Object-Aware Query Perturbation for Cross-Modal Image-Text Retrieval

Naoya Sogi, Takashi Shibata, and Makoto Terao

Visual Intelligence Research Laboratories, NEC Corporation, Kanagawa, Japan
naoya-sogi@nec.com, t.shibata@ieee.org, m-terao@nec.com

Abstract. The pre-trained vision and language (V&L) models have substantially improved the performance of cross-modal image-text retrieval. In general, however, V&L models have limited retrieval performance for small objects because of the rough alignment between words and the small objects in the image. In contrast, it is known that human cognition is object-centric, and we pay more attention to important objects, even if they are small. To bridge this gap between the human cognition and the V&L model's capability, we propose a cross-modal image-text retrieval framework based on "object-aware query perturbation." The proposed method generates a key feature subspace of the detected objects and perturbs the corresponding queries using this subspace to improve the object awareness in the image. In our proposed method, object-aware cross-modal image-text retrieval is possible while keeping the rich expressive power and retrieval performance of existing V&L models without additional fine-tuning. Comprehensive experiments on four public datasets show that our method outperforms conventional algorithms.

Keywords: cross-modal retrieval · vision and language · object centric

1 Introduction

Cross-modal image-text retrieval is one of the mainstream tasks in pattern recognition [1, 29] and has various applications including e-commerce [17, 38] and video surveillance [26]. Recent pre-trained vision-and-language (V&L) models [5, 36, 37, 43, 52, 60] have caused a paradigm shift. Those pre-trained models substantially outperform legacy cross-modal image-text retrieval by leveraging massive amounts of training data while equipping advantages such as zero-shotness and generalizability.

Nevertheless, those V&L models are not any panacea; those V&L models have limited performance for small objects due to the rough alignment between text and the fine-grained localization of these small targets in the image. An example of retrieval results obtained by a sophisticated V&L model, BLIP2 [35], on the Flickr 30K [42] dataset is shown in Fig. 1(a). The matching between the target objects and the input query text is weak because the objects, e.g., the person and the rollerblades, are small, resulting in incorrect retrieval results. Although the drawback of the retrieval performance degradation related to those



Fig. 1: Example results by our method with BLIP2 [35]. In BLIP2, the matching between the target objects and the input text is weak because the objects are small, resulting in incorrect retrieval results.

small objects is critical in actual applications, it has been hidden behind the overwhelming performance gains of the recent pre-trained V&L models on the public benchmark datasets.

In contrast, humans can effectively understand visual scenes by an ability that lies in their object-centered (or compositional) perception [31,61]. Owing to this human object-centered perception, the human visual function is highly robust to the size of the target object. For example, objects critical to understanding a scene, e.g., a small rescue caller in an image of a disaster scene, will be gazed at regardless of the target object's size.

The lack of such object-awareness in V&L models is a major issue, especially for human-centered vision tasks, e.g., image retrieval. Although legacy image retrieval algorithms using object detection have also been proposed [22,32], these methods cannot inherit the strengths of recent pre-trained V&L models. There is a strong demand for a general framework that bridges the gap between human perception and the V&L models while inheriting the potential capabilities of pre-trained V&L models including zero-shotness and absolute high performance.

This paper proposes an object-aware query-perturbation for cross-modal retrieval as a solution to the above demand. Our Query-Perturbation (Q-Perturbation) increases object awareness of V&L models by focusing on object information of interest even when the objects in an image are relatively small. An example of retrieval results by the proposed method is shown in Fig. 1(b). In contrast to the existing methods, our retrieval framework, i.e., a V&L model with Q-Perturbation, can perform accurate retrieval even for images that capture small objects. The core mechanism of Q-Perturbation is to enhance queries with keys corresponding to object regions, at cross-attention modules in a V&L model.

Naively enhancing queries in the existing V&L model breaks the original weights, resulting in poor performance. Our query perturbation prevents weight breaking by enhancing queries within a subspace constructed using keys corresponding to target objects. That is, queries are first decomposed by subspaces representing object information and then enhanced using the decomposed information, which is object information retained in original queries. This process

naturally selects queries to be enhanced as the process uses only object information in original queries, i.e., a query is not enhanced if it does not have object information. As a result, our Query Perturbation improves the object-awareness of a V&L model while inheriting the impressive performance of a V&L model.

The proposed method is applicable to a variety of V&L models and can avoid increased computational cost due to data updates and catastrophic forgetting due to re-training because the proposed method is training-free and easy to implement. Comprehensive experiments on public datasets demonstrate the effectiveness of the proposed method[1].

The contributions of this paper are including: 1) We propose an object-aware query perturbation (Q-Perturbation) for cross-modal image-text retrieval. 2) We construct the object-aware retrieval framework by plugging Q-Perturbation into state-of-the-art V&L models, e.g., BLIP2 [35], COCA [60], and InternVL [6]. 3) Comprehensive experiments on public data demonstrate the effectiveness of the proposed method. In addition, we propose a new metric that mitigates the dataset bias regarding object size.

2 Related works

Cross-Modal Image-Text Retrieval. The study of cross-modal image-text retrieval is a fundamental task in vision, and many existing methods exist [3,7,9, 16,18,20,23,24,32,40,50,51,53,57,58,62,65,66]. A standard approach acquires a text-language common space from image and text datasets prepared as training data in advance [10, 15, 16, 28, 34, 46, 48]. To acquire an accurate image-text common space, several approaches have been introduced to improve the loss function and distance space, such as metric learning [16,66,67] and probabilistic distribution representation [10, 30, 54]. Various extensions have been proposed for fine-grained retrieval [2,32,40,56,64] by introducing object detection [22,32], graph-based relationships between objects [12, 41, 63], re-weighting strategy [2, 55, 56], and the attention mechanism [7, 65]. These existing approaches for fine-grained retrieval suggest that object awareness is an essential cue for locally detailed cross-modal retrieval. This paper focuses on object awareness for the pre-trained V&L models [13,19,25,35–37,49]. We propose a simple-yet-effective framework that can efficiently improve the performance of image-text retrieval for an image containing small objects that are semantically important.

Pre-trained Vision & Language Model. In recent years, cross-modal image-text retrieval using the V&L model has been proposed as a new paradigm [13, 19, 25, 35–37, 49]. Vision language pre-training, such as CLIP [43], trains vision language alignments from large numbers of image-text pairs through a self-supervised task. