Dual Associations Semantic Enhancement for Image-Text Matching

Xinlin Zhao[1],Tao Yao[1,*] , Yafei Bu[1], Li Liu[1], Yuling Zhang[1], Linliang Zhang[1,*]
[1]School of Computer and Artificial Intelligence, Ludong University,
Yantai, 264025, Shandong, China
*Corresponding author(s).
Contributing authors: linzizhaoxinlin@163.com; yaotao@ldu.edu.cn;
c133567@126.com; liulildu@163.com; zhangyuling@ldu.edu.cn;
linliangzhang@my.swjtu.edu.cn.

Abstract. Image-text matching faces the significant challenge in effectively mitigating the large visual-semantic discrepancy among different modalities. Existing studies mainly address this by projecting multi-modal data into a common subspace for measuring their semantic similarities. However, most of them often overlook the inherent asymmetry for directional associations, i.e., the differences between image-to-text and text-to-image affinities, which often results in inaccurate retrieval results. To tackle the challenge, we propose a Dual Associations Semantic Enhancement (DASE) model to capture bidirectional image-text semantic associations. Specifically, we first build a two-layer GCN fusion network to construct and mine the semantic associations for each modality. And then, due to the inherent asymmetry of directional associations, a Dual Associations Alignment Module (DAAM) is designed to capture the dual associations between visual and textual modalities, enabling comprehensive cross-modal fine-grained interaction. Finally, global alignment is incorporated with the local alignment to achieve full semantic matching across heterogeneous modalities in a unified embedding space. Experimental results on two publicly available datasets demonstrate that the proposed DASE model achieves significant performance improvements in image-text matching tasks compared to baseline methods, validating its effectiveness and superiority.

Keywords: Image-text Matching, Dual Associations, Cross-modal Retrieval.

1 Introduction

With the rapid development of multi-modal data, information retrieval across various types of data has attracted widespread attention. Of these, vision and language are the two most important ways for humans to acquire and convey information. Accordingly, image-text matching has quickly become a key research focus, significantly supporting the development of multi-modal intelligence. This technology consists of using an image to fetch semantically relevant texts and a text to obtain semantically associated images. Image-text matching has been successfully utilized across diverse domains,

such as image captioning[1] and visual question answering[2]. However, it is undeniable that image-text matching still faces severe challenges in narrowing the inherent gap between modalities and accurately learning the semantic correspondence between image and text for measuring their similarity.

The last decade has witnessed remarkable progress in this domain. According to different sampling anchors and matching strategies, existing methods can be roughly divided into two paradigms: global-based matching methods and local-based matching methods. Global-based matching leverages mainstream architectures like ResNet and GRU to construct a two-branch network, which independently extracts global visual and textual features; subsequently, matching similarity is directly estimated in a shared space. However, these approaches with global-based matching often capture semantic associations between modalities in a coarse manner, with the risk of ignoring modal-specific information and fine-grained semantic associations. To capture the fine-grained semantic associations in multi-modal, local-based matching methods have received widespread attention. These methods[3, 4, 5,] typically aggregate image regions at the target object level and all words in a sentence to construct local matching affinity matrix across all regions to guide retrieval between image and text. Recent research[6, 7] have focused on approaches for identifying region-word correspondences to enhance matching details. So scene graphs have rapidly emerged as an efficient tool for high-level semantic understanding tasks by modeling the objects and their relationships[8].

Although numerous image-text matching approaches have been proposed by researchers, existing approaches still face the following limitations. First, current methodologies often overlook the inherent asymmetry for directional associations between image-to-text and text-to-image. Specially, the regions matched by anchoring on text words and those anchored on image patches are different, and the matching similarity scores are also different. For example, as illustrated in **Fig. 1**, an image patch depicting a wooden wall exhibits a strong association with the word "wooden" in text, while the word "wooden" can concurrently exhibit associations with the wooden wall, wooden bench, wooden desk, and wooden chair in the image. This inconsistency implies that the associations from image-to- text are not be necessarily equivalent associations from text-to-image. Therefore, it will impair the cross-modal semantic correspondence and hinder the reduction of the modality gap, if the image-to-text and text-to-image associations are treated as one association. To fully mine the inherent information, most methods inadequately utilize intra-modal neighborhood information among image regions and textual words, which contain critical contextual semantic relationships, thus reducing image-text matching accuracy.

