neighborhood topology constructed in an offline way and utilizes it at query time for search on the manifold. For example, Iscen et al. [16] captured the image manifold in the region-level embedding space and proposed a fast spectral ranking method [15] to improve the computational efficiency. In addition, some research has been devoted to employing neighborhood relations, especially the k-Reciprocal Nearest Neighbors, for re-ranking. For example, Zhong et al. [48] encoded the k-reciprocal neighbors of a query into a single vector and then conducted re-ranking under the Jaccard distance. Liu et al. [22] built a graph based on neighborhood affinities and leveraged GCN for manifold learning. Recently, Ouyang et al. [28] reconstructed affinity features with a set of anchor images for contextual similarity reasoning. As one of the rare efforts to explore re-ranking for image-text retrieval, the k-reciprocal nearest neighbor searching scheme [38] was designed as post-processing to enhance retrieval performance. Despite the thrilling progresses, existing methods depend heavily on prior knowledge, lacking sufficient flexibility to adaptively model query-neighbor and neighbor-neighbor relations. Besides, they overlook the sparsity and asymmetry issues widely existing in image-text retrieval.

3 METHODOLOGY

The overall framework of our proposed LeaPRR is depicted in Figure 3. In the following, we will first introduce the problem definition for cross-modal ITR re-ranking and then elaborate on Pillar-based Representation and Neighbor-aware Graph Reasoning. Finally, we will describe the Learning strategy in the pillar space.

3.1 Problem Definition

Given a query \( q \) (an image or a text) and the database in another modality, i.e., \( D = \{d_1, \ldots, d_C\} \), we first calculate the initial similarity \( \phi(q, d_i) \) between the query and each item \( d_i \) in \( D \) via a base ITR backbone \( \phi \). It can be implemented by a one- or two-tower network. The former [19] commonly treats \( \phi \) as the cosine similarity between embeddings extracted from two branches, while the latter views \( \phi \) as an integrated inter-modal interaction module, such as cross-attention [6, 39]. Afterward, we can attain an initial ranking list \( R(q, D) \) by sorting similarities in descending order.

The goal of ITR re-ranking is to refine the initial ranking list so those items that semantically match well with the query are ranked higher and vice versa in the refined ranking list, denoted by \( R^+(q, D) \). In this work, we focus on re-ranking top-K items in \( R(q, D) \) since users always attach more attention to top-ranked retrieval results in practical scenarios.

We consider two retrieval settings: Image-to-Text (I2T), i.e., \( q : I, D = DT \), and Text-to-Image (T2I), i.e., \( q : T, D = DI \). Considering the high consistency except for the retrieval direction, we will elaborate our method under the I2T setting without loss of generality. Besides, we will use items to denote all retrieved texts (or images) for a given image (or text) query. neighbors to refer to top-ranked items, and entities to represent any queries or items.

3.2 Pillar-based Representation

For two-tower I2T frameworks, we can easily access intermediate features of each modality and leverage them to enrich the query or item representation for re-ranking [13, 35]. Nevertheless, this is not possible with one-tower I2T frameworks due to the existence of cross-modal interaction modules. In the light of this, it is unrealistic to use the intermediate representations to optimize the ranking results of both two-tower and one-tower frameworks. Therefore, we resort to cross-modal similarities for image-text retrieval re-ranking. However, how to represent an image or a text and compare them only depending on similarities, becomes a problem to be solved.

As the saying goes, "birds of a feather flock together." We can describe a person with inherent appearance characteristics, such as height, weight, skin color, face shape, and hairstyle. Meanwhile, we can also depict this person by means of the relation between him and other people. For instance, it is reasonable to believe that two people are similar in appearance if they have common family members, or similar in interest if with common idols and/or friends. This inspired us that we can not only represent an entity using its
internal content/attributes but also can do it with external relations between it and other entities.

Although others can be employed as references to represent an item, it does not mean anyone can take responsibility. As an illustration, those people that the person is unfamiliar with would not be helpful as references to discriminate him against. Therefore, the relations between an item and the selected references should not be very sparse. We will detail how to select such references and exploit them as references to represent an item, in the following.

