AGREE: Aligning Cross-Modal Entities for Image-Text Retrieval Upon Vision-Language Pre-trained Models

Xiaodan Wang
School of Computer Science,
Fudan University
Shanghai, China
xiaodanwang20@fudan.edu.cn

Lei Li
School of Data Science and
Engineering, East China Normal
University
Shanghai, China
leili@stu.ecnu.edu.cn

Zhixu Li[*]
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University
Shanghai, China
zhixuli@fudan.edu.cn

Xuwu Wang
Xiangru Zhu
School of Computer Science,
Fudan University
Shanghai, China
{xwwang18,xrzhu19}@fudan.edu.cn

Chengyu Wang
Jun Huang
Alibaba Group
Hangzhou, Zhejiang, China
{chengyu.wcy,huangjun.hj}@alibaba-inc.com

Yanghua Xiao[*]
Shanghai Key Laboratory of Data
Science, School of Computer Science,
Fudan University
Shanghai, China
shawyh@fudan.edu.cn

ABSTRACT

Image-text retrieval is a challenging cross-modal task that arouses much attention. While the traditional methods cannot break down the barriers between different modalities, Vision-Language Pre-trained (VLP) models greatly improve image-text retrieval performance based on massive image-text pairs. Nonetheless, the VLP-based methods are still prone to produce retrieval results that cannot be cross-modal aligned with entities. Recent efforts try to fix this problem at the pre-training stage, which is not only expensive but also unpractical due to the unavailable of full datasets. In this paper, we novelly propose a lightweight and practical approach to align cross-modal entities for image-text retrieval upon VLP models only at the fine-tuning and re-ranking stages. We employ external knowledge and tools to construct extra fine-grained image-text pairs, and then emphasize cross-modal entity alignment through contrastive learning and entity-level mask modeling in fine-tuning. Besides, two re-ranking strategies are proposed, including one specially designed for zero-shot scenarios. Extensive experiments with several VLP models on multiple Chinese and English datasets show that our approach achieves state-of-the-art results in nearly all settings.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; *Retrieval models and ranking.*

KEYWORDS

Image-Text Retrieval, Vision-Language Pre-training, VLP

ACM Reference Format:
Xiaodan Wang, Lei Li, Zhixu Li[*], Xuwu Wang, Xiangru Zhu, Chengyu Wang, Jun Huang, and Yanghua Xiao[*]. 2023. AGREE: Aligning Cross-Modal Entities for Image-Text Retrieval Upon Vision-Language Pre-trained Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539597.3570481

1 INTRODUCTION

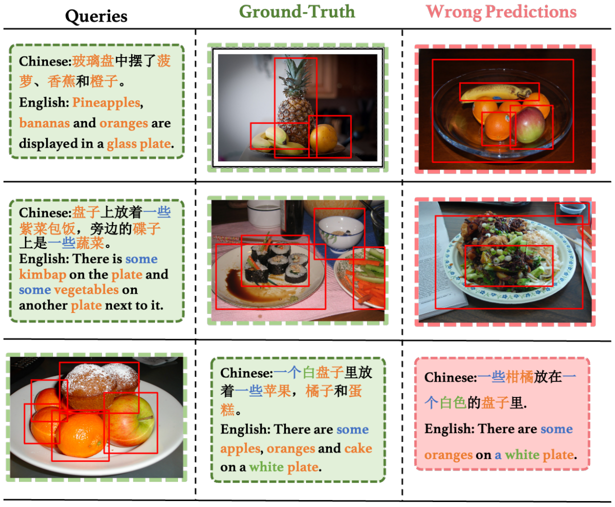
Image-text retrieval is a challenging cross-modal task, which requires retrieving semantically correlated text (or image) samples for a given image (or text) sample from a predefined database [7, 17]. With the development of social media, image-text retrieval becomes a more and more important and hot research topic in several communities including multimedia [8], computer vision [40], and natural language processing [5].

The key challenge in image-text retrieval lies in how to do representation learning for image and text data, and then measure the cross-modal similarity between their representations. To address the challenge, one line of traditional methods tend to encode samples of different modalities in a unified embedding space [25], while the other line of traditional efforts prefer to encode images and texts separately and then compute the distance between images and texts through metric learning [5, 7, 45]. However, without sufficient training data, traditional methods cannot break down the barriers between the representation learning of different modalities, which also limits the learning on cross-modal matching mechanisms. Recently, Vision-Language Pre-trained (VLP) models [1, 3, 30] show increasing power in greatly improving the performance of many cross-modal tasks under either zero-shot or fine-tuning scenarios. Based on massive image-text pairs, the VLP models learn more adequate cross-modal association information through various self-supervised pre-training tasks [28, 31] and the Transformer structure [11, 12], thus greatly alleviating the shortcomings of traditional image-text retrieval methods.

[*]Corresponding Authors. Yanghua Xiao is also affiliated with Fudan-Aishu Cognitive Intelligence Joint Research Center, and a member of AI Communication at Institute for Global Communications and Integrated Media, Fudan University.
Our code will be released to the EasyNLP at https://github.com/alibaba/EasyNLP.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570481



Chinese:玻璃盘中摆了菠萝、香蕉和橙子。
English: Pineapples, bananas and oranges are displayed in a glass plate.

Chinese:盘子上放着一些紫菜包饭，旁边的碟子上是一些蔬菜。
English: There is some kimbap on the plate and some vegetables on another plate next to it.

Chinese:一个白盘子里放着一些苹果，橘子和蛋糕。
English: There are some apples, oranges and cake on a white plate.

Chinese:一些柑橘放在一个白色的盘子里.
English: There are some oranges on a white plate.

Figure 1: Three mismatched image-text retrieval cases of Wukong$_{ViT-L}$ [11] fine-tuned on COCO-CN [29]

However, the current VLP models have not yet realized *fine-grained cross-modal semantic matching* mechanism. Although some fine-grained cross-modal association, like image object-token correspondence [1, 27, 30] and region patch-token correspondence [46], are explored, VLP-based image-text retrieval models may still output incorrect retrieval results where entities cannot be aligned between the query data and the retrieved data. See the three mismatched cases in Figure 1: in the first case, the "pineapple" in the query does not appear in the predicted image. Similarly, in the second case, the model only pays attention to the matching of "vegetables" and "plates", but ignores another important entity "kimbap" in the query. Also, it has a misjudgment about the number of plates. In the third case, the predicted text does not contain "apple" and "cake" which can be apparently observed in the query image.

Recent efforts [3, 30, 48] try to fix this problem by emphasizing the correspondence of cross-modal entities (i.e., the visual entities in images and the entity mentions in texts) in the pre-training stage of VLP. Some work [3, 48] constructs scene graphs from images and texts separately and then align them, while some other works leverage external knowledge such as external object detector [30] or multi-lingual datasets [49] for improving fine-grained cross-modal matching. Despite their success, the improvement in the pre-training stage of VLP requires to re-train the large models with extremely high computational costs. Moreover, the full image-text pair datasets are usually not publicly available, which further decreases the practical values of these approaches.

