

– Supplementary Material –
**CODER: Coupled Diversity-Sensitive
Momentum Contrastive Learning
for Image-Text Retrieval**

In this appendix, we provide additional details which were omitted in the main manuscript owing to the limited space. From the perspective of algorithm details, we first give more technical details of representation modules. For instance-level representation, the feature aggregators will be introduced. For concept-level representation, the details of concept selection and the creation of statistical commonsense graph are described. Then, more designed motivation and illustration of our proposed Diversity-sensitive Contrastive Learning (DCL) loss will be given. After that, we also report more experimental results, including the performance of models with different data encoder, influence of different diversity estimation functions, impact of hyper-parameters, data distribution visualization of joint embedding space, performance comparison with different contrastive objectives, and bidirectional image-text retrieval results.

1 Methodology

1.1 Aggregator for Instance-level Representation

For simplicity, here we only describe the image feature aggregator for visual modality, since the same goes for the textual branch. Specifically, we employ the Generalized Pooling Operator proposed in [1], which leverages the encoder-decoder architecture to build the image feature aggregator $g_{vis}(\cdot)$: (1) A positional encoding function that turns position index of local features into a vector. (2) A decoding module that takes the positional encoding output to produce pooling weights.

Position Encoder. To represent each position index l by a dense vector, the positional encoding strategy in Transformer [15] is adopted:

$$\mathbf{p}_l^i = \begin{cases} \sin(u_j, l), & \text{if } i = 2j, \forall i, \\ \cos(u_j, l), & \text{if } i = 2j + 1, \forall i. \end{cases} \quad (1)$$

where $u_j = \frac{1}{10000^{2j/d_p}}$ and d_p denotes the dimension for positional encoding.

Position Decoder. Given the dense vector $\mathbf{p}_l \in \mathbb{R}^{d_p}$, we feed them into a sequence model, which outputs the corresponding pooling weights $\theta = \{\theta\}_{l=1}^L$. The decoder function contains a bidirectional-GRU (BiGRU) and a two-layer perceptron (MLP):

$$\{\mathbf{h}\}_{l=1}^L = BiGRU(\{\mathbf{p}_l\}_{l=1}^L), \theta_k = MLP(\mathbf{h}_l) \quad (2)$$

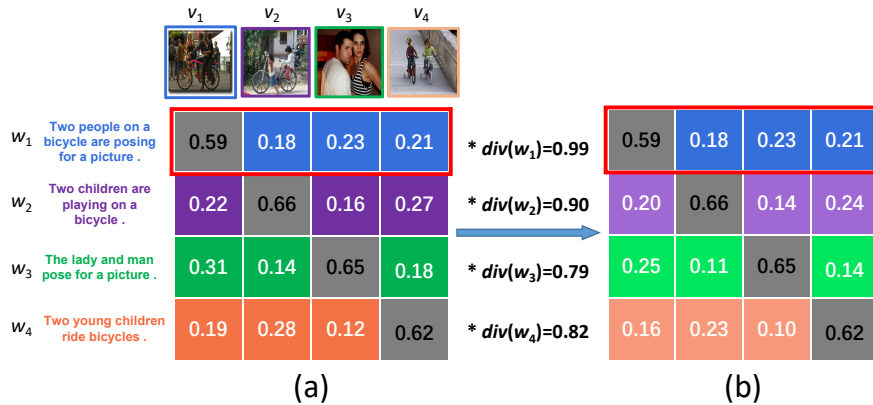


Fig. 1. Conceptual illustration of how diversity affects the cross-modal alignment. Sub-figure (a) depicts the image-text similarity matrix and Sub-figure (b) illustrates the similarity matrix being sensitive to the semantic diversity. After the incorporation of diversity score, our DCL loss will focus on handling the samples with high diversity.

where \mathbf{h}_l is the hidden states output by Bi-GRU at the position index k . Then, we can aggregate the local image features into a holistic instance-level image representation \mathbf{v}^I :

$$\mathbf{v}^I = g_{vis}(\{\mathbf{o}_l\}_{n=1}^L) = \sum_{l=1}^L \theta_l * \mathbf{o}_l \quad (3)$$

Similarly, we can obtain the textual feature aggregator $g_{text}(\cdot)$ and global instance-level text representation \mathbf{w}^I .

1.2 Concept-level Representation

Concept Initialization Our statistical commonsense knowledge is extracted from certain meaningful concepts and their semantic correlations, which are collected from the texts of the whole image-caption dataset. In order to filter out most meaningless and infrequent concepts, we follow [5, 7, 16] to select the representative words with top- q appearing frequencies in the concept vocabulary, which are roughly categorized into three types, *i.e.*, *Object*, *Motion*, and *Property*. Then, following [16], according to the appearance frequency over dataset, the ratio of the concepts with type of (*Object*, *Motion*, *Property*) is set to (7:2:1). Afterwards, we adopt the glove [13] to initialize them and denote them as \mathbf{X} .

Commonsense Aided Concept Representation. To model the statistical commonsense knowledge, we follow [16] to utilize the co-occurrence relationship between concepts to build one correlation graph. To be more specific, we construct a conditional probability matrix \mathbf{P} to model the relation between different

concepts, with each element \mathbf{P}_{ij} denoting the appearance probability of concept C_i when concept C_j appears: $\mathbf{P}_{ij} = \mathbf{B}_{ij}/N_i$, where $\mathbf{B} \in \mathbb{R}^{q \times q}$ is the concept co-occurrence matrix, \mathbf{B}_{ij} represents the co-occurrence times of C_i and C_j , and N_i is the appearance times of C_i in the corpus.

Afterwards, to further prevent the correlation matrix from being over-fitted and improve its generalization ability, we follow [3, 16] to apply binary operation to the rescaled matrix \mathbf{P} :

$$\mathbf{H}_{ij}^{sc} = \begin{cases} 0, & \text{if } \mathbf{P}_{ij} < \epsilon, \\ 1, & \text{if } \mathbf{P}_{ij} \geq \epsilon, \end{cases} \quad (4)$$

where ϵ denotes a threshold parameter filters noisy edges. Given the LCC representations \mathbf{X}^l and statistical commonsense graph \mathbf{H}^{ss} , we employ one Graph Convolution Network (GCN) [8] to process them, after one-layer convolution operation, the statistical commonsense aided concept (SCC) representations can be computed as:

$$\mathbf{Y} = \rho(\tilde{\mathbf{A}}_{sc} \mathbf{X}^l \mathbf{W}_{sc}) \quad (5)$$

where $\tilde{\mathbf{A}}_{sc} = \mathbf{D}_{ss}^{-\frac{1}{2}} \mathbf{H}^{sc} \mathbf{D}_{sc}^{-\frac{1}{2}} + \mathbf{I}$ denotes the normalized symmetric matrix and \mathbf{W}_{sc} is the learnable weight matrix.

Commonsense Aided Concept-level Representation. To generate concept-level representations, we generate representations (\mathbf{v}_C^q and \mathbf{w}_C^q) by using another group of feature aggregators ($g_{vis}(\cdot)$ and $g_{text}(\cdot)$) to combine local features $\{\mathbf{o}_l\}_{l=1}^L$ and $\{\mathbf{e}_t\}_{t=1}^T$, respectively. Note that the weights of both visual feature aggregators for \mathbf{v}_I and \mathbf{v}_C^q are shared, and we empirically find this operation helps to make our method converge better. Afterwards, \mathbf{v}_C^q and \mathbf{w}_C^q are taken as input vectors to query from the SCC representations \mathbf{Y} . As consequence, the output scores for different concepts allow us to uniformly utilize the linear combination of the SCC representations to represent both modalities. Mathematically, the concept-level representation \mathbf{v}^C and \mathbf{w}^C can be calculated as:

$$\begin{aligned} \mathbf{v}^C &= \sum_{i=1}^g a_i^v \mathbf{y}_i; \quad a_i^v = \frac{e^{\lambda \mathbf{v}_C^q \mathbf{W}^v \mathbf{y}_i^T}}{\sum_{i=1}^g e^{\lambda \mathbf{v}_C^q \mathbf{W}^v \mathbf{y}_i^T}}. \\ \mathbf{w}^C &= \sum_{j=1}^g a_j^w \mathbf{y}_j; \quad a_j^w = \frac{e^{\lambda \mathbf{w}_C^q \mathbf{W}^w \mathbf{y}_j^T}}{\sum_{j=1}^g e^{\lambda \mathbf{w}_C^q \mathbf{W}^w \mathbf{y}_j^T}} \end{aligned} \quad (6)$$

where $\mathbf{W}^v \in \mathbb{R}^{F \times F}$ and $\mathbf{W}^w \in \mathbb{R}^{F \times F}$ denote the learnable parameter matrix, \mathbf{a}_i^v and \mathbf{a}_j^w denote the visual and textual score corresponding to the concept \mathbf{z}_i , respectively. λ controls the smoothness of the softmax function.

1.3 Illustration of Diversity in DCL

In this section, we describe how our proposed semantic *diversity* affects the cross-modal alignment. First, we briefly review the mathematical definitions of

diversity and DCL loss defined in the main manuscript. Specifically, we take as example that visual feature \mathbf{v}_i is an anchor sample and Q text features $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_Q\}$ are to be compared (among which only \mathbf{w}_i is a matching sample for \mathbf{v}_i), to illustrate how we estimate diversity of an anchor sample. The cosine similarity of $\text{cosine}(\mathbf{v}_i, \mathbf{w}_j)$ is defined as S_{ij} . Then, the diversity of anchor \mathbf{v}_i (\mathbf{w}_i) is defined as:

$$\begin{aligned} SD(i) &= \sqrt{E(S_{ij}) - [E(S_{ij})]^2}, i \neq j, j \in [1, Q]; \\ div_{std}(\mathbf{v}_i) &= 1/\sigma(\epsilon/SD(\mathbf{v}_i)); \\ div_{std}(\mathbf{v}_i) &= div_{std}(\mathbf{v}_i) / \max\{div_{std}(\mathbf{v}_1), \dots, div_{std}(\mathbf{v}_Q)\}, \end{aligned} \quad (7)$$

where $E(\cdot)$ is the mathematical expectation function and $\sigma(\cdot)$ denotes the Sigmoid function that normalizes the reciprocal of SD value to a uniform scale. $div_{std}(\mathbf{v}_i)$ denotes the diversity score of \mathbf{v}_i calculated from the candidate textual samples to be compared with. $\epsilon = 0.1$ is a tuning parameter. Lastly, we divide each diversity score $div_{std}(\mathbf{v}_i)$ by the maximum value of them in mini-batch for normalization. Similarly, the diversity of \mathbf{w}_j can be obtained.

Except for the definition above, we also explore another method to define diversity, which is built based on employing *information entropy*. It is commonly used to measure the information volume conveyed by variables. To incorporate cross-modal similarity into information entropy computation, we first utilize soft-max function to convert similarity score to probability form. Then, we can use information entropy to estimate the diversity of anchor sample. Formally, the information entropy based diversity of anchor \mathbf{v}_i (\mathbf{w}_i) is defined as:

$$\begin{aligned} P_{ij} &= \frac{e^{S_{ij}}}{\sum_{j=1}^Q e^{S_{ij}}}; \\ H(\mathbf{v}_i) &= -P_{ij} \cdot \sum_{j=1}^Q \log_2(P_{ij}), i \neq j; \\ div_{ent}(i) &= 1/\sigma(\epsilon/H(i)); \\ div_{ent}(\mathbf{v}_i) &= div_{ent}(\mathbf{v}_i) / \max\{div_{ent}(\mathbf{v}_1), \dots, div_{ent}(\mathbf{v}_Q)\}, \end{aligned} \quad (8)$$

where $H(\cdot)$ represents the function for calculating information entropy. $div_{ent}(\mathbf{v}_i)$ denotes the information entropy based diversity score of \mathbf{v}_i calculated from the candidate textual samples to be compared with. The effect comparison between two types of diversity will be presented in Section 2.2.