3.2.1 Multimodal Pillar Selection. Given the initial ranking list \( R(I, D_T) = \{T_1, ..., T_N\} \) of I2T retrieval, we select top-\( L (L < N) \) texts as the above-mentioned references. In our work, these selected references are defined as pillars. For the sake of simplicity, we re-denote \( R(I, D_T) \) as \( R^{I \rightarrow T} \), and utilize \( R^{I \rightarrow T}_{L} = \{T_1, ..., T_L\} \) to represent the inter-modal pillars of the query \( I \). To further enhance the representation ability, we propose to extract intra-modal pillars from the intra-modal database \( D_I \). Concretely, we apply the image encoder \( \psi_I \) to extract features and then calculate pairwise similarities between them. After that, based on the intra-modal ranking list \( R(I, D_I) = \{I_1, ..., I_M\} \), \( R^{I \rightarrow I}_{L} \) for short, we collect a set of intra-modal pillars\(^1\) of query \( I \) as \( R^{I \rightarrow I}_{L} = \{I_1, ..., I_L\} \).

\(^1\)In this paper, we use the same number of intra- and inter-modal pillars for the sake of simplicity. In practice, it is acceptable to consider different numbers of pillars.

3.2.2 Pillar-based Encoding. On the basis of pillars, we encode the query image \( I \) and its top-\( K \) neighbor texts \( \{T_1, ..., T_K\} \) as follows:

\[
\begin{align*}
\upsilon &= \left[ \psi^L_I(I, T_j), \psi^L_I(I, I_j) \right], \\
\upsilon_k &= \left[ \psi^L_I(T_k, T_j), \psi^L_I(T_k, I_j) \right],
\end{align*}
\]

where \( \odot \) denotes the concatenation operator, \( \upsilon \in \mathbb{R}^{2L} \) and \( \upsilon_k \in \mathbb{R}^{2L} \) respectively represent pillar-based representations of the query \( I \) and its \( k \)-th neighbor \( T_k \). \( \phi \) can be any base backbone that calculates image-text similarities, and \( \pi \) is the intra-modal similarity calculation model. Specifically, \( \pi(I, I_j) = \cos(\psi_I(I), \psi_I(I_j)) \), \( \pi(T_k, T_j) = \cos(\psi_T(T_k), \psi_T(T_j)) \), where \( \psi_I \) and \( \psi_T \) are image encoder and text encoder, respectively.

3.3 Neighbor-aware Graph Reasoning

One of the keys to re-ranking is to model the high-order neighbor relations in a neighborhood including a query and its neighbors. Prior approaches try to achieve it from different aspects. For instance, QE-based models [2, 35] exploit query-neighbor interaction to expand query representation by aggregating neighbors. In contrast, diffusion-based [16, 30] models perform one-way information flow by building a connected graph. Nevertheless, these models are designed based on some rules with limited flexibility. To tackle
this issue, we tailor a graph-based neighbor propagation to capture query-neighbor and neighbor-neighbor relations adaptively.

After finishing pillar encoding, we are able to represent the query and its neighbors in a pillar space. In this space, we view each item \( (i.e., \) a query or a neighbor) as a node and then connect them according to their affinity scores.

### 3.3.1 Neighbor-based Affinity

To exploit neighbor relations from the raw feature space, we first collect top-\( C \) neighbors from both intra- and inter-modal ranking lists for each item and then merge them into one set, as,

\[
\tilde{R}_i = \begin{cases} R_{i,C}^{I \rightarrow I} \cup R_{i,C}^{I \rightarrow T} & \text{node } i \text{ is an image}, \\ R_{i,C}^{T \rightarrow I} \cup R_{i,C}^{T \rightarrow T} & \text{node } i \text{ is a text}, \\ \end{cases}
\]

where \( \tilde{R}_i \) denotes the neighbor set of the \( i \)-th node. \( R_{i,C}^{I \rightarrow I} \) refers to the set of top-\( C \) neighbors of the \( i \)-th node in the direction \( I \rightarrow I \), and the same is true for other directions. Like \( I \rightarrow I \), we also apply a text encoder to obtain \( R_{i,C}^{T \rightarrow T} \).