In this paper, we novelly propose a lightweight and practical approach, **AGREE**, to AliGn cRoss-modal EntitiEs for image-text retrieval upon VLP models ONLY at the fine-tuning and re-ranking stages. Particularly, several optimization methods are designed and adopted to enhance cross-modal entity alignment for both fine-tuning and zero-shot scenarios. In the fine-tuning step, we first obtain visual entity-image pairs from the external knowledge base Visual Genome [19], which can then be used to learn the alignment between visual entities and their corresponding images through contrastive learning and image region mask modeling. Secondly, we construct a sentence only with textual entities and their visualizable properties (such as color and number) contained in each text, and then learn the alignment between the sentence and its corresponding image through contrastive learning and textual entity mask modeling. Last but not the least, we emphasize the importance of cross-modal entity alignment by randomly masking entities either in the image or in the text to let the model be more sensitive to the missing of aligned entities across modalities. In the re-ranking step, we take the top-k (e.g. k=10) retrieval results to do reverse image-text retrieval, whose results are then taken into account for re-ranking. Specifically, for the zero-shot scenarios, which do not have the fine-tuning step, we also take the top-k retrieval results to calculate the similarity between the entities from images and texts, which are also considered in re-ranking. Our experiments demonstrate that the proposed methods benefit both fine-tuning scenarios and zero-shot scenarios.

The main contributions of this paper are threefold:

- We novelly propose a lightweight and practical approach to align cross-modal entities for image-text retrieval upon VLP models only at the fine-tuning and re-ranking stages.
- We employ external knowledge and tools to construct fine-grained vision-text pairs, and then emphasize cross-modal entity alignment through contrastive learning and entity-level mask modeling for fine-tuning. Besides, two strategies are designed for re-ranking, including one specially designed for zero-shot.
- Extensive experiments with several VLP models on multiple Chinese and English datasets show that our approach achieves state-of-the-art results in nearly all settings.

2 RELATED WORK

Image-text retrieval attracts growing attention in recent years [5, 7, 8, 17, 40]. In the following, we first cover traditional methods without using Vision-Language Pre-trained (VLP) models [1, 3, 30], and then introduce the mainstream methods based on VLPs.

Traditional Image-Text Retrieval. Traditional image-text retrieval methods usually learn a shared embedding space to directly compare features of different modalities [21, 25], or to learn an objective in the embedding space, enabling the distance of matched pairs closer and mismatched pairs far away [5, 7, 45]. Some work also combines both of the advantages through knowledge distillation [22], to transfer fine-grained alignment learning from cross-attention to dual-encoder fast retrieval model [10], balancing the retrieval efficiency and accuracy.

For fine-grained local alignment on the text side, a general approach is to exploit a multi-label detector to detect semantic concepts, and then fuse these concepts with the global representation of the image [13], or construct scene graphs to utilize consensus based on the concept co-occurrence relationship [41]. Another type of frequently-used method pays more attention to local image features, by extracting region features for fine-grained vision-semantic matching [17]. Such methods generally treat all fine-grained information of images (region) and texts (word) in a unified manner and obtain a correlation score of fine-grained cross-modal information [21] with the help of attention mechanism, which often requires complex cross-modal interaction. Another disadvantage brought by this kind of approach lies in the limitations of predefined labels, and the dataset-based co-occurrence statistics may also introduce strong inductive bias.

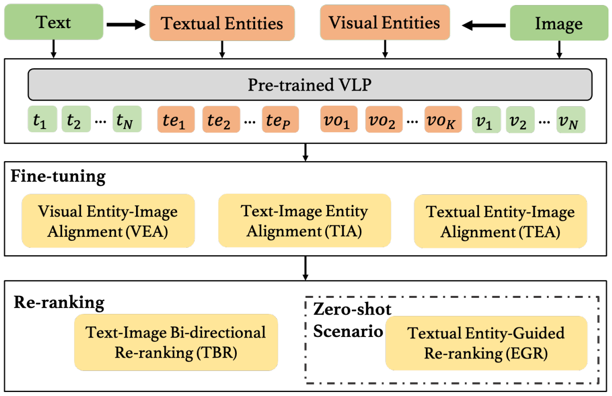

Text
Textual Entities
Visual Entities
Image
Pre-trained VLP
t_1 t_2 ··· t_N te_1 te_2 ··· te_P vo_1 vo_2 ··· vo_K v_1 v_2 ··· v_N
Fine-tuning
Visual Entity-Image Alignment (VEA)
Text-Image Entity Alignment (TIA)
Textual Entity-Image Alignment (TEA)
Re-ranking
Zero-shot Scenario
Text-Image Bi-directional Re-ranking (TBR)
Textual Entity-Guided Re-ranking (EGR)

Figure 2: Overview framework of AGREE

VLP-based Image-Text Retrieval. The emergence of VLP models greatly improves the performance of image-text retrieval [9, 11, 36]. While single-stream VLP models [1, 23] concatenate images and texts as one input, dual-stream VLP models [24, 36, 39, 46] project the representations of images and texts into a unified embedding space by contrastive learning.

To focus more on fine-grained cross-modal alignment, some work uses patch-token late-similarity calculation [46], or utilizes region-level image information [9, 27, 30]. Some other work incorporates prior knowledge to enhance pretraining [3, 30, 48, 49]. For instance, [48] finds a novel way to analyze the textual syntactic structure and introduce scene graphs to fuse with the image's local features, and [3] integrates the cross-modal and intra-modal knowledge simultaneously in a unified scene graph, all of which commit to fine-grained interaction and knowledge injection during the pretraining stage. Despite their success, the improvements in the pre-training stage are quite expensive and usually impractical for those who do not have the complete training dataset. Thus, this paper innovatively proposes a lightweight and practical approach to align cross-modal entities for image-text retrieval upon VLP models only at the fine-tuning and re-ranking stages.

3 METHODOLOGY

This section first gives a formal definition to the image-text retrieval task, and then presents the overview framework of our approach, followed with the proposed cross-modal entity alignment methods adopted in the fine-tuning and re-ranking stages respectively.

3.1 Task Formulation

Given a collection of image-text pairs in $D = \{V, T\}^N$, where $V = \{v_m\}_{m=1}^{N}$ and $T = \{t_m\}_{m=1}^{N}$ are the sets of images and texts respectively, and each image $v \in V$ is associated with one or several $t \in T$, denoted as $v \approx t$, the task of image-text retrieval can be performed in two settings: 1) Given a query text sample $t_q \in T$, the task aims to retrieve all the $v \in V$ satisfying $v \approx t_q$; or 2) Given a query image sample $v_q \in V$, the task aims to retrieve all the $t \in T$ satisfying $t \approx v_q$.

3.2 Framework Overview

The overview framework of AGREE is depicted in Figure 2. Initially, we identify the textual entities from the texts and the visual entities from the images, which are then encoded together with the original texts and images by the pre-trained VLP models. After that, we go to the fine-tuning stage, where three different modules are designed to learn the alignment between cross-modal entities:

- **Visual Entity-Image Alignment (VEA)** obtains visual entity-image pairs from Visual Genome [19], which are used to learn the alignment between visual entities and their corresponding images by contrastive learning and image region mask modeling.
- **Textual Entity-Image Alignment (TEA)** constructs a sentence only with textual entities and their visualizable properties (such as color and number) contained in each text, and then learns the alignment between the sentence and its corresponding image through contrastive learning and textual entity mask modeling.
- **Text-Image Entity Alignment (TIA)** further emphasizes the importance of cross-modal entity alignment by randomly masking entities grounded in the image to let the model be more sensitive to the missing of aligned entities across modalities.

Then we go to the re-ranking stage, which expects to refine the top-k ranking results with designed re-ranking strategies below:

- **Text-Image Bidirectional Re-ranking (TBR)** takes the top-k (e.g. k=10) retrieval results to do reverse image-text retrieval, whose results are then taken into account for re-ranking.
- **Textual Entity-Guided Re-ranking (EGR)** is specifically designed for the zero-shot scenarios, which takes the top-k retrieval results to calculate the similarity between the entities from images and texts, and then considers the similarities to refine the ranking result.

3.3 Entity Alignment in Fine-tuning

The fine-tuning of AGREE is built upon an image-text contrastive learning paradigm [36], which expects to shorten the distance between related images and texts, and far push those irrelevant ones in the embedding space.