Furthermore, given $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ and $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_Q\}$, based on the diversity score defined in Eq.7, our proposed DCL loss L_{DCL} is defined as:

$$\begin{aligned} L_{DCL}(\mathbf{V}, \mathbf{W}) &= l_{DCL}(\mathbf{W}, \mathbf{V}) + l_{DCL}(\mathbf{V}, \mathbf{W}) \\ l_{DCL}(\mathbf{V}, \mathbf{W}) &= \frac{\mu}{N} \sum_{n=1}^N [\log(\sum_{q \neq n} \exp(\frac{(S_{nq} - \gamma)}{\mu \cdot div_{std}(\mathbf{v}_n)}) + 1) - \log(S_{nn} + 1)]; \\ l_{DCL}(\mathbf{W}, \mathbf{V}) &= \frac{\mu}{Q} \sum_{q=1}^Q [\log(\sum_{n \neq q} \exp(\frac{(S_{qn} - \gamma)}{\mu \cdot div_{std}(\mathbf{w}_q)}) + 1) - \log(S_{qq} + 1)]; \end{aligned} \quad (9)$$

where $div_{std}(\mathbf{v}_n)$ and $div_{std}(\mathbf{w}_q)$ denotes diversity of \mathbf{v}_n and \mathbf{w}_q , respectively and they are used to adaptively weight each negative sample.

Then, we take the similarity measuring matrix of text-to-image as example. As shown in Figure 1(a), for query sample \mathbf{w}_1 with higher diversity, *i.e.* semantic ambiguity, the calculated similarity difference between it and other three mismatched samples is very small. According to Eq. 7, the diversity of \mathbf{w}_1 is larger than those of others. As a result, seeing Figure 1(b), the similarity ratio between positive pair and negative pairs of \mathbf{w}_1 remain unchanged. By contrast, those of other samples all become larger than before. From Eq. 9, we can know that this adaptive weighting strategy will lead to imposing harder punishment on sample \mathbf{w}_1 than others. And it is consistent with our designing principle.

Note that distinct from previous works [2, 4] that focus on mining hard negatives specifically for a single sample, the diversity in DCL is defined from a more holistic view, which is calculated based on the statistical information of data distribution. Consequently, our DCL loss aims at reducing the cross-modal distribution discrepancy, which captures more hierarchical semantic structure in joint space by alleviating the negative influence brought by samples with high diversity. Furthermore, we will display how the diversity in DCL loss affects data distribution in the joint embedding space by t-SNE visualizing in Section 2.4.

2 Experiments

2.1 More Results and Comparisons for Image-Text Retrieval

The additional experimental results are presented in Table 1. Note that the results of [1] are reported by our replicated number. Since the instance-level representation part of our model is built according to [1], a solid performance of baseline is needed to reasonably evaluate the impact of our contributions. Thus, we report our replicated results by using their open-sourced code with no change, and mark them with \star symbol in Table 1 & 2. To further assure the fairness of comparisons, we divide the experiments into two groups. One group of approaches adopt “Faster-RCNN + BiGRU” as image and text encoders, meanwhile another group of methods is uniformly built based on “Faster-RCNN + BERT” as encoders. The experimental results on Flickr30K test set are presented in Table 1. First, in contrast to other methods adopting “Faster-RCNN + BiGRU” architecture, the “R@sum” achieved by our CODER surpasses the second best performance by 13.6%. Secondly, compared with those employing “Faster-RCNN + BERT” for encoding multi-modal data, our method outperforms the best competitor by 13.7% on the “R@sum” metric.