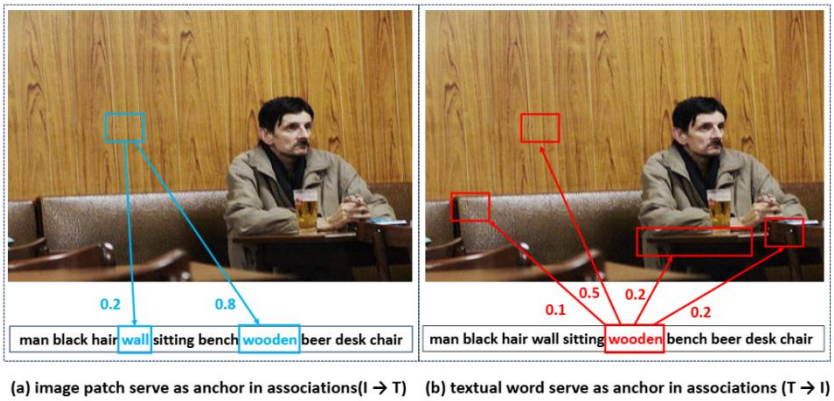

Fig. 1. Exemplifying the inherent asymmetry of directional associations between image-to-text and text-to-image from the perspective of visual patch and textual word.

To tackle the aforementioned problems, we propose a Dual Associations Semantic Enhancement (DASE) model for image-text matching. Specially, to effectively bridge the semantic gap for performing precise matching, a dual associations alignment module (DAAM) is designed to capture bidirectional cross-modal semantic associations. To better capture neighbourhood information, a two-layer GCN fusion module is applied to explicitly mine intricate semantic relationships for enhancing the discrimination of feature representations.

The primary advancements of this framework are crystallized as follows:

- We propose a attention mechanisms-based cross-modal dual associations module (DAAM) that significantly addresses the inherent asymmetry in cross-modal associations between visual and textual modalities.
- To capture intricate intra-modal contextual correlations, we design a two-layer GCN fusion network to aggregate neighborhood information and expand receptive field.
- Extensive experiments comprehensively verify that our proposed DASE has remarkably superior performance over the current baseline methods in image-text matching.

2 Methodology

In this section, we elaborate on the proposed cross-modal Dual Associations Semantic Enhancement (DASE) model. As depicted in **Fig. 2**, the framework comprises three core modules: Feature Extraction Module, Two-layer GCN Fusion Module, and Dual Associations Alignment Module. A detailed description of each module is presented below.

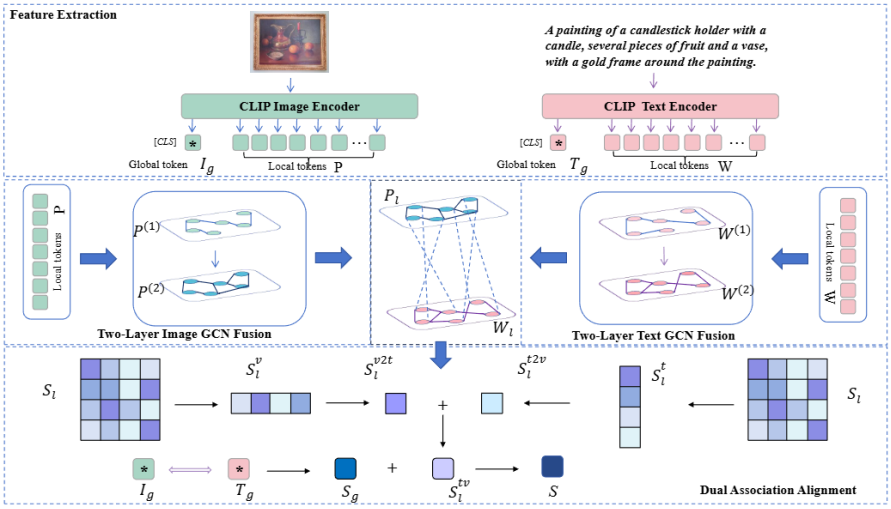

Fig. 2. Illustration of the proposed Dual Associations Semantic Enhancement (DASE) for image-text matching. The proposed DASE comprises three components including feature extraction module, two-layer GCN fusion module, and dual associations alignment module.

2.1 Notation Definition

Several notations in the article are first defined. The main notations are illustrated in **Table 1**.

Table 1. Important Notation Utilized in This Manuscript.