Afterward, we calculate the edge weight \( C_{ij} \) between the \( i \)-th node and the \( j \)-th node depending on their common neighbors:

\[
C_{ij} = \frac{|\tilde{R}_i \cap \tilde{R}_j|}{\sum_{k=1}^{K}|\tilde{R}_i \cap \tilde{R}_k|},
\]

where \(| \cdot |\) denotes the element number of the set. To alleviate the influence of noisy edges, we introduce a sparse factor \( \lambda \) and the edge weight is updated as follows,

\[
\hat{C}_{ij} = \begin{cases} C_{ij}, & C_{ij} > \frac{\lambda}{1 + K}, \\ 0, & \text{otherwise}. \\ \end{cases}
\]

Following the row-wise normalization of \( \hat{C} \), we could obtain the neighbor-based affinity matrix, denoted as \( \hat{A} \in \mathbb{R}^{(1 + K) \times (1 + K)} \).

### 3.3.2 Learning-based Affinity

To endow the model with sufficient flexibility, we drive it to learn informative relations. Specifically, we first stack the pillar-based representations of the query and its neighbors to build a matrix \( F = [v_1, t_1; \ldots; v_K, t_K] \in \mathbb{R}^{(1 + K) \times 2L} \). And then we calculate the affinity matrix \( \hat{A} \in \mathbb{R}^{(1 + K) \times (1 + K)} \) as,

\[
\hat{A} = \text{softmax}(f_0(F) \cdot f_K(F)^	op),
\]

where \( f_0 \) and \( f_K \) respectively denote two fully-connected (FC) layers and softmax() is performed over each row. \( \hat{A}_{ij} \) represents a learning-based affinity score between node \( i \) and node \( j \).

### 3.3.3 Neighbor Propagation

Combining the above two types of affinities, we derive the final affinity matrix by \( \hat{A} = (\hat{A} + \hat{A})/2 \) and perform the neighbor propagation in the first layer of GCN as,

\[
F^{(1)} = g(A \cdot f_Y(F)) + F,
\]

where \( f_Y \) is an FC layer and \( g \) denotes a multi-layer perceptron for feature transformation followed by residual connection. After the first layer of GCN, we can re-compute the learning-based affinity matrix using Eqn. (5) and then perform the following propagation as Eqn. (6) does. The output of the last layer GCN, denoted as \( F^* \), is treated as the refined pillar-based representation matrix.

### 3.4 Learning from Pillar Space

Under pairwise supervision between images and texts, our model is optimized with the following three constraints.

#### 3.4.1 Global Contrastive Loss

We adopt the contrastive loss [14] to enforce the query close to positives and far away from negatives as much as possible in a global perspective, which is defined as,

\[
\mathcal{L}^{I \rightarrow T}_c = -\log \frac{\sum_{i \in \mathcal{P}} \exp(s(I, T_i)/\tau)}{\sum_{i \in \mathcal{P}} \exp(s(I, T_i)/\tau)},
\]

where \( \mathcal{P} \) denotes the positive set of \( I \), \( \tau \) is a scalar temperature factor, and \( s(I, T_i) = \cos(v^*, t_i^*) \) refers to the cosine similarity between the refined pillar-based representations of the query image \( I \) and its \( i \)-th textual neighbor \( T_i \).

#### 3.4.2 Local Triplet Loss

Different from the contrastive loss that forces similarities of all positive pairs essentially, the hinge-based triplet loss [37] focuses on local relative proximity and allows the variance between positives to some extent, making it less greedy. In this paper, considering negative samples in top-\( K \) neighbors are already hard, we do not conduct the hardest negative mining but turn to the hardest positive instead, which makes it complementary to the contrastive loss. Formally, we formulate the triplet loss as,

\[
\mathcal{L}^{I \rightarrow T}_1 = \sum_{i \in \mathcal{N}} \left[ 1 - s(I, \hat{T}) + s(I, T_i) \right]^+, \]

where \( \alpha \) is the scalar margin, \( \mathcal{N} \) represents the negative set of \( I \), and \( \mathcal{P} \cup \mathcal{N} = R^{I \rightarrow T}_i \). \( x^+ = \max(x, 0) \) is the hinge function, and \( \hat{T} \) is the hardest positive defined as \( \hat{T} = \arg\min_{i \in \mathcal{P}} s(I, T_i) \).