As shown in Table 1, on MSCOCO 1k test set, our CODER also significantly outperforms all other compared methods. Although employing the “Faster-RCNN + BiGRU” as image and text encoders, there is still a performance gap between CODER and best competitor SMFEA [6] on the R@sum metric, *e.g.* 4.0% improvement. Moreover, the retrieval performances on MSCOCO 5K test set are listed in Table 2. Comparing with best competitor GPO (BERT) [1], our CODER

Table 1. Comparisons of experimental results on MSCOCO 1K test set and Flickr30k test set, employing different image and text encoders (denoted by bold section title).

Methods	Image Encoder	MSCOCO 1K					Flickr30K								
		Text Retrieval			Image Retrieval		R@sum	Text Retrieval			Image Retrieval		R@sum		
		R@1	R@5	R@10	R@1	R@5		R@10	R@1	R@5	R@10	R@1		R@5	R@10
Faster-RCNN + BiGRU															
SCAN [9] (2018)	Faster-RCNN	72.7	94.8	98.4	58.8	88.4	94.8	507.9	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN [10] (2019)	Faster-RCNN	76.2	94.8	98.2	62.8	89.7	95.1	516.8	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CVSE [16] (2020)	Faster-RCNN	74.8	95.1	98.3	59.9	89.4	95.2	512.7	73.5	92.1	95.8	52.9	80.4	87.8	482.5
MMCA [18] (2020)	Faster-RCNN	74.8	95.6	97.7	61.6	89.8	95.2	514.7	74.2	92.8	96.4	54.8	81.4	87.8	487.4
GSMN [11] (2020)	Faster-RCNN	76.1	95.6	98.3	60.4	88.7	95.0	514.0	74.4	91.5	95.3	54.1	79.9	86.6	481.8
SMFEA [6] (2021)	Faster-RCNN	75.1	95.4	98.3	62.5	90.1	96.2	517.6	73.7	92.5	96.1	54.7	82.1	88.4	487.5
WGL [17] (2021)	Faster-RCNN	75.4	95.5	98.6	60.8	89.3	95.3	514.9	74.8	93.3	96.8	54.8	80.6	87.5	487.8
GPO (BiGRU) [1] (2021) *	Faster-RCNN	76.2	95.4	98.5	60.1	89.8	95.2	515.2	74.8	93.5	97.0	55.1	83.8	89.4	493.6
CODER (BiGRU)	Faster-RCNN	78.9	95.6	98.6	62.5	90.3	95.7	521.6	79.4	94.9	97.7	59.0	85.2	91.0	507.2
Faster-RCNN + BERT															
DSRAN [19] (2021)	Faster-RCNN	77.1	95.3	98.1	62.9	89.9	95.3	518.6	75.3	94.4	97.6	57.3	84.8	90.9	500.3
GPO (BERT) [1] (2021) *	Faster-RCNN	78.6	96.2	98.7	62.9	90.8	96.1	523.3	78.1	94.1	97.8	57.4	84.5	90.4	502.3
DIME (i-t) [14] (2021)	Faster-RCNN	77.9	95.9	98.3	63.0	90.5	96.2	521.8	77.4	95.0	97.4	60.1	85.5	91.8	507.2
CODER (BERT)	Faster-RCNN	82.1	96.6	98.8	65.5	91.5	96.2	530.6	83.2	96.5	98.0	63.1	87.1	93.0	520.9

Table 2. Comparisons of experimental results on MSCOCO 5K test set, employing different image and text encoders (denoted by bold section title).