The overall architecture of the fine-tuning is depicted in Figure 3, where both *global similarity* and *entity similarity* are calculated and then fused. While the *global similarity* is the similarity directly calculated between the embedding of image and text, the *entity similarity* is the cross-modal entity alignment that emphasizes similarity between an image-text pair based on three novelly proposed modules including VEA, TEA, and TIA. Particularly, VEA inputs the entity labels with corresponding images obtained from the external Multi-modal Knowledge Base (MMKB) and outputs the similarity between visual image and labels with two sub-modules VEM and MVC. TEA consists of 2 sub-modules TEE and MEC, which receives the text with textual entities and the image as input, and outputs the similarity between textual entities with the image. TIA also accepts the original image and text with entities, but learns to calculate the similarity between text entities with grounded images entities.

$$L = \frac{1}{2} \sum_{k=1} b(L_k^V + L_k^T) \tag{1}$$

Here we denote visual entities extracted from images as V_{obj}, and textual entities extracted from texts as T_{ent}. After encoded by VLP models, the representation of visual entities are $v_{io_i} = g(x_i; \gamma^{\alpha}) \in \mathbb{R}^{d_i}$ and that of the textual entities are $t_{te_j} = g(x_j; \gamma^{\beta}) \in \mathbb{R}^{d_t}$. Under the same contrastive learning paradigm for all fine-tuning modules, we sample b image-text pairs $\{i_k^V, t_k^T\}_{k=1}^{b}$ with an image set V and a text set T in a training batch, and for an image $i_k^V \in V$ in the selected samples, the text $t_k^T \in T$ is treated as its positive pair, while other texts are as in-batch negative samples. The contrastive loss of images and texts can be expressed as Equation (1), where the L_k^V and L_k^T refer to the image-to-text and text-to-image contrastive loss respectively. Taking image-to-text as an example, the loss function can be formulated as Equation (2) where $s_{j,k}^V$ denotes the k-th image to the j-th text. It is symmetric with the text-to-image part.

Following the contrastive learning paradigm, entities $t_{te_k}^T$ extracted from text t_k^T can be utilized as positive samples of the corresponding image i_k^V indicating textual-image entity-level alignment, while entities not mentioned in text t_k^T are regarded as negative samples. So as to the label of visual object $i_{io_k}^V$ from image i_k^V, while labels that are not detected from the image are negative ones.

$$L_k^V(i_k^V, \{t_j^T\}_{j=1}^b) = -log(exp(s_{k,k}^V)/\sum_j exp(s_{j,k}^V)) \tag{2}$$

In the following, we give more technical details of the three modules designed for calculating the entity similarity between each image-text pair.

Visual Entity-Image Alignment. Unlike many existing VLP models relying heavily on object detection models for fine-grained interactions [1, 9, 27, 30, 34, 39], we simply use the detected labels as meditation and re-establish an object-image library as MMKB for visual knowledge to align with their visual images. We choose Visual Genome (VG) [19], and design simple heuristic rules to filter images. Details for MMKB construction are described in Section 4.2. During fine-tuning, a label set of visual entities for in-batch N image with k entities $VO = \{io_m\}_{m=1}^k$ are collected, and related images of entities are found from our filtered MMKB. We explore 2 tasks to learn the entity-image alignment of each visual entity, following the paradigm of image-text contrastive learning. The overall loss function can be expressed as Equation $Loss_{VEA} = \frac{1}{2}(Loss_{VEM} + Loss_{MVC})$, where $Loss_{VEM}$ and $Loss_{MVC}$ are two sub-modules.

$$Loss_{VEM} = \frac{1}{2}\sum_{i=1} k(L_i^{VO} + L_i^{TO}) \tag{3}$$

1)Visual Entity Matching. The image of a detected object vo_i in the training batch is regarded as the positive sample of object image io_m from MMKB. With consideration that short labels of entities such as "sheperd dog" is inconsistent with the long and complete sentences in pre-training data, we use a unified rule-based method to construct the prompt [36] for the entity-level text samples to align with the images from the visual side. The prompt we use is expressed as *"a photo contains {entity}"*. Here, we optimize to match the label text of the visual object to_m and its image io_m, and formulate the loss function consistent with the global training objective Equation (2).

In Equation (3), $TO = \{to_m\}_{m=1}^k$ refers to the set of object texts and to_m is the corresponding object caption for recalled image io_m.

The distance of visual entity label and entity image is calculated through the embedding of `[CLS]` token from visual and textual entity encoders, denotes as $T_{to_{cls}}$ and V_{cls} in Figure 3. The simple framework enables the model to have the capacity of aligning object images with their correct labels.

2) Masking Visual Entity Consistency Alignment. Inspired by the pretraining task of VLP models to randomly mask some parts of the image for Masked Regions Classification [23, 39] or Masked Regions Features Regression [1, 48], we adopt masking strategy to learn representations of visual entities, but in a different way. We draw on the difference of similarity scores calculated between the label prompt with the original image and the image with masked entities, and minimize the margin ranking loss for visual entity consistency learning in Equation 4, where $y = 1$ and s_{io_k,to_k} denotes similarity between image and text. The visual embedding of an image with masked entity regions denotes $V_{wo/io_{cls}}$ in Figure 3. $Loss_{MVC}$ is to expect the score of the original image and object label higher, to stress more on those missing visual entities.

$$Loss_{MVC} = \sum_{k=1} max(0, -y \cdot (s_{io_k,to_k} - s_{io_{k(wo/io)},to_k})) \tag{4}$$

Textual Entity-Image Alignment. Motivated by the case in Figure 4, we re-consider the asymmetry of the visual and textual information and pay extra attention to the entity-level information in the text to align with the corresponding image.

1) Textual Entity Emphasizing Alignment. We first emphasize the entity-level information in the captions with stress on tokens of entities. Given an image-text pair $\{i_k^V, t_k^T\}$, p multi-level entity information from t_k^T are extracted, including named entities and the attributes (colors and numerical information in particular), denoted as $TE = \{te_m\}_{m=1}^p$. As shown in Figure 3, we extract "a white boat" (numerical), "a man" (entity), "blue clothes" (attribute), etc. Then the constructed entity prompts are treated as additional positive samples of image i_k^V for contrastive learning. The embedding of prompt label is expressed as $T_{te_{cls}}^1, T_{te_{cls}}^2, \ldots, T_{te_{cls}}^p$, to calculate the similarity with image embedding V_{cls}. We adopt average pooling for multiple entities in the same text, to look out for the importance of all entities simultaneously but not only consider the alignment with part of the entities. The loss function consistent with Equation (2) is expressed as Equation (5). Specifically, L_i^{TE} denotes $L_k^V(\{i_j^V\}_{j=1}^b, te_k^{TE}) = -\frac{1}{b}log\frac{exp(s_{k,k}^T)}{\sum_j exp(s_{j,k}^{TE})}$ for each entity text te_m and P denotes the number of entities limited for each caption.

$$Loss_{TEE} = \frac{1}{2P}\sum_{i=1} k(L_i^V + L_i^{TE}) \tag{5}$$

2) Mask Entity Consistency Alignment. By masking the textual entity tokens, we further consistently align images with text entities. Though inspired by [50], we do not give an exact vocabulary and classify entities like most models do, but adopt a more lightweight way to learn a unified cross-modal representation of textual entities. We re-compute the similarity between the original image i_k^V and text with masked entities $t_{wo/te}$, expecting the similarity $s_{i_k,t_{k(wo/te)}}$ between the image and the corrupted sentence smaller than that of the original text s_{i_k,t_k}. The embedding of text with masked text denotes as $T_{wo/te_{cls}}$ in Figure 3. Similar to the TEE module, average pooling is adopted. Within a batch of b image-text pairs samples, the loss function can be formulated as Equation (6).

$$Loss_{MEC} = \sum_{k=1} max(0, -y \cdot (s_{i_k,t_k} - s_{i_k,t_{k(wo/te)}})) \tag{6}$$

The objective of textual entity-image alignment is uniformed optimized as $Loss_{TEA} = \frac{1}{2}(Loss_{TEE} + Loss_{MEC})$.