#### 3.4.3 Mutual Modality Alignment

Until now, we have illustrated the mechanism of the I2T submodel. Similarly, we can derive the T2I submodel by respectively assigning the query and database as the text \( T \) and \( \mathcal{D}_T \). To ensemble these two submodels, a naive way is to directly average two similarities, as most of prior base backbones [17, 21, 34] do. In this way, however, two submodels are trained independently and unable to promote each other. Inspired by mutual learning [40, 45], we devise a mutual modality alignment scheme to spur two submodels to teach and promote each other together.

Taking I2T as an example, we first carry out cross-modal positive sampling for the query \( I \) and each neighbor \( T_i \) in \( \mathcal{D}_T \), obtaining \( T' = \delta(I) \) and \( T_i' = \delta(T_i) \), where \( \delta(\cdot) \) is the sampling function. Concretely, for a given entity, we first find its positive matches in another modality by the pairwise ground-truth information and then uniformly sample one if there are multiple positives. Afterward, we respectively calculate two similarity distributions centered with \( I \) and \( T' \), as,

\[
\mathcal{P}_I = \text{softmax}(s(I, T'_1)/\tau, \ldots, s(I, T'_K)/\tau),
\]

\[
\mathcal{Q}_{T'} = \text{softmax}(s(T'_1, I)/\tau, \ldots, s(T'_K, I)/\tau).
\]

Finally, the alignment loss \( \mathcal{L}_a \) is defined based on Kullback–Leibler divergence \( D_{KL} \) as,

\[
\mathcal{L}_a^{I \rightarrow T} = D_{KL}(\mathcal{P}_I \| \mathcal{Q}_{T'}).\]
We validate our method on two large-scale datasets: Flickr30K [42] and MS-COCO (5K) [20]. Following the existing work [7, 19, 34], we conduct evaluation by the standard recall metric R@K (K = 1, 5, and 10) with 5 sentences. Following the split of [17, 19, 34], we use 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. In particular, we adopt the challenging evaluation setting and collect intra-modal pillars, i.e., $R_{I \rightarrow T}$ and $R_{T \rightarrow I}$ for all base backbones and experiments.

### 4 EXPERIMENTS

In this section, we conducted experiments on two benchmark datasets to answer the following research questions:

- **RQ1**: How does LeaPRR perform in the image-text retrieval task compared with state-of-the-art baselines?
- **RQ2**: How does each component of LeaPRR affect the retrieval performance?
- **RQ3**: How are the generalization and transferability of LeaPRR when facing different base ITR backbones and datasets?

#### 4.1 Datasets

We validate our method on two large-scale datasets: Flickr30K [42] and MS-COCO [20].

- **Flickr30K [42]**: It comprises 31,783 images collected from Flickr, and each image corresponds to 5 human-annotated sentences. We employ the same split as [19, 33] did, specifically, 29,783 training images, 1,000 validation images, and 1,000 testing images.

- **MS-COCO [20]**: It includes 123,287 images and each is annotated with 5 sentences. Following the split of [17, 19, 34], we use 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. In particular, we adopt the challenging evaluation setting and collect intra-modal pillars, i.e., $R_{I \rightarrow T}$ and $R_{T \rightarrow I}$ for all base backbones and experiments.

#### 4.2 Experimental Settings

##### 4.2.1 Evaluation Protocols

The sparse factor in Eqn. (4) is set as 0.8. The number of GCN layers is set to 2 for neighbor propagation. The margin $\alpha$ in local triplet loss is set to 2 for neighbor propagation. The margin $\alpha$ in local triplet loss and the temperature factor $\tau$ are set as 0.2 and 1.0, respectively. In addition, we used the SGD optimizer with a momentum of 0.9 and batch size of 512. Our model is trained for 30 epochs with a learning rate of 0.001. The checkpoint with the highest rSum on the validation set is chosen for testing.