Methods	Image Encoder	MSCOCO 5K						
		Text retrieval			Image Retrieval		R@sum	
		R@1	R@5	R@10	R@1	R@5		R@10
Faster-RCNN + BiGRU								
SCAN [9] (2018)	Faster-RCNN	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN [10] (2019)	Faster-RCNN	53.0	81.1	89.4	40.5	70.6	81.1	415.7
MMCA [18] (2020)	Faster-RCNN	54.0	82.5	90.7	38.7	69.7	80.8	416.4
SMFEA [6] (2021)	Faster-RCNN	54.2	-	89.9	41.9	-	83.7	-
GPO (BiGRU) [1] (2021) *	Faster-RCNN	55.2	83.1	91.0	39.3	69.9	81.1	419.6
CODER (BiGRU)	Faster-RCNN	58.5	84.3	91.5	40.9	70.8	81.4	427.2
Faster-RCNN + BERT								
DSRAN [19] (2021)	Faster-RCNN	53.7	82.1	89.9	40.3	70.9	81.3	418.2
DIME (i-t) [14] (2021)	Faster-RCNN	56.1	83.2	91.1	40.2	70.7	81.4	422.7
GPO (BERT) [1] (2021) *	Faster-RCNN	57.3	84.5	91.6	41.1	71.9	82.6	429.0
CODER (BERT)	Faster-RCNN	62.6	86.6	93.1	42.5	73.1	83.3	441.3

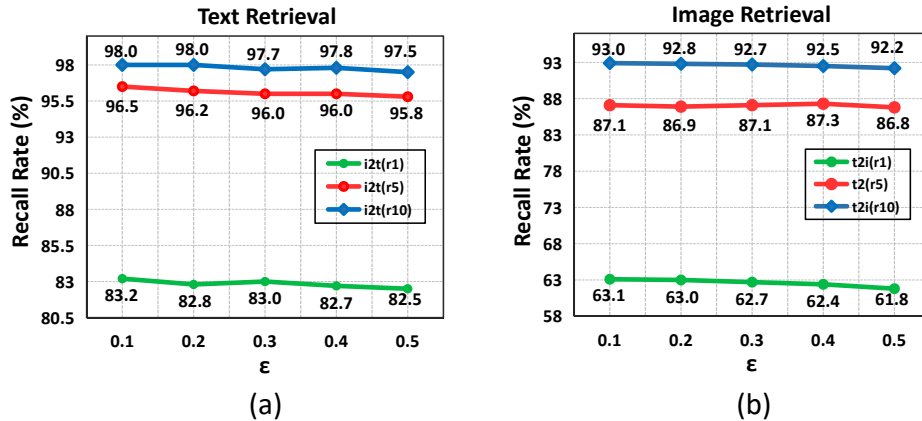
outperforms it by 5.3% improvement for text retrieval and 1.4% for image retrieval on R@1 criteria. The above results are obtained under totally fair conditions with the same data encoders, thus they can solidly validate the superiority of our method for image-text retrieval.

2.2 Impact of Different Functions for Diversity Estimation

In this section, we explore the effect of different diversity estimation functions. As shown in Table 3, the experimental results based on estimation functions of $div_{std}(\cdot)$ and $div_{ent}(\cdot)$ are listed. For comparison, the results without diversity

Table 3. Impact of different diversity estimation functions in DCL loss on Flickr30K test set. Explicit diversity estimation is abbreviated as “EE”.

EE	Diversity Estimation Function	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
×	-	81.5	95.8	98.1	61.3	85.7	91.0
✓	$div_{std}(\cdot)$	83.2	96.5	98.0	63.1	87.1	93.0
✓	$div_{ent}(\cdot)$	83.0	96.2	98.1	62.4	86.9	92.5

**Fig. 2.** Impact of varied controlling parameters ϵ on Flickr30K test set. Sub-figure (a) shows image-to-text retrieval performance with different values of ϵ in DCL loss. Sub-figure (b) depicts the corresponding text-to-image retrieval performance.

estimating are also presented. From Table 3, we can see our proposed two types of diversity estimation functions can both bring about substantial performance boost. It further validates our train of thought for diversity estimation is reasonable. Besides, the performance of model using $div_{ent}(\cdot)$ is slightly inferior to that with $div_{std}(\cdot)$. The potential reason may be that the softmax function adopted in $div_{ent}(\cdot)$ will made the original data distribution of cross-modal similarity to be more smooth.