Notation	Description
$\mathcal{V}$	image set
$\mathcal{T}$	text set
I	an image
T	a text
I_g	image global feature extracted by visual encoder
T_g	text global feature extracted by textual encoder
P	image local feature extracted by visual encoder
W	text local feature extracted by textual encoder
P_l	image local feature aggregated by two-layer GCN fusion
W_l	text local feature aggregated by two-layer GCN fusion
n	number of image patches
m	number of text tokens
d	embedding dimension

2.2 Feature Extraction Module

Visual Feature Representation. For a input image $I \in \mathcal{V}$, it is first partitioned into n discrete patches. This discretization transforms the image into a discrete sequence for facilitating subsequent feature extraction and then a learnable [*CLS*] token is prefixed to the sequence. We employ the standard 12-layer architecture of the CLIP model as the image encoder. This encoder generates both image-level feature embeddings and patch-level feature embeddings. Specially, the [*CLS*] token learns a coarse-grained global representation $I_g \in \mathbb{R}^d$, while the remaining patch features form a fine-grained local representation $P = \{ p_i \mid i \in [1,2,\dots,n], p_i \in \mathbb{R}^d \}$.

Textual Feature Representation. For the textual representation, a given sentence $T \in \mathcal{T}$ is initially tokenized by Tokenizer, and then the textual sequence is filled with a [*CLS*] token. The finally generated sequence is input to the textual encoder. Similar to the image encoder, leveraging CLIP's large-scale pretrained knowledge, the text encoder efficiently generates discriminative textual representations, *i.e.* a coarse-grained global representation $T_g \in \mathbb{R}^d$ and a fine-grained local representation $W = \{ w_j \mid j \in [1,2,\dots,m], w_j \in \mathbb{R}^d \}$, where m denotes the number of tokens.

2.3 Two-layer GCN Fusion Module

Owing to the potential defect that the fine-grained local feature representations of images and texts extracted by patches and tokens often neglect of contextual semantic relationships, we design a two-layer GCN fusion network for the local feature representations of images and texts. Specially, intricate intra-modal contextual correlations are captured by aggregating neighborhood information and expanding the receptive field, thereby generating more discriminative feature representations, which contributes to enhancing the accuracy of image-text matching.

Limited by fixed adjacency matrices, traditional graph convolutional networks (GCNs) have the limitations of relying on prior assumptions and weak generalization ability. Thus, a learnable dynamic relationship mechanism is designed to dynamically tune the connection weights between nodes based on the input features.

Concretely, for an image-text pair (*I*, *T*), the image local feature matrix P is set as $P^{(0)}$. The fusion is realized to generate a semantic-enhanced feature representation:

$$P^{(\gamma+1)} = D_4^T \left(\left(\frac{1}{n} \left(D_1^T P^{(\gamma)} \right)^T \left(D_2^T P^{(\gamma)} \right) \right) \left(D_3^T P^{(\gamma)} \right) \right) + P^{(\gamma)} \tag{1}$$

where γ denotes the number of layer, $\mathrm{P}^{(\gamma)}$ and $\mathrm{P}^{(\gamma+1)}$ define the input and output of the γ-th layer in fusion network. $\mathrm{D}_1 \in \mathrm{R}^{\mathrm{n}\times\mathrm{d}}$, $\mathrm{D}_2 \in \mathrm{R}^{\mathrm{n}\times\mathrm{d}}$, $D_3 \in R^{n\times d}$ and $\mathrm{D}_4 \in \mathrm{R}^{\mathrm{d}\times\mathrm{n}}$ are four learnable projection matrices. In the intra-modal graph, the vertices correspond to all patches (words) within an image (text) based on the input feature matrices. And the adjacency relations of this graph are defined by $\frac{1}{\mathrm{n}}\left(\mathrm{D}_1^{\mathrm{T}}\mathrm{P}^{(\gamma)}\right)^{\mathrm{T}}\left(\mathrm{D}_2^{\mathrm{T}}\mathrm{P}^{(\gamma)}\right)$, where a scaled normalization operation is applied to mitigate numerical instability. Thereafter, the feature matrix P, after being multiplied by the learnable matrix $\mathrm{D}_3 \in \mathrm{R}^{\mathrm{n}\times\mathrm{d}}$, is then multiplied by the previously obtained adjacency matrix to yield a new feature matrix. This

step is followed by a projection via the learnable parameter matrix $D_4 \in R^{d\times n}$ to map aggregated features back to the original channel dimension. To mitigate vanishing gradients and preserve input feature information, a residual connection is added between the input $P^{(\gamma)}$ and the aggregated features $D_4^T\left(\left(\frac{1}{n}\left(D_1^T P^{(\gamma)}\right)^T\left(D_2^T P^{(\gamma)}\right)\right)\left(D_3^T P^{(\gamma)}\right)\right)$.