As information in a single modality should be correlated with another complementary modality, we enhance the textual entities of image-text pairs to align with their visual representations in images instead of introducing additional knowledge for entities.

Text-Image Entity Alignment. To further bridge the gap between modalities and compensate for the aligning defects due to disordered vocabularies between heterogeneous information, we

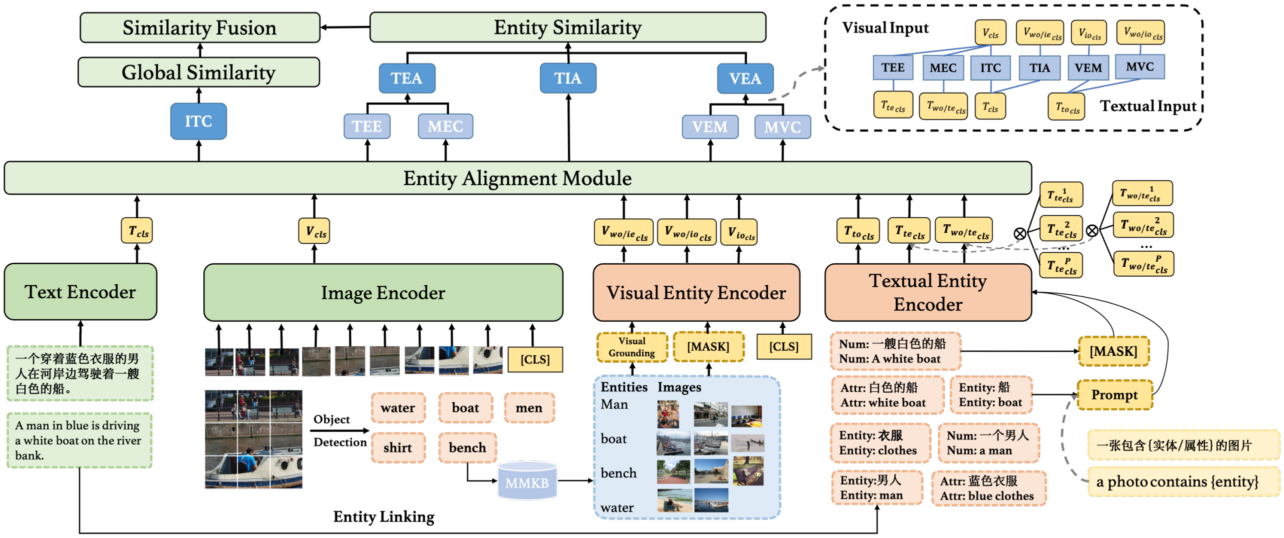


Figure 3: Architecture for AGREE fine-tuning framework. The inputs of the visual entity encoder are images of entities selected from Visual Genome [19]. The inputs of the textual entity encoder contain entities from the texts using entity linking, as well as label entities from Visual Genome, depending on the task. The box on the right top of the figure indicates visual and textual inputs for each module.

utilize a pre-trained visual grounding model as anchors to identify the region of each entity in the image for detected textual entities $TE = \{te_m\}_{m=1}^{p}$. The grounded entity is then masked in the image.

$$Loss_{TIA} = \sum_{k=1} max(0, -y \cdot (s_{i_k,t_k} - s_{i_{k(wo/ie)},t_k})) \qquad (7)$$

We still maximize the dissimilarity between the image with masked regions $i_{wo/ie}$, whose embedding is denoted as $V_{wo/ie_{cls}}$ in Figure 3, and $s_{i_{k(wo/ie)}}$ in Equation 7 is computed between $V_{wo/ie_{cls}}$ and T_{cls}. In TIA, we only focus on the consistency of entities in texts and images, since entity-image alignment on the visual side has been learned in VEA. Following the training objective mentioned above, the loss function is expressed as Equation (7). We jointly optimize VEA, TEA, and TIA. Each image and text requires three times of forward propagation, without introducing additional encoders or parameters. The overall training objective is $Loss = Loss_{CLIP} + \frac{1}{3}(Loss_{VEA} + Loss_{TEA} + Loss_{TIA})$.

3.4 Entity-Alignment in Re-ranking

Entity-Guided Re-ranking. To further boost the performance of VLP models with fine-grained entity-level interaction, we transform the strategy of the TEA module in Section 3.3 stage into an entity-alignment score for re-ranking. Following the same procedure, we on the one side convert extracted entities for text t_k^T into a prompt-based caption, and calculate with the candidate image i_m^V as textual entity alignment score $Score_{TEE} = \sum_{i=1}^{p} s_{i_m,te_i}$, and on the other side, entities in the text are replaced with [MASK] for textual entity consistency score as $Score_{MEC} = \sum_{i=1}^{p} max(0, -y \cdot (s_{i_m,t_i} - s_{i_m,t_{i(wo/te)}}))$. The entity-guided re-ranking score $Score_{EGR}$ is calculated with the combination of $Score_{TEE}$ and $Score_{MEC}$.

EGR only mimics the entity-level aligning process into image-text similarity scores, which is more compatible with VLP models. We adjust the coefficients of $Score_{All}$ and $Score_{EGR}$ on the validation set and apply them to the test set. The final score for ranking is expressed as $Score_{Final} = \alpha \cdot Score_{All} + (1 - \alpha) \cdot Score_{EGR}$. Images and texts are first pre-ranked with $Score_{All}$ for k candidates selection, and $Score_{EGR}$ is then used to re-rank the k candidates.

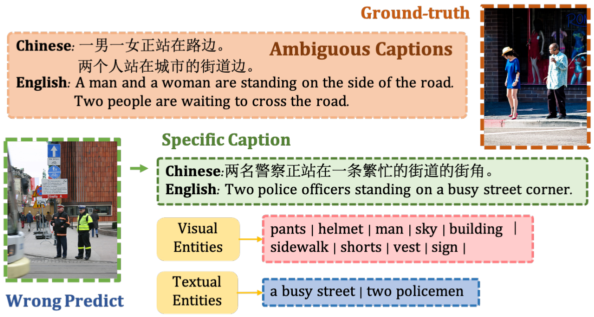

Ground-truth
Chinese: 一男一女正站在路边。
两个人站在城市的街道边。
English: A man and a woman are standing on the side of the road.
Two people are waiting to cross the road.
Ambiguous Captions
Specific Caption
Chinese:两名警察正站在一条繁忙的街道的街角。
English: Two police officers standing on a busy street corner.
Visual Entities
pants | helmet | man | sky | building |
sidewalk | shorts | vest | sign |
Textual Entities
a busy street | two policemen
Wrong Predict

Figure 4: An example indicates the inconsistency of images and texts, that captions are always concise but important, while images usually contain abundant entities.