In this paper, we directly applied the visual and textual branches of the pre-trained two-tower model CAMERA [33] published in the open-source community to calculate intra-modal similarities and collected intra-modal pillars, i.e., $R_{I \rightarrow T}$ and $R_{T \rightarrow I}$ for all base backbones and experiments.

#### 4.3 Performance Comparison (RQ1)

This section demonstrates the comparison with other state-of-the-art re-ranking methods, including query expansion based methods (AQE [8], AQEwD [12], and QAE [35]), and them combined with database-side augmentation (DBA) [2], including average augmentation (ADBA), with-decay augmentation (ADBAwD), and weighted augmentation (ADBAwD), respectively, diffusion based methods (DFS [16], FSR [15]), and reciprocal nearest neighbor based top-K of the ranking list and rSum means the sum of all $R@K$ in both image-to-text and text-to-image retrieval directions for overall performance evaluation. The higher R@K and rSum are better.

- **RQ2**: How does each component of LeaPRR affect the retrieval performance?
- **RQ3**: How are the generalization and transferability of LeaPRR when facing different base ITR backbones and datasets?
ones (KRNN [38]). These re-ranking baselines are implemented on top of two representative state-of-the-art base backbones, including a two-tower base backbone VSE∞ [7] and a one-tower base backbone DIME∗ [34]. Note that we directly utilized the published pre-trained parameters to calculate image-text similarities to train and evaluate LeaPRR and other re-ranking baselines. Besides, we inherit ensemble similarities (i.e., averaging two similarities calculated by two single base backbones) for challenging evaluation if available. As query expansion re-ranking baselines (e.g., AQE, AQEwD, and their corresponding DBA versions) require accessing intermediate intra-modal representations, we can only apply them to two-tower base backbones.

The comparisons are summarized in Table 1. From this table, we have the following observations:

- Directly applying Query Expansion methods [12, 35] to the two-tower base backbone, i.e., VSE∞, is not helpful. Concretely, after executing QE-based re-ranking for VSE∞, the performance drops, especially on R@5 and R@10. The main reasons are as follows: 1) The intrinsic sparsity of ITR (discussed in Section 1) may induce top-ranked negative neighbor information to be integrated into the query, therefore overwhelming the informative representation of the original query; and 2) the modality asymmetry makes it difficult to fuse neighbors from a different modality. Besides, performing DBA before QE (e.g., αDBA + αQE) fails to redeem this but even makes it worse, which is mainly attributable to the sparsity and asymmetry issues that have not been tackled.

- By combining DBA and QE at the same time for re-ranking, the performance on some metrics (e.g., recall values at T2I on Flickr30K) gets improved. Among them, 2αDAWd + AQEwD [12] gains the better R@1 (5.0% higher than VSE∞) and 2αDA + αQE [35] performs better on R@5 (2.0% gain) and R@10 (1.8% gain). The reason may be that the combination strategy relieves the asymmetry problem. In other words, the query could absorb certain informative clues from intra-modal neighbors.

- The diffusion approach DFS [16] gains marginal overall performance improvement (0.3% and 0.1% regarding rSum on Flickr30K and MS-COCO (5K), respectively) compared with the base VSE∞ [7]. However, it fails to offer consistent improvement for DIME∗ [34] due to the possible more complicated manifold structure caused by cross-modal interactions in this one-tower architecture.

- The reciprocal neighbor-based method KRNN [38], acting as a strong baseline, surpasses all aforementioned baselines in terms of most metrics. This demonstrates the effectiveness of capturing and modeling reciprocal query-neighbor relations for ITR.

### 4.4 In-depth Analysis (RQ2 & RQ3)

#### 4.4.1 Ablation Study

On top of the challenging state-of-the-art base backbone DIME∗, we conduct a series of experiments on Flickr30K to verify the effectiveness of each component of LeaPRR and explore their impacts as follows.

- **Impact of Pillar-based Representation.** Considering our proposed model performs a series of learning in the pillar space, the quality of this space plays a key role in the overall performance. In this section, we hence first justified the effectiveness of our multimodal pillar selection strategy by designing the following variants: 1) **Bottom Pil.**, collecting L items from the tail of the intra- and inter-modal ranking list to constitute the pillar set; 2) **Random Pil.**, randomly and uniformly sampling L items from the intra- and inter-modal databases; 3) **w/o Inter-Pil.**, removing inter-modal pillars; and 4) **w/o Intra-Pil.**, excluding intra-modal pillars.