For the textual local feature matrix, W is set as $W^{(0)}$. The text fusion is shown in the following formula:

$$W^{(\gamma+1)} = D_4^T\left(\left(\frac{1}{n}\left(D_1^T W^{(\gamma)}\right)^T\left(D_2^T W^{(\gamma)}\right)\right)\left(D_3^T W^{(\gamma)}\right)\right) + W^{(\gamma)} \tag{2}$$

We define $P_l = P^{(2)}$ and $W_l = W^{(2)}$. Through these two-layer operations, the fusion is realized to generate a semantic-enhanced local image features P_l and local text features W_l by embedding abundant adjacency information into each modal fine-grained semantic representations.

2.4 Dual Associations Alignment Module

To bridge semantic gaps caused by focusing solely on unidirectional association while overlooking the dual associations between image-to-text and text-to-image, we propose a Dual Associations Alignment Module (DAAM). Additionally, coarse-grained feature representations are integrated to achieve comprehensive cross-modal interactions of semantic information.

For an image-text pair (*I*, *T*), P_l and W_l represent local image features and local text features obtained via the two-layer GCN fusion module respectively. We calculate the similarity of image-text on fine-granularity by matrix multiplication.

$$S_l = P_l W_l^T \tag{3}$$

Where $S_l \in R^{n\times m}$, $S_l(i,j)$ denotes the affinity between the i-th image patch and j-th text token.

To establish discriminative associations between images and texts in different directions, attention mechanism is deployed twice based on the obtained fine-grained similarity scores S_l. Specially, the first attention mechanism is applied across both visual and textual perspectives to obtain fine-granularity image-level and text-level similarity vectors. To further acquire the fine-grained instance-level similarity score, we proceed with the second attention operation.

Concretely, for the image perspective, the first attention can be defined as:

$$S_l^v = \sum_{i=1}^{n} \frac{\exp(S_l(i,\cdot)/\tau)}{\sum_{j=1}^{n}(\exp(S_l(j,\cdot)/\tau))} S_l(i,\cdot) \tag{4}$$

where a row-wise softmax is applied to each row of S_l, (*i.e.*, the similarity between one image patch and all text tokens) to convert values into probability distributions. This is followed by a weighted summation to generate the image-level local similarity vector $S_l^v \in R^{\mathbb{1}\times\mathbb{1}}$. τ denotes the temperature parameter.

For the text perspective, the first attention can be defined as:

$$S_l^t = \sum_{i=1}^{m} \frac{\exp(S_l(\cdot, i)/\tau)}{\sum_{j=1}^{m}(\exp(S_l(\cdot, j)/\tau))} S_l(\cdot, i) \tag{5}$$

where a column-wise softmax is applied to each column of S_l, (*i.e.*, the affinity between one text token and all image patches), followed by a weighted summation to produce the text-level local similarity vector $S_l^t \in R^{n\times 1}$.

The second attention operation further enhances interactions between textual words and image and between image patches and text to obtain the final two bidirectional similarity associations between image-to-text and text-to-image.

$$S_l^{v2t} = \sum_{i=1}^{m} \frac{\exp\left(S_l^i(1, i)/\tau\right)}{\sum_{j=1}^{m}(\exp(S_l^v(1, j)/\tau))} S_l^i(i) \tag{6}$$

$$S_l^{t2v} = \sum_{i=1}^{n} \frac{\exp(S_l^t(i, 1)/\tau)}{\sum_{j=1}^{n}(\exp(S_l^t(j, 1)/\tau))} S_l^t(i) \tag{7}$$

where the bidirectional associations $S_l^{v2t} \in R^{1}$and $S_l^{t2v} \in R^{1}$ are the instance-level similarities. Then, the final fine-grained similarity is obtained by averaging the fine-grained instance-level dual associations:

$$S_l^{vt} = (S_l^{v2t} + S_l^{t2v})/2 \tag{8}$$

And the global similarity matrix is also computed via matrix multiplication using the global image feature $I_g \in R^d$ and global text feature $T_g \in R^d$:

$$S_g = I_g^T T_g \tag{9}$$

where $S_g \in R^{1}$ is the affinity score on coarse-granularity.