Text-Image Bidirectional Re-ranking. The inconsistency of redundancy between rich visual information and concise textual knowledge may lead to misjudgment by incomplete information in one modality, particularly for those VLP models without fine-grained interactions. Therefore, we propose to compensate for the inconsistency by TBR, which introduces mutual information from complementary modality as additional supervision signals by reverse retrieval. TBR only relies on cross-modal samples themselves. Specifically, following [4], we see the text samples with the highest similarity $\{t^T_{rank_1}, t^T_{rank_2}, \dots, t^T_{rank_k}\}$ as reciprocal neighbors of image i_m^V, and oppositely retrieve the most similar images to each text from the candidate pool. Here, we employ ranking position only instead of the similarity score. Then the top-k candidates of image i_m^V are re-ranked with newly computed positions as $rank_{t^T_{rank_i}} = (rank_{t^T_{rank_i}-to-i^V_m} + i)/2$ for $t^T_{rank_i}$. It is the same for text-to-image retrieval. The simple but effective self-supervision only re-visits the ranking position, without the need for extra data, but ensures the visual and textual alignment to a certain extent.

Our re-ranking strategy compensates for the lack of fine-grained interaction, and avoids the wrong decisions made only through partial information. The TBR module is also applied for fine-tuning results for post-processing of image-text consistency alignment.

4 EXPERIMENTS

4.1 Experimental Setup

Table 1: Statistics of each image-text retrieval dataset.

	Flickr30K [47]		**Flickr30K-CN [20]**		**COCO-CN [29]**		**MUGE [32]**	
	#img.	#sen.	#img.	#sen.	#img.	#sen.	#img.	#sen.
train	29,000	145,000	29,783	148,915	18,341	20,065	129,380	248,786
val	1,014	5,070	1,000	5,000	1,000	1,100	29,806	5,008
test	1,000	5,000	1,000	5,000	1,000	1,053	30,399	5,004

To demonstrate the improvement for the cross-modal retrieval task of our proposed method, we conduct experiments on both Chinese and English VLP models with datasets in the two languages.

For Chinese, we experiment on COCO-CN [29], Flickr30k-CN [20] and MUGE [32] datasets. COCO-CN is re-splitted from MSCOCO [33] with human annotations, while Flickr30k-CN [20] is machine translated from Flickr30k [47] with human-translated for the test set. MUGE[1] is an image-text dataset under e-commerce scenarios, and only considers the text-to-image retrieval task. Since the test set of MUGE is not released, we use the validation set instead. Experiments in English are conducted on Flickr30k [47]. Detailed datasets statistics are shown in Table 1.

We report standard retrieval metrics for evaluation: recall at rank K denotes R-K and MR (mean recall of R-1, R-5, and R-10).

4.2 Implementation Details

Encoders. Following [11, 36], AGREE adopts the dual-encoder architecture, with a text encoder and an image encoder. We use the same encoder with sharing weights to encode visual and textual entities. Hidden state sequences obtained from image and text encoders are used to perform global or entity-level computations.
1)Text Encoder. We follow the same architecture of [11, 36], and use WordPiece [43] for Chinese tokenization and BPE [37] for tokenizing English. Besides, the special `[MASK]` token is introduced to mark the masked entity tokens. Given an input text of N_t tokens, the text encoder outputs an embedding sequence $\{t_{start}, t_1, \ldots, t_{N_t}, t_{end}\}$, where t_{start} denotes `[CLS]` for Chinese and `[SOS]` for English, and t_{end} denotes `[SEP]` or `[EOS]`, respectively.
2) Image Encoder. We adopt ViT [6] as the visual encoder, with images uniformly re-scaled to 224×224 and split into patches. Each patch is then linearly projected with positional embeddings and a `[CLS]` token. ViT-B/32 model includes 12 layers with a patch size of 32, and ViT-L/14 model includes 24 layers with patch size of 14.

The similarity between image and text is computed between special token t_{end} of texts and `[CLS]` of images. We choose CLIP [36] for English, and the CLIP$_{\text{ViT-L}}$ in Wukong [11] (pre-trained only with global similarity) for Chinese as the VLP model.
Fine-tuning Modules *1) Visual Entity Extraction.* We extract visual entities in images using model fine-tuned [39] on Visual Genome [19]. We select 38,848 images of 1600 image entities from Visual Genome as the MMKB. We strictly restrict the size of entities in images should not be smaller than 90×90, to avoid the interference which may be caused by too small visual objects.
2) Visual Grounding. We exploit the pre-trained GLIP [26] [2] model for visual grounding in the TIA module. The extracted textual entities are queries, to locate the regions of entities in the image.
3) Textual Entity Linking NLP tools are utilized respectively to obtain entities and attributes in Chinese and English. We then combine attributes with corresponding entities as phrases for full semantics. Specifically, we adopt LAC [16] for pre-processing Chinese captions, and spacy[3] for English text.

4.3 Model Training

To the best of our knowledge, neither Wukong [11] nor CLIP [36] publicly provides fine-tuning parameters for downstream tasks. As contrastive learning based VLP models usually require an extremely large batch size, such as [46] using 5120 for fine-tuning, which is also quite expensive for small laboratories or individual researchers. Therefore, we reproduce the fine-tuning results of both models using a smaller batch size with the re-implementation version in PyTorch of Wukong [11] and CLIP [12] in EasyNLP[4]. We employ NVIDIA V100 32G GPUs with a total batch size of 32 for ViT-L/14 and AdamW as the optimizer. The learning rate is set to 10^{-5}, and 0.001 for weight decay, for 50 epochs fully fine-tuning. Compared to [46], our settings are more acceptable and applicable.

4.4 Image-text Retrieval Results

Table 3: Fine-tuning results on Flickr30k [47].

Method	Flickr30k [47]								
	Image2Text				Text2Image				MR
	R-1	R-5	R-10	MR	R-1	R-5	R-10	MR	
E2E-VLP[44]†	86.2	97.5	-	-	73.6	92.4	-	-	-
SOHO[14]	86.5	98.1	-	-	72.5	92.7	-	-	-
Unicoder-VL[23]†	86.2	96.3	99.0	93.8	71.5	90.9	94.9	85.8	89.8
ROSITA [3]†	88.9	98.1	99.3	95.4	74.1	92.4	94.1	86.9	91.2
VILLA-L [9]†	87.9	97.5	98.8	94.7	76.3	94.2	96.8	89.1	91.9
UNITER-L [1]†	87.3	98.0	99.2	94.8	75.6	94.1	96.8	88.8	91.8
ERNIE-ViL [48]	89.2	97.3	99.1	95.2	75.1	93.4	96.3	88.3	91.7
LightningDOT [38]	87.2	98.3	99.0	94.8	75.6	94.0	96.6	88.7	91.8
VISTA-L [2]†	89.5	98.4	**99.6**	95.8	75.8	94.2	96.9	89.0	92.4
LoopITR [22]	89.6	98.6	99.5	95.9	77.2	94.3	97.6	89.7	92.8
Our Baseline	90.7	**98.9**	99.5	96.4	77.8	94.2	96.7	89.6	93.0
AGREE (FT only)	91.6	98.7	99.2	96.5	78.1	95.1	**97.8**	90.3	93.4
AGREE	**92.1(↑1.4)**	98.7	99.2	**96.7**	**82.8(↑4.0)**	**95.9**	**97.8**	**92.1**	**94.4**

Fine-tuning. Fine-tuning results using settings in Section4.3 with pre-trained models publicly released by Wukong [11] and CLIP [36] are shown in the tables (denotes as **Our Baseline**), with the best result for each metric in **bold** and the second best underlined. For a fair comparison, we compare with models in their large settings in Table 3 respectively)[5] and also report the reported fine-tuning results in Wukong [11] besides our baseline. The specific growth value over "Our Baseline" is highlighted on the tables. We show results with the full AGREE framework and AGREE without re-ranking (denotes as **AGREE (FT only)**), to demonstrate the effects of AGREE in the fine-tuning stage. We compare AGREE against methods with complex attention interactions using object detectors, including Unicoder-VL [23], VILLA [9], UNITER [1], as well as

[1] https://tianchi.aliyun.com/muge

[2] https://github.com/microsoft/GLIP

[3] https://spacy.io/spaCy

[4] https://github.com/alibaba/EasyNLP

[5] "L" denotes the large settings of models with different variants.