2.3 Hyper-Parameter Analysis for Diversity Estimation

In this part, we investigate the affect of controlling parameter ϵ of diversity in Eq.7 on retrieval performance. As shown in Figure 2, with the variant ϵ , the retrieval results vary moderately, indicating our model is robust to ϵ within a proper range. Additionally, the increase of ϵ value implicates the narrower variation range of diversity score. Thus, from Figure 2, we can infer the proper sensitiveness of DCL loss on parameter ϵ also leads to performance gain. Overall,

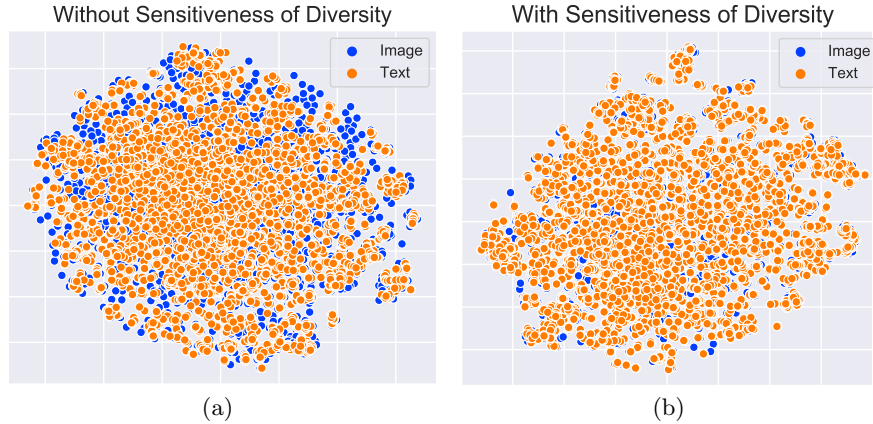


Fig. 3. T-SNE visualization of the image-text representations generated by (a) baseline model with L_{DCLI} loss and (b) full CODER model on Flickr30K test set (1000 images and 5000 texts).

these results reveal that diversity plays critical role in DCL loss for learning more discriminative cross-modal embeddings.

2.4 T-SNE Visualization of Cross-Modal Representation

To better understand how our DCL loss affects the cross-modal joint embedding space, we utilize t-SNE [12] to visualize the learned representations from Flickr30K test set, including 1000 images and 5000 texts. Specifically, Figure 3(a) displays the feature distribution of the baseline model (referring to the model #1 in Table 4 defined in the main manuscript, employing the L_{DCLI} loss as learning objective), and those of full CODER model is illustrated in Figure 3(b). We can see that the data distribution in Figure 3(b) is obviously more desirable than that in Figure 3(a), which lies in two main points: 1) The distribution discrepancy between images and texts is alleviated significantly. 2) The learned joint space is characterized by being structured and hierarchical rather than being irregular and scattered. Considering the unique difference between both models is the varied configuration of DCL loss, we believe the main factor improving the data distribution is the combination between coupled memory banks ($L_{M,DCL}^I$) and diversity estimation. Benefiting from the large-scale negative interactions from the former, we can achieve more accurate diversity estimation for DCL. It is able to regularize the joint embedding space by alleviating the influence of sample with high diversity, such as some visual instances existing on the left side of Figure 3(a).





