To achieve comprehensive cross-modal interactions of semantic information, the final similarity score is computed by integrating local dual associations and global similarities:

$$S(I, T) = \left(S_g + S_l^{vt}\right)/2 \tag{10}$$

2.5 Loss Function

To further mitigate heterogeneity gaps, contrastive loss function is leveraged to optimize feature space alignment via further closing semantically similar instances while simultaneously increasing the distance between dissimilar ones. The contrastive loss framework ensures precise matching and discrimination of cross-modal pairs. For a training batch containing N image-text pairs, the loss formulations are defined as:

$$L_{con}^v = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(< I_i, T_i >)}{\sum_{j=1}^{N} \exp\left(< I_i, T_j >\right)} \tag{11}$$

$$L_{con}^t = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(< I_i, T_i >)}{\sum_{j=1}^{N} \exp\left(< I_j, T_i >\right)} \tag{12}$$

$$L_{con} = L_{con}^v + L_{con}^t \tag{13}$$

Here, $<*,*>$ denotes vector inner product. Clearly, L_{con}^v represents the image-to-text contrastive loss. L_{con}^t measures the text-to-image contrastive loss. The cross-modal semantic contrastive loss L_{con} is defined as the combination of the two losses L_{con}^v and L_{con}^t.

3 Experiments

In this section, we compare DASE's outperformance with baseline methods. Evaluations are performed on two widely adopted multi-modal benchmarks: Flickr30K and MSCOCO.

3.1 Implementation Details

All experimental procedures are performed on an NVIDIA RTX 3090 GPU. The batch size is experimentally determined to 128. On the Flickr30K dataset, the DASE is trained 40 epochs using a fixed learning rate of 2e-5 maintained throughout the optimization process. For the MSCOCO dataset, a phased learning rate strategy is implemented: a starting rate of 5e-5 is applied during epochs 1–10 to accelerate coarse feature learning, followed by a reduced rate of 5e-6 for epochs 11–20 to facilitate fine-grained parameter optimization.

Key hyperparameters are configured as follows: the embedding dimension d = 768, the number of image patches n is fixed at 49, and the maximum length of text token m is limited to 25.

Model performance is evaluated using Recall at K(R@K, where K = 1, 5, 10) and the rSum metric. R@K assesses the fraction of ground-truth matches retrieved within the top-K candidates, while rSum aggregates all R@K values to provide a holistic performance measure.

3.2 Methods for Comparison

To verify the efficacy of our proposed approach, we compared our model DASE with current baseline methods, including SCAN[4], GSMN[11], PFAN[3], CVSE[10], SGRAF[12], GPO[13], VSE++[9], NAAF[14], DRCE-A[16], MV-VSE[15], CHAN[17], USER[18] and DVSE[19]. Our proposed method DASE is an unsupervised approach, and to secure impartiality and persuasiveness of comparison results, the aforementioned methods selected are all unsupervised as well. For each retrieval scenario, two tasks are executed: text-to-image (T2I) and image-to-text (I2T) retrieval. For all experimental tables, bold numbers indicate highest score over each column.

Table 2. DASE Results of image-text matching on the Flickr30K Dataset.

	I2T			T2I			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum
VSE++(BMVC'18)	64.6	90.0	92.7	45.6	76.4	84.4	449.7
SCAN(ECCV'18)	67.4	90.3	95.8	48.6	77.7	85.2	465.0
PFAN(IJCAI'19)	70.0	91.8	95.0	50.4	78.7	86.1	482.0
CVSE(ECCV'20)	73.5	92.1	95.8	52.9	80.4	87.8	482.5
GSMN(CVPR'20)	76.4	94.3	94.3	57.4	82.3	89.0	496.8
SGRAF(AAAI'21)	77.8	94.1	97.4	58.5	83.0	88.8	499.6
GPO(CVPR'21)	81.7	95.4	97.6	61.4	85.9	91.5	513.5

NAAF(CVPR'22)	81.9	96.1	98.3	61.0	85.3	90.6	513.2
MV-VSE(IJCAI'22)	82.1	95.8	97.9	63.1	86.7	92.3	517.5
DRCE-A(TCSVT'23)	75.9	93.7	96.2	56.0	80.9	87.6	490.3
CHAN(CVPR'23)	80.6	96.1	97.8	63.9	87.5	92.6	518.5
USER(CVPR'24)	82.6	**97.0**	98.3	63.1	86.7	92.1	519.9
DVSE(TMM'25)	82.1	96.4	98.3	63.4	87.6	92.3	520.1
Ours	**82.8**	96.9	**98.6**	**65.2**	**89.0**	**93.8**	**526.3**

3.3 Results and Analysis

Performance on Flickr30K The performance comparison between DASE and baseline methods on the Flickr30K dataset is reported in **Table 2**. Our method achieves the best overall performance with an rSum of 521.2\%, surpassing the strongest baseline CORA by 1.1\%. For image-to-text retrieval, DASE obtains R@1, R@5, and R@10 scores of 80.9\%, 95.2\%, and 97.1\%, respectively, showing competitive performance compared with recent advanced methods. For text-to-image retrieval, DASE achieves 65.2\%, 89.0\%, and 93.8\% on R@1, R@5, and R@10, respectively, consistently outperforming all baseline methods. In particular, compared with CORA, DASE improves T2I retrieval by 2.9\%, 1.9\%, and 1.2\% on R@1, R@5, and R@10. These results demonstrate that DASE can learn effective and discriminative cross-modal representations, especially for text-to-image retrieval. Since Flickr30K is smaller than MSCOCO, the superior overall performance further indicates that DASE has good generalization ability and is not strongly dependent on large-scale training data.