Table 2: Zero-shot re-ranking results on Chinese datasets with pre-trained weights provided by Wukong [11].

Dataset	Method	ViT-B/32									ViT-L/14								
		Image2Text				Text2Image				MR	Image2Text				Text2Image				MR
		R-1	R-5	R-10	MR	R-1	R-5	R-10	MR		R-1	R-5	R-10	MR	R-1	R-5	R-10	MR	
COCO-CN [29]	Our Baseline	52.4	79.7	89	73.7	49.1	79.5	88.6	72.4	73.1	50.9	79.5	89.8	73.4	48.6	76.9	86.4	70.7	72.0
	+ EGR	53.5	79.7	88.2	73.8	48.8	79.9	88.9	72.5	73.2	51.5	79.7	90.4	73.9	49.6	77.4	86.9	71.3	72.6
	+ TBR	55.1	**82.2**	**91.1**	76.1	52.8	**81.9**	**90.2**	75	**75.6**	54.2	81.9	91.1	75.7	53.2	79.9	88.7	73.9	74.8
	AGREE (RR only)	**56.4(↑4.0)**	81.6	89.8	75.9	**53.7(↑4.6)**	81.8	89.8	**75.1**	75.5(↑2.4)	**54.5(↑3.6)**	**82.8**	**91.5**	**76.3**	**54.2(↑5.6)**	**80.2**	**88.7**	**74.4**	**75.3(↑3.3)**
Flickr30k-CN [20]	Our Baseline	74.1	94.2	97.7	88.7	51.5	78.2	85.8	71.8	80.3	72.4	**91.8**	96.3	86.8	47.2	74.2	82.9	68.1	77.5
	+ EGR	75.9	93.8	97.6	89.1	51.5	78.3	85.7	71.8	80.5	72.1	**91.8**	**96.6**	86.8	47.2	74.1	82.9	68.1	77.5
	+ TBR	76.1	**94.4**	97.7	89.4	**58.5**	82.2	87.5	**76.1**	82.7	73.3	**91.8**	96.3	87	**53.5**	**78.9**	**84.8**	**72.4**	**79.8**
	AGREE (RR only)	**77(↑2.9)**	94.2	**97.8**	**89.7**	58.4(↑6.9)	**82.3**	**87.7**	**76.1**	**82.9(↑2.6)**	**73.4(↑1.0)**	**91.8**	96.4	**87.2**	53.4(↑6.2)	78.8	**84.8**	72.3	**79.8(↑2.3)**
MUGE [32]	Our Baseline	-	-	-	-	37.3	64	73.5	58.3	58.3	-	-	-	-	43.4	69.4	78.1	63.7	63.7
	AGREE (RR only)	-	-	-	-	**38.5(↑1.2)**	**64.9**	**73.9**	**59.1**	**59.1(↑0.8)**	-	-	-	-	**44.5(↑1.1)**	**70.2**	**78.3**	**64.4**	**64.4(↑0.7)**

methods incorporating external knowledge for pre-training including ERNIE-ViL [48] and ROSITA [3]. Innovative research including VISTA [2] which aggregates scene text, and LoopITR [22] which combines the advantages of cross-attention and dual encoder are also considered. We also compare with three Chinese VLP models CLIP, FILIP, and Wukong from [11] for COCO-CN [29] as Table 4. For COCO-CN [29], performance on cross-lingual VLP models including M^3P [35] and UC^2 [49] are also displayed.

We can explore the contributions of cross-modal entities in cross-modal retrieval from the results. Compared with fine-tuning on large-scale VLP models using only global similarity like Wukong [11] and CLIP [36], AGREE shows great improvements on COCO-CN, and is significantly higher than the fine-tuning results on VLP models with patch-token fine-grained contrastive learning framework [46] in Wukong (named $\text{FILIP}_{\text{ViT-L}}$), which proves the effectiveness of our entity-based strategy. For the English dataset, results on Flickr30k also obtain great improvement, higher than VILLA [9] which incorporates adversarial learning, and Unicoder-VL [18] which includes several pre-training tasks for in-depth image-text representation learning. Our approach surpasses UNITER [1] on both Chinese and English datasets, and higher than knowledge injected pre-training methods [3, 48][6]. Our improvement is mainly reflected on the results of R-1, on both Chinese and English datasets. On the premise of VLP models' strong fitting ability pre-trained on a large amount of data, our improvement mainly lies in optimizing the re-ranking results. The ranking of the ground-truths can be moved to the front, so as to improve the mean recall. The subsequent empirical analysis will further illustrate that.

Zero-shot Re-ranking. We also test the effectiveness of the re-ranking strategy in AGREE under zero-shot scenarios, and show results on two modules only (EGR or TBR) as well as their combinations (denote as **AGREE(RR only)**) in Table 2. Experimental results on different datasets and image encoders show that the MR of both image-to-text and text-to-image can be improved by 3%, and is brought by a more significant increase on R-1 with an average of about 5%. For the dataset MUGE [32] which only includes text-to-image retrieval task, the reverse retrieval of image-to-text also leads to apparent improvement. Moreover, the boosting of text-to-image is more obvious compared with that of image-to-text in most of our experiments, which further confirms our observation in Section 3.4 that the inconsistency lies in the richness of image-text information. With a richer and more specific description of the image through the image-to-text reverse retrieval result, AGREE can better assist the text to find a more suitable image, as shown in Figure 4.

Table 4: Fine-tuning results on COCO-CN [29] .

Method	COCO-CN [29]								
	Image2Text				Text2Image				MR
	R-1	R-5	R-10	MR	R-1	R-5	R-10	MR	
UNITER[1]†	-	-	-	-	-	-	-	-	87.3
M^3P [35]*	-	-	-	-	-	-	-	-	86.2
UC^2 [49]*	-	-	-	-	-	-	-	-	88.4
$\text{CLIP}_{\text{ViT-L}}$ [11]	68.3	93.0	97.3	86.2	70.1	92.2	96.4	86.2	86.2
$\text{FILIP}_{\text{ViT-L}}$ [11]‡	69.1	91.3	96.9	85.8	72.2	92.4	97.2	87.3	86.7
$\text{Wukong}_{\text{ViT-L}}$ [11]‡	**73.3**	94.0	**98.0**	**88.4**	**74.0**	**94.4**	**98.1**	**88.8**	**88.6**
Our Baseline	69.9	93.5	97.6	87.0	70.5	92.4	96.6	86.5	86.8
AGREE (FT only)	71.9	93.8	97.6	87.8	71.1	93.2	97.2	87.2	87.5
AGREE	73.0(↑3.1)	**94.6**	97.6	**88.4**	73.4(↑2.9)	93.6	97.3	88.1	88.3

4.5 Ablation Studies

To demonstrate the importance of visual and textual entities for image-text retrieval, we present different variants of AGREE in Table 5, with $CLIP_{ViT-L-14}$ in [11] as our baseline, to show the effect of AGREE on COCO-CN [29].

Table 5: Ablation studies on COCO-CN [29].