From Table 2, we observe that the first two variants Bottom Pil. and Random Pil. get performance degradation as compared to the full model, indicating that inferior pillars would deteriorate retrieval performance. Nevertheless, the performance achieved by these two variants is still comparable with the base backbone DIME∗ [34], demonstrating that our LeaPRR can tolerate bad pillars and avoid performance collapse. Besides, the improvements achieved by the full model LeaPRR compared with two single-pillar variants w/o Intra-Pil and w/o Inter-Pil. verifies the effectiveness of the top-L pillars selected from the intra- or inter-modal database. The remarkable performance improvement achieved by DIME∗+LeaPRR strongly attests to the effectiveness of our proposed selection strategy and the complementary intra-modal and inter-modal pillars.

- **Impact of Neighbor-aware Graph Reasoning.** To assess how the neighbor-aware graph module in Section 3.3 contributes to the performance, we design three variants to compare with the base backbone DIME∗ and LeaPRR by removing the neighbor-based affinity, the learning-based affinity, and both. As displayed in Figure 4(c), the full model achieves the best I2T and T2I retrieval performance compared with other variants, which demonstrates

<table>
<thead>
<tr>
<th>Method</th>
<th>Image-to-Text R@1 R@5 R@1 R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIME∗</td>
<td>81.0 95.9 63.6 88.1</td>
</tr>
<tr>
<td>+ LeaPRR (Bottom Pil.)</td>
<td>78.8 94.4 64.8 87.3</td>
</tr>
<tr>
<td>+ LeaPRR (Random Pil.)</td>
<td>81.8 95.3 62.9 87.3</td>
</tr>
<tr>
<td>+ LeaPRR (w/o Inter-Pil.)</td>
<td>80.3 96.2 66.7 89.0</td>
</tr>
<tr>
<td>+ LeaPRR (w/o Intra-Pil.)</td>
<td>84.3 95.9 67.9 89.1</td>
</tr>
<tr>
<td>+ LeaPRR</td>
<td>86.2 96.7 71.1 89.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Text-to-Image R@1 R@5 R@1 R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIME∗</td>
<td>81.0 95.9 63.6 88.1</td>
</tr>
<tr>
<td>+ LeaPRR (w/o Cont. Loss)</td>
<td>86.1 95.9 70.4 89.4</td>
</tr>
<tr>
<td>+ LeaPRR (w/o Trip. Loss)</td>
<td>85.5 96.1 69.9 88.7</td>
</tr>
<tr>
<td>+ LeaPRR (w/o MMA, only t+i)</td>
<td>83.0 95.0 70.1 89.3</td>
</tr>
<tr>
<td>+ LeaPRR (w/o MMA, only i+t)</td>
<td>86.7 97.8 66.9 89.2</td>
</tr>
<tr>
<td>+ LeaPRR (w/o MMA, i+t &amp; t+i)</td>
<td>86.0 96.8 68.7 89.5</td>
</tr>
<tr>
<td>+ LeaPRR</td>
<td>86.2 96.7 71.1 89.5</td>
</tr>
</tbody>
</table>

### Table 2: Ablation Study on Flickr30K regarding different pillar selection strategies. Pil. denotes Pillar.
the effectiveness of the neighbor-aware graph reasoning module. Besides, we can observe w/o NA performs better than w/o LA, showing learning to estimate edge weights and construct the graph in the pillar space is more important, because it endows LeaPRR with flexibility. Finally, by comparing the two retrieval directions, more performance improvement by contrasting the full model and w/o LA & NA is achieved in I2T. One of the reasons is that the positive distribution is sparser in T2I than I2T. From this perspective, it verifies the existence of the sparsity challenge in Section 1.