Table 3. DASE Results of image-text matching on the MSCOCO-1K Dataset.

	I2T			T2I			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum
VSE++(BMVC'18)	64.6	90.0	95.7	52.0	84.3	92.0	478.6
SCAN(ECCV'18)	72.7	94.8	98.4	58.8	88.4	94.8	507.9
PFAN(IJCAI'19)	76.5	96.3	99.0	61.6	89.6	95.2	518.2
CVSE(ECCV'20)	74.8	95.8	98.3	59.9	89.4	95.2	512.7
GSMN(CVPR'20)	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SGRAF(AAAI'21)	79.6	96.2	98.5	63.2	90.7	96.1	524.3
GPO(CVPR'21)	79.7	96.4	98.9	64.8	91.4	96.3	527.5
NAAF(CVPR'22)	80.5	96.5	98.8	64.1	90.7	96.5	527.2
MV-VSE(IJCAI'22)	80.4	96.6	99.0	64.9	91.2	96.0	532.6
DRCE-A(TCSVT'23)	77.6	95.6	98.6	61.7	89.2	95.3	528.1
CHAN(CVPR'23)	81.4	96.9	98.9	66.5	92.1	96.7	532.6
USER(TIP'24)	**82.8**	96.8	98.8	66.1	90.6	95.6	530.5
DVSE(TMM'25)	80.6	96.8	98.7	65.8	91.3	96.1	529.2
Ours	**82.8**	**97.3**	**99.3**	**68.6**	**92.6**	**96.9**	**537.5**

Table 4. DASE Results of image-text matching on the MSCOCO-5K Dataset.

Methods	I2T			T2I			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++(BMVC'18)	41.3	71.1	81.2	30.3	59.4	72.4	355.7
SCAN(ECCV'18)	50.4	82.8	90.0	38.6	69.3	80.4	410.9
PFAN(IJCAI'19)	53.7	83.2	91.0	39.7	69.1	79.8	416.5
CVSE(ECCV'20)	-	-	-	-	-	-	-
GSMN(CVPR'20)	-	-	-	-	-	-	-
SGRAF(AAAI'21)	58.8	84.8	92.1	41.6	70.9	81.5	429.7
GPO(CVPR'21)	58.3	85.3	92.3	42.4	72.7	83.2	434.3
NAAF(CVPR'22)	58.9	85.2	92.0	42.5	70.9	81.4	430.9
MV-VSE(IJCAI'22)	59.1	86.3	92.5	42.5	72.8	83.1	436.3
DRCE-A(TCSVT'23)	56.2	83.0	90.9	40.3	69.5	80.6	420.5
CHAN(CVPR'23)	59.8	87.2	93.3	44.9	74.5	**84.2**	443.9
USER(TIP'24)	**63.7**	**87.4**	**93.5**	44.8	73.4	82.7	445.5
DVSE(TMM'25)	60.5	86.3	92.9	43.8	73.8	83.2	440.5
Ours	63.4	87.2	92.9	**46.0**	**74.6**	83.9	**448.0**

Performance on MSCOCO Table 3 and **Table 4** report the quantitative comparison between DASE and competing methods on the MSCOCO-1K and MSCOCO-5K test sets, respectively. On the MSCOCO-1K test set, as shown in Table~\ref{tab3}, DASE achieves the best overall performance with an rSum of 537.1\%, outperforming the strongest baseline CORA by 4.4\%. For image-to-text retrieval, DASE obtains R@1, R@5, and R@10 scores of 82.7\%, 97.3\%, and 99.3\%, respectively, improving the previous best results by 0.3\%, 0.5\%, and 0.3\%. For text-to-image retrieval, DASE achieves R@1, R@5, and R@10 scores of 68.6\%, 92.6\%, and 96.6\%, respectively. Compared with CORA, DASE improves R@1 and R@5 by 2.4\% and 0.7\%, while achieving a comparable best result on R@10. These results demonstrate that DASE can learn more discriminative cross-modal representations and effectively enhance retrieval accuracy on both retrieval directions.