	Query	Baseline	CODER
Flickr30k		<ol style="list-style-type: none"> 1. A pride softball team member hits the ball and runs towards first base while the umpire and catcher watch the ball. 2. A girl dressed in a red uniform is hitting a softball with a bat while a catcher and home plate umpire look on. 3. A woman runs after making a hit in women's softball, the catcher rises to her feet. 4. A baseball catcher trying to tag a base runner in a baseball game. 	<ol style="list-style-type: none"> 1. A pride softball team member hits the ball and runs towards first base while the umpire and catcher watch the ball. 2. A woman runs after making a hit in women's softball, the catcher rises to her feet. 3. Girl hits a ball and the catcher looks on. 4. A girl dressed in a red uniform is hitting a softball with a bat while a catcher and home plate umpire look on.
		<ol style="list-style-type: none"> 1. Three male field hockey players are running onto the field while the goalie is standing in the goal looking on. 2. The three field hockey players dressed in orange make for the ball. 3. A group of guys are playing roller hockey. 4. A large goalie towers over his opposing teammates. 	<ol style="list-style-type: none"> 1. Three male field hockey players are running onto the field while the goalie is standing in the goal looking on. 2. The three field hockey players dressed in orange make for the ball. 3. A team in orange uniforms are near a goal and a goalkeeper in green. 4. A large goalie towers over his opposing teammates.
	Two large dogs chase after another dog that has a ball in his mouth and runs from them.	  	  
	A bicyclist near town is racing in a race while wearing yellow and a helmet.	  	  
MSCOCO		<ol style="list-style-type: none"> 1. A man holding a book and a phone in his hands. 2. person lying on ground reading book and holding cell phone. 3. Person laying on ground looking at book and phone. 4. A person holding up a smart phone in a public space. 	<ol style="list-style-type: none"> 1. A man holding a book and a phone in his hands. 2. person lying on ground reading book and holding cell phone. 3. A person laying down with a book in one hand and a cell phone in another. 4. Person laying on ground looking at book and phone.
		<ol style="list-style-type: none"> 1. The two giraffes are walking in their pen. 2. Two giraffes in a grassy area with a fence and trees next to them. 3. A couple of giraffes that are standing in a fence. 4. Two giraffes standing in a brush covered area. 	<ol style="list-style-type: none"> 1. Two giraffes in a grassy area with a fence and trees next to them. 2. The two giraffes are walking in their pen. 3. Two giraffes roaming around an enclosed area on a sunny day. 4. A couple of giraffes that are standing in a fence.
	A tray topped with two sandwiches, pie and a plate of coleslaw.	  	  
	A row of motorcycles parked in front of a building.	  	  

Fig. 4. The qualitative bi-directional retrieval results on Flickr30k and MSCOCO datasets. For text retrieval, the ground-truth and non ground-truth describing sentences are marked in red and black, respectively. For image retrieval, the number in the upper left corner denotes the ranking order, and the ground-truth images are annotated with green check mark.

2.5 Retrieval Result Visualization

To further validate the effectiveness of our method, in Figure 4, we choose several images and texts as queries and exhibit their retrieval results. Note that we take CODER adopting BTR loss [4] instead of our DCL loss as baseline model. As shown in Figure 4, comparing with baseline, the CODER model with the aid of DCL loss is able to return better image-text retrieval results.

References

1. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: CVPR (2021) 1, 5, 6
2. Chen, T., Deng, J., Luo, J.: Adaptive offline quintuplet loss for image-text matching. In: ECCV (2020) 5
3. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: CVPR (2019) 3
4. Faghri, F., Fleet, D.J., Kiros, J., Fidler, S.: Vse++: improved visual-semantic embeddings. In: BMVC (2018) 5, 9

5. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR (2015) [2](#)
6. Ge, X., Chen, F., Jose, J.M., Ji, Z., Wu, Z., Liu, X.: Structured multi-modal feature embedding and alignment for image-sentence retrieval. ACMMM (2021) [5](#), [6](#)
7. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: CVPR (2018) [2](#)
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2016) [3](#)
9. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018) [6](#)
10. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: ICCV (2019) [6](#)
11. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: CVPR (2020) [6](#)
12. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2605 (2008) [8](#)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014) [2](#)
14. Qu, L., Liu, M., Wu, J., Gao, Z., Nie, L.: Dynamic modality interaction modeling for image-text retrieval. In: SIGIR (2021) [6](#)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) [1](#)
16. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: ECCV (2020) [2](#), [3](#), [6](#)
17. Wang, Y., Zhang, T., Zhang, X., Cui, Z., Huang, Y., Shen, P., Li, S., Yang, J.: Wasserstein coupled graph learning for cross-modal retrieval. In: ICCV (2021) [6](#)
18. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR (2020) [6](#)
19. Wen, K., Gu, X., Cheng, Q.: Learning dual semantic relations with graph attention for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology **31**, 2866–2879 (2021) [6](#)