Method	ViT-L/14							
	Image2Text				Text2Image			
	R-1	R-5	R-10	MR	R-1	R-5	R-10	MR
Our Baseline	69.9	93.5	97.6	87.0	70.5	92.4	96.6	86.5
+VEA	70.6	93.6	97.5	87.2	70.2	93.1	96.8	86.7
+TEA	71.4	93.4	97.0	87.3	70.6	93.3	97.0	86.9
+TIA	70.3	92.9	97.8	87.2	71.5	92.5	96.9	87.0
+VEA+TEA	70.3	92.9	97.8	87.0	70.4	92.5	96.8	86.5
+TIA+TEA	70.6	93.2	97.2	87.0	70.6	92.5	97.2	86.7
+TBR	71.6	93.9	97.5	87.7	72.3	93.4	97.3	87.7
AGREE (FT only)	71.9	93.8	97.6	87.8	71.1	93.2	97.2	87.2
AGREE	**73.0(↑3.1)**	**94.6**	**97.6**	**88.4(↑2.2)**	**73.4(↑2.9)**	**93.6**	**97.3**	**88.1(↑1.9)**

Impacts of Entity Learning. Retrieval performance on individual module shows the significant improvement brought by visual and textual entities, which indicates that aligning cross-modal entities is crucial to lift effects on image-text retrieval. The results with only VEA or TEA module on COCO-CN [29] also reveal the effectiveness in utilizing entity information, better learning image-text representations and alignments. It is also worth noticing the great improvement brought by TIA module. We believe that the gain is related to visual grounding, which makes the correspondence between entities from images and texts more specific.

Combination of Modules. As the combination of modules in Table 5, an interesting observation is that although exploiting VEA or TEA alone leads to an increase, it is slightly decreased when they are combined (e.g. *+VEA+TEA*, even though still improved compared to *Our baseline*). However, the addition of TIA will eliminate the

[6] Since we claim that AGREE is a fine-tuning framework on the basis of dual-stream VLP models and thus applicable to most VLP models in dual-stream architecture and brings benefit to further fine-grained interactions, we only demonstrate its effectiveness on CLIP-style models. Thus, we do not directly compare with dual-stream models using larger datasets or in-house data for pre-training, such as FILIP [46], ALIGN [15], SimVLM [42]. For a fair comparison, in Table 4 and Table 3, we use † to mark the models which adopt multi-modal fusion interaction modules, and mark the methods utilizing multilingual data for training with *. VLP models with complex similarity calculations between images and texts are highlighted with ‡.

contradictory phenomenon. It is the same with the comparison with *+TIA +TEA*. Thus we can conclude that the representation learning of entities from multiple modalities is very important in AGREE. When aligning visual and textual entities simultaneously, it is quite necessary to build a bridge (as TIA in AGREE) to align the image-text entities. To a certain extent, this ensures the consistency between images and texts, and eliminates the contradiction caused by utilizing image-text entities for augmentation at the same time.

4.6 Few-shot Experiments

We also verify the effects of AGREE under few-shot scenarios. As shown in Table 6, we randomly split the training set of COCO-CN into 5%, 15%, 25% and 50%, and perform experiments on the test split the same as Table 1. For each volume of data, we conduct three experiments and report their average. When the amount of data is quite small, experimental results show a significant improvement with AGREE (FT only), (e.g.about 1.6% improvement on R-1 for 5% train data). After adopting the optimization strategy of fine-tuning and re-ranking at the same time, the MR score of a smaller amount of data can achieve or even exceed the results with a larger amount of data using the original fine-tuning method (e.g. AGREE with 25% data compared with 50% using baseline). We consider this an exciting result, which provides an efficient way of fine-tuning. Focusing on the alignment of entity-level information during fine-tuning can greatly reduce training data dependencies. In this way, we are even able to achieve better results with less data.

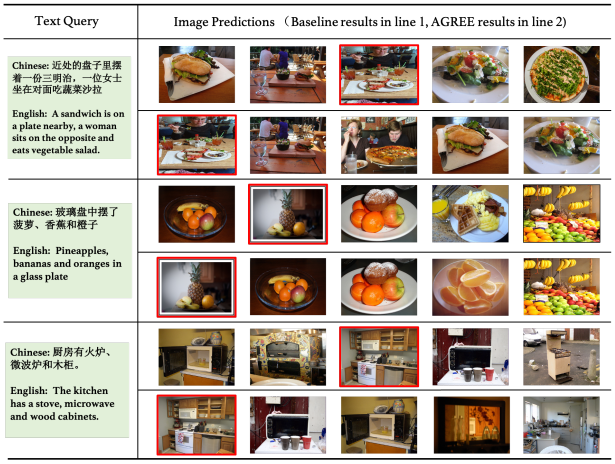
Table 6: Few-shot results on COCO-CN [29] with different volume of training data.

	Method	COCO-CN [29]							
		Image2Text				Text2Image			
		R-1	R-5	R-10	MR	R-1	R-5	R-10	MR
5%	Our Baseline	57.2	85.2	93.0	78.5	59.3	84.8	93.0	79.0
	AGREE (FT only)	58.6	87.1	93.4	79.7	61.7	85.9	94.0	80.6
	AGREE	**61.8(↑4.6)**	**88.4**	**94.3**	**81.5**	**63.7(↑4.6)**	**86.7**	**93.9**	**81.4**
15%	Our Baseline	62.4	89.0	95.5	82.3	65.9	88.9	94.9	83.2
	AGREE (FT only)	62.4	90.1	95.4	82.6	65.6	89.8	95.3	83.6
	AGREE	**65.7(↑3.3)**	**90.5**	**96.2**	**84.1**	**67.2(↑1.3)**	**90.4**	**95.8**	**84.5**
25%	Our Baseline	65.5	90.2	96.1	83.9	66.1	89.5	94.9	83.5
	AGREE (FT only)	64.7	91.5	96.2	84.1	66.5	90.7	95.5	84.2
	AGREE	**67.0(↑1.5)**	**91.9**	**96.6**	**85.2**	**68.6(↑2.5)**	**91.2**	**95.4**	**85.1**
50%	Our Baseline	67.3	91.6	97.0	85.3	67.8	91.5	96.1	85.2
	AGREE (FT only)	68.0	92.0	96.9	85.6	68.6	91.7	96.7	85.7
	AGREE	**69.8(↑2.5)**	**92.9**	**97.1**	**86.6**	**70.3(↑2.5)**	**92.2**	**96.4**	**86.3**

4.7 Case Study

Several examples are presented to illustrate the effectiveness of AGREE, and to reveal the importance of aligning cross-modal entities for the performance improvement of image-text retrieval. The examples of text-to-image and image-to-text of COCO-CN [29] are shown in Figure 4.7 in Figure 4.7 respectively, with ground-truth framed in the red line. As in Figure 4.7, the Top-5 images with AGREE obviously contain more entities corresponding to the query and the correct samples have a higher rank. For example, the Top-1 result of query *"Pineapples, bananas and oranges in a glass plate"* does not include the important entity *"pineapples"*. AGREE helps to re-establish this correspondence between entities, resolving the problem with ineffective fine-grained interactions from VLP models as in Figure 1. It is the same as cases in Figure 4.7, where the image query in the first row is more accurately matched to the text with entities in the image, including *"a plate of pizza"*, *"bottles"*, *"pots"*.

We pay more attention to the consistency of multiple entities in both images and texts, thus optimizing the ranking result.


Text Query
Image Predictions （Baseline results in line 1, AGREE results in line 2)

Chinese: 近处的盘子里摆着一份三明治，一位女士坐在对面吃蔬菜沙拉

English: A sandwich is on a plate nearby, a woman sits on the opposite and eats vegetable salad.

Chinese: 玻璃盘中摆了菠萝、香蕉和橙子

English: Pineapples, bananas and oranges in a glass plate

Chinese: 厨房有火炉、微波炉和木柜。

English: The kitchen has a stove, microwave and wood cabinets.