On the more challenging MSCOCO-5K test set, as illustrated in **Table 4**, DASE also shows strong competitiveness. Although its image-to-text retrieval performance and overall rSum are slightly lower than CORA, DASE achieves the best R@1 and R@5 scores for text-to-image retrieval, reaching 46.0\% and 74.4\%, respectively. These results indicate that DASE has a clear advantage in retrieving relevant images from textual queries under large-scale evaluation settings. Overall, the consistent gains on MSCOCO-1K and the strong text-to-image performance on MSCOCO-5K verify the effectiveness and robustness of DASE in cross-modal semantic alignment.

From the above experimental results, the following conclusions can be drawn: (1) DASE has made good performance in image-text matching. We argue that the results are primarily attributable to the fusion of neighborhood information by two-layer GCN and the reliable cross-modal interaction through dual associations. (2) The retrieval performance from image-to-text remains consistently superior to that from text-to-image.

The phenomenon is mainly related to the imbalance of modality data, which can serve as a focus of future research.

Ablation Study Ablation studies are systematically performed on the Flickr30K benchmark to quantify the individual contributions of core components in the DASE architecture. Three variants of the model are specified below:

- w/o image GCN: This eliminates the image two-layer GCN fusion module.
- w/o text GCN: This discards the text two-layer GCN fusion module.
- w/o dual associations: This employs non-dual associations to substitute the original dual associations mechanism.

Table 5. Ablation studies on Flickr30K.

	I2T			T2I			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum
w/o image GCN	78.2	95.0	97.2	64.0	88.9	93.5	516.8
w/o text GCN	78.6	95.5	98.2	64.6	88.8	93.7	519.4
w/o dual associations	72.2	93.4	97.4	62.7	**89.8**	93.2	508.7
Ours	**82.8**	**96.9**	**98.6**	**65.2**	89.0	**93.8**	**526.3**

The ablation study results on Flickr30K are reported in **Table 5** . Compared with the variant w/o image GCN, the complete DASE model improves I2T R@1 from 78.2\% to 80.9\% and T2I R@1 from 64.0\% to 65.2\%, with the overall rSum increasing by 4.4\%. This demonstrates that the image-side two-layer GCN fusion module contributes to learning more discriminative visual representations. Compared with w/o text GCN, DASE achieves gains of 2.3\% and 0.6\% on I2T R@1 and T2I R@1, respectively, and improves rSum from 519.4\% to 521.2\%, indicating that the text-side GCN fusion module is also beneficial for cross-modal semantic alignment.

Furthermore, removing the dual associations leads to the most significant performance degradation. Compared with w/o dual associations, DASE improves I2T R@1 by 8.7\% and T2I R@1 by 2.5\%, while increasing rSum by 12.5\%. These results suggest that the dual association mechanism plays a crucial role in capturing fine-grained cross-modal correspondences. Although a few recall values such as I2T R@10 do not reach the highest score among all variants, the complete DASE model achieves the best overall rSum and consistently strong R@1 performance in both retrieval directions, verifying the effectiveness of combining two-layer GCN fusion with dual associations.

3.4 Visualization Results

The efficacy of the DASE framework is qualitatively validated through visualization experiments on cross-modal retrieval outcomes. Specifically, the experiment focuses

on displaying the top-3 retrieved results for both image and text queries as key evidence for evaluating DASE's performance. For clarity, a distinct color-coding scheme is adopted: correct retrievals are marked in green, and incorrect ones in red. As shown in **Fig. 3**, the visualization results demonstrate DASE's superior retrieval capability, which aligns closely with the quantitative findings presented earlier.