- **Impact of Learning Objectives.** To thoroughly analyze the effectiveness of the three learning objectives presented in Section 3.4, we carried out ablation experiments by eliminating each of them. Concretely, we separately remove the global contrastive objective (w/o Cont. Loss), the local triplet objective (w/o Trip. Loss), and mutual modality alignment (w/o MMA). Besides, as to MMA, we further explored the contribution of each branch by building the following variants: 1) w/o MMA, only i-t. 2) w/o MMA, only t-i. and 3) w/o MMA, i-t & t-i (i.e., separately training two single models and then combining them).

By analyzing the comparison results reported in Table 3, we have the following observations: 1) The contrastive objective and the triplet objective are able to enhance the learning of LeaPRR from different perspectives and collaborate with each other to improve the performance. Considering most of the raw content information has been abandoned in the pillar space, sufficient optimization driven by these two constraints is crucial; 2) Compared with the contrastive one, the triplet objective seems to contribute more, which reflects the importance of focusing on local hard structures (e.g., relations among the query, hard positives, and hard negatives); 3) Two single submodels still gain obvious performance improvement in terms of all recall metrics strongly validate the effectiveness of our proposed re-ranking framework; 4) The MMA module plays a critical role in combining two single submodels, aligning cross-modal local structures, and mutual enhancement; And 5) considerable overall performance (rSum) improvement can be gained, though some of metrics (e.g., R@5 in I2T) may sacrifice in ensemble models.

### 4.4.2 Parameter Sensitivity.
To further delve into how LeaPRR performs with regard to pillars and neighbors, we conducted extensive experiments using varying numbers of pillars and neighbors.

- **Impact of the Pillar Number (L).** As shown in Figure 4(a), we evaluated and compared the retrieval performance by selecting various numbers of top-ranked pillars with a fixed exponential interval. From the compared results, we found that saturation points exist in all metrics, roughly at $L = 32$ or 64, except for I2T R@5. Too few pillars are inadequate to represent data samples, while too many pillars have a higher risk to introduce noisy pillars. From this knowable, an appropriate number of pillars is beneficial to make pillar representations rich and discriminative. More importantly, almost all the settings improve the base backbone from different depths, further indicating the effectiveness and robustness.

- **Impact of the Neighbor Size (K).** To investigate the sensitivity of LeaPRR to neighbors, we evaluated its performance under different numbers of neighbors, i.e., $K$, in each submodel and reported the corresponding results in Figure 4(b).

These results show that: 1) The overall performance, i.e., rSum, improves with the increasing neighbors at the beginning. When the number increases to a range of values for each submodel, the performance reaches the saturation point and then decreases. This indicates that exploiting a proper number of neighbors can enhance performance and increase the probability that bottom-ranked positives are recalled. And 2) compared with the i-t submodel, the t-i one gets saturated and starts to rapidly deteriorate earlier. It may attribute to the intrinsic ratio of positives to negatives in existing datasets. As described in Section 4.1, each image in both datasets corresponds to five sentences, i.e., more negative samples exist in the graph and the top-K neighbor set for the t-i submodel than for the i-t one. More importantly, regardless of the number of neighbors, all the settings manage to achieve overall performance improvement over the base backbone, which also verifies the effectiveness and robustness of LeaPRR.

### 4.4.3 Generalization Analysis.
As discussed in Section 1, generalization is essential for re-ranking. To verify whether LeaPRR can perform well for different base ITR architectures, we plug it into five representative ones: two one-tower (SCAN* and DIME*) and three two-tower (VSRN*, CAMERA*, and VSEoo) architectures and compare it with KRNN [38]. Results on Flickr30K and MS-COCO (5K) are reported in Table 4. The superior performance, especially the R@1 performance, on both datasets validates the strong generalization capability of our method and the superiority over the current state-of-the-art KRNN, regardless of base backbones.

### 4.4.4 Transferability Analysis.
The transferability includes backbone transferability and dataset transferability. These two reflect the zero-shot ability across different backbones and datasets, respectively. The bulk of previous efforts on ITR focuses on representing

<table>
<thead>
<tr>
<th>Method</th>
<th>Image-to-Text R@H</th>
<th>Text-to-Image R@H</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAN* [7]</td>
<td>69.0</td>
<td>69.0</td>
</tr>
<tr>
<td>+ KRNN [38]</td>
<td>65.6</td>
<td>63.4</td>
</tr>
<tr>
<td>+ LeaPRR (Ours)</td>
<td>70.0</td>
<td>68.0</td>
</tr>
<tr>
<td>VSRN* [19]</td>
<td>71.5</td>
<td>70.8</td>
</tr>
<tr>
<td>+ KRNN [38]</td>
<td>70.0</td>
<td>68.4</td>
</tr>
<tr>
<td>+ LeaPRR (Ours)</td>
<td>75.3</td>
<td>73.2</td>
</tr>
<tr>
<td>CAMERA* [33]</td>
<td>70.0</td>
<td>69.1</td>
</tr>
<tr>
<td>+ KRNN [38]</td>
<td>68.4</td>
<td>67.6</td>
</tr>
<tr>
<td>+ LeaPRR (Ours)</td>
<td>74.5</td>
<td>73.0</td>
</tr>
<tr>
<td>VSEoo [38]</td>
<td>71.7</td>
<td>71.7</td>
</tr>
<tr>
<td>+ KRNN [38]</td>
<td>69.6</td>
<td>69.0</td>
</tr>
<tr>
<td>+ LeaPRR (Ours)</td>
<td>76.0</td>
<td>75.0</td>
</tr>
<tr>
<td>DMA* [34]</td>
<td>80.0</td>
<td>79.3</td>
</tr>
<tr>
<td>+ KRNN [38]</td>
<td>78.4</td>
<td>77.6</td>
</tr>
<tr>
<td>+ LeaPRR (Ours)</td>
<td>86.2</td>
<td>85.0</td>
</tr>
</tbody>
</table>

Table 4: Quantitative results on the Flickr30K and MS-COCO (5K) datasets based on five base ITR architectures for generalization evaluation. The performance improvement compared with base backbones achieved by the proposed LeaPRR is marked in green.
and aligning visual and textual content. Due to the existence of a large discrepancy between content spaces, these methods suffer from limited transferability. In contrast, our proposed LeaPRR goes beyond raw content information and embraces pillars to exploit relations between entities for representation. We believe such pillar-based representations are more robust and transferable.

- **Backbone Transferability.** To investigate the backbone transferability, we train LeaPRR using similarities calculated by a source base backbone in the training set, and then are evaluated on similarities of the test set calculated by target base backbones. A rSum means the improvement achieved by LeaPRR compared with corresponding base backbones after being transferred.

- **Dataset Transferability.** To further delve into the transferability across datasets, we also carried out corresponding experiments, the results of which are reported in Table 5. The comparison and performance improvement over the base backbone illustrates the outstanding dataset transferability of LeaPRR. It is worth noting that the improvement for Flickr30K → MS-COCO is highly comparable to that of learning on MS-COCO, specifically 11.0 vs. 12.7, although the scale of Flickr30K is only a quarter of that of MS-COCO. We ascribe the strong transferability to the extensive soundness of the pillar-based representation, i.e., representing an entity with its top-ranked neighbors, as discussed in Section 3.2.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed to model the higher-order neighbor relations for image-text retrieval in the re-ranking paradigm. To this end, we first delved into the limitations of prior re-ranking methods in the context of multi-modality and proposed four challenges for image-text retrieval re-ranking. To tackle the four issues, we then reformulated multi-modal re-ranking, constructed a new pillar space in the context of multi-modality and proposed four challenges for image-text retrieval in the re-ranking paradigm. To this end, we first delved into the limitations of prior re-ranking methods in the context of multi-modality and proposed four challenges for image-text retrieval re-ranking. To tackle the four issues, we then reformulated multi-modal re-ranking, constructed a new pillar space with top-ranked neighbors, and proposed a learning-based framework, which can fit in with one-tower and two-tower frameworks and flexibly explore complex neighbor relations among entities. Comprehensive experiments demonstrate the effectiveness, superiority, generalization, and transferability of the proposed method.

In the future, to further promote re-ranking performance, we plan to introduce abundant multimodal external knowledge for the modeling of the higher-order neighbor relations. Besides, inspired by the proposed re-ranking paradigm, we plan to explore the transferability between different downstream retrieval tasks from a broader perspective.