Figure 5: Text-to-image Top-5 retrieval result examples.

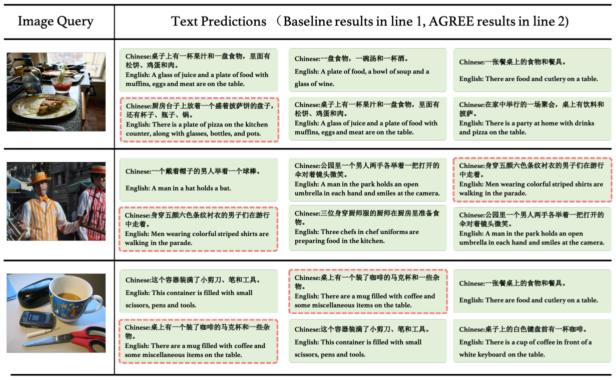


Image Query
Text Predictions （Baseline results in line 1, AGREE results in line 2)

Chinese:桌子上有一杯果汁和一盘食物，里面有松饼、鸡蛋和肉。
English: A glass of juice and a plate of food with muffins, eggs and meat are on the table.

Chinese:一盘食物，一碗汤和一杯酒。
English: A plate of food, a bowl of soup and a glass of wine.

Chinese:一张餐桌上的食物和餐具。
English: There are food and cutlery on a table.

Chinese:厨房台子上放着一个盛着披萨饼的盘子，还有杯子、瓶子、锅。
English: There is a plate of pizza on the kitchen counter, along with glasses, bottles, and pots.

Chinese:桌子上有一杯果汁和一盘食物，里面有松饼、鸡蛋和肉。
English: A glass of juice and a plate of food with muffins, eggs and meat are on the table.

Chinese:在家中举行的一场聚会，桌上有饮料和披萨。
English: There is a party at home with drinks and pizza on the table.

Chinese:一个戴着帽子的男人举着一个球棒。
English: A man in a hat holds a bat.

Chinese:公园里一个男人两手各举着一把打开的伞对着镜头微笑。
English: A man in the park holds an open umbrella in each hand and smiles at the camera.

Chinese:身穿五颜六色条纹衬衣的男子们在游行中走着。
English: Men wearing colorful striped shirts are walking in the parade.

Chinese:身穿五颜六色条纹衬衣的男子们在游行中走着。
English: Men wearing colorful striped shirts are walking in the parade.

Chinese:三位身穿厨师服的厨师在厨房里准备食物。
English: Three chefs in chef uniforms are preparing food in the kitchen.

Chinese:公园里一个男人两手各举着一把打开的伞对着镜头微笑。
English: A man in the park holds an open umbrella in each hand and smiles at the camera.

Chinese:这个容器装满了小剪刀、笔和工具。
English: This container is filled with small scissors, pens and tools.

Chinese:桌上有一个装了咖啡的马克杯和一些杂物。
English: There are a mug filled with coffee and some miscellaneous items on the table.

Chinese:一张餐桌上的食物和餐具。
English: There are food and cutlery on a table.

Chinese:桌上有一个装了咖啡的马克杯和一些杂物。
English: There are a mug filled with coffee and some miscellaneous items on the table.

Chinese:这个容器装满了小剪刀、笔和工具。
English: This container is filled with small scissors, pens and tools.

Chinese:桌子上的白色键盘前有一杯咖啡。
English: There is a cup of coffee in front of a white keyboard on the table.

Figure 6: Image-to-text Top-3 retrieval result examples.

5 CONCLUSION

In this paper, we propose a lightweight and applicable approach AGREE, to align cross-modal entities for image-text retrieval at the fine-tuning and re-ranking stages. We employ external knowledge and tools to construct extra fine-grained vision-text pairs, and then emphasize cross-modal entity alignment through contrastive learning and entity-level mask modeling in fine-tuning. Two re-ranking strategies are also proposed including one specially designed for zero-shot scenarios. We conduct extensive experiments with several VLP models on multiple Chinese and English datasets, and the results show that our approach achieves state-of-art results under various settings for both fine-tuning and zero-short scenarios. Our experiments under few-shot scenarios also verify that AGREE can significantly reduce data dependency. We hope our work could inspire future research in the visual and linguistic communities.

6 ACKNOWLEDGEMENT

This research has been funded by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), National Natural Science Foundation of China (No.62072323), and Shanghai Science and Technology Innovation Action Plan (No. 22511104700) and supported by Alibaba Group through Alibaba Innovative Research Program.

REFERENCES

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In European conference on computer vision. Springer, 104–120.

[2] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. 2022. ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5184–5193.

[3] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration. In Proceedings of the 29th ACM International Conference on Multimedia. 797–806.

[4] Agni Delvinioti, Hervé Jégou, Laurent Amsaleg, and Michael E Houle. 2014. Image retrieval with reciprocal and shared nearest neighbors. In VISAPP, Vol. 2. 321–328.

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 1218–1226.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017).

[8] Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual Cross-modal Pretraining for Multimodal Retrieval. In Proceedings of NAACL. 3644–3650.

[9] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. Advances in Neural Information Processing Systems 33 (2020), 6616–6628.

[10] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2022. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. TACL 10 (2022), 503–521.

[11] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark. (2022).

[12] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6163–6171.

[13] Yan Huang, Qi Wu, Wei Wang, and Liang Wang. 2018. Image and sentence matching via semantic concepts and order learning. IEEE transactions on pattern analysis and machine intelligence 42, 3 (2018), 636–650.

[14] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12976–12985.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning. PMLR, 4904–4916.

[16] Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese lexical analysis with deep bi-gru-crf network. arXiv preprint arXiv:1807.01882 (2018).

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of CVPR. 3128–3137.

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning. PMLR, 5583–5594.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123, 1 (2017), 32–73.

[20] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In Proceedings of ACM MM. 1549–1557.

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In Proceedings of the European conference on computer vision (ECCV). 201–216.

[22] Jie Lei, Xinlei Chen, Ning Zhang, Mengjiao Wang, Mohit Bansal, Tamara L Berg, and Licheng Yu. 2022. LoopITR: Combining Dual and Cross Encoder Architectures for Image-Text Retrieval. arXiv preprint arXiv:2203.05465 (2022).

[23] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In Proceedings of AAAI, Vol. 34. 11336–11344.

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS 34 (2021), 9694–9705.

[25] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In Proceedings of CVPR. 4654–4662.

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10965–10975.

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. arXiv preprint arXiv:2012.15409 (2020).

[28] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2022. UNIMO-2: End-to-End Unified Vision-Language Grounded Learning. arXiv preprint arXiv:2203.09067 (2022).

[29] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on Multimedia 21, 9 (2019), 2347–2360.

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In ECCV. Springer, 121–137.

[31] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021).

[32] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823 (2021).

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In ECCV. Springer, 740–755.

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems 32 (2019).

[35] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 3977–3986.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. PMLR, 8748–8763.

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of ACL. 1715–1725.

[38] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In Proceedings of NAACL. 982–997.

[39] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of EMNLP.

[40] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In Proceedings of ACM MM. 154–162.

[41] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In European Conference on Computer Vision. Springer, 18–34.

[42] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021).

[43] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016).

[44] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. arXiv preprint arXiv:2106.01804 (2021).

[45] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. 2019. Deep adversarial metric learning for cross-modal retrieval. World Wide Web 22, 2 (2019), 657–672.

[46] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021).

[47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL 2 (2014), 67–78.

[48] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In Proceedings of AAAI, Vol. 35. 3208–3216.

[49] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In Proceedings of CVPR. 4155–4165.

[50] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER. In Proceedings of ACL. 2251–2262.