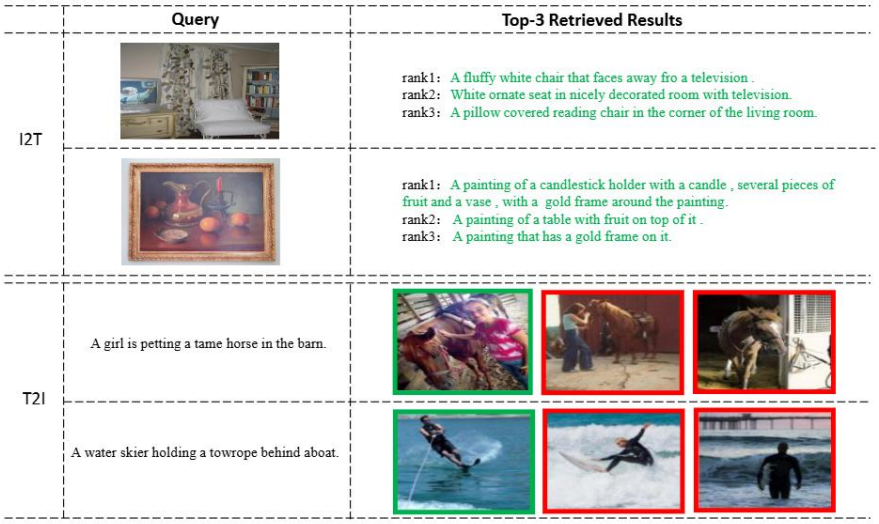

Fig. 3. Visualization of cross-modal alignment is conducted on the MSCOCO datasets.

4 Conclusion

In this work, we present a Dual Associations Semantic Enhancement (DASE) for image-text matching, which effectively addresses the semantic gap between heterogeneous modalities. In contrast to existing methods that largely neglect fine-grained directional associations across modalities, DAAM explicitly models bidirectional image-text interactions, constructing detailed fine-grained cross-modal alignments. Moreover, to achieve full semantic matching across heterogeneous modalities in a unified embedding space, global alignment is incorporated with the local alignment. Additionally, two-layer GCN fusion networks are introduced to capture intricate intra-modal contextual correlations, generating more discriminative feature representations, thus enhancing the accuracy of image-text matching. Extensive experiments validate the effectiveness of DASE in cross-modal retrieval tasks. Overall, our work provides an innovative approach to enhance alignment consistency through dual cross-modal associations modeling, which holds significant implications for advancing image-text matching research.

5 Acknowledgments

This work was supported by the Natural Science Foundation of Shandong Province (No. ZR2023MF031).

References

1. Liu, B., Wang, D., Yang, X., Zhou, Y., Yao, R., Shao, Z., Zhao, J.: Show, deconfound and tell: Image captioning with causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18041–18050 (2022)
2. Wang, Y., Liu, M., Wu, J., Nie, L.: Multi-granularity interaction and integration network for video question answering. IEEE Transactions on Circuits and Systems for Video Technology 33(12), 7684–7695 (2023)
3. Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., Fan, X.: Position Focused Attention Network for Image-Text Matching (2019). https://arxiv.org/abs/1907.09748
4. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216 (2018)
5. Wu, D., Zhang, L., Chen, Y.: Syntactic-guided optimization of image–text matching for intra-modal modeling. The Journal of Supercomputing 81(2), 367 (2025)
6. Ge, X., Chen, F., Xu, S., Tao, F., Jose, J.M.: Cross-modal semantic enhanced interaction for image-sentence retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1022–1031 (2023)
7. Pan, Z., Wu, F., Zhang, B.: Fine-grained image-text matching by cross-modal hard aligning network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19275–19284 (2023)
8. Wang, L., Yang, P., Wang, X., Xu, Z., Dong, Y.: Scene graph fusion and negative sample generation strategy for image-text matching. The Journal of Supercomputing 81(1), 138 (2025)
9. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
10. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV, pp. 18–34. Springer, Berlin, Heidelberg (2020).
11. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10921–10930 (2020)
12. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1218–1226 (2021)
13. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15789–15798 (2021)
14. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15661–15670 (2022)

15. Li, Z., Guo, C., Feng, Z., Hwang, J.-N., Xue, X.: Multi-view visual semantic embedding. In: IJCAI, vol. 2, p. 7 (2022)
16. Wang, Y., Su, Y., Li, W., Xiao, J., Li, X., Liu, A.-A.: Dual-path rare content enhancement network for image and text matching. IEEE Transactions on Circuits and Systems for Video Technology 33(10), 6144–6158 (2023)
17. Z. Pan, F. Wu, and B. Zhang, Fine-grained image-text matching by cross-modal hard aligning network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 1927-519284.
18. Y. Zhang, Z. Ji, D. Wang, Y. Pang, and X. Li, "USER: Unified semantic enhancement with momentum contrast for image-text retrieval," (J), IEEE Trans. Image Process., vol. 33, pp. 595–609, 2024.10.1109/TIP.2023.3348297
19. W. Wei, Z. Gui, C. Wu, A. Zhao, D. Peng, and H. Wu, "Dynamic visual semantic sub-embeddings and fast re-ranking for image-text retrieval," (J), IEEE Trans. Multimed., vol. 27, pp. 3781–3796, 2025.10.1109/TMM.2025.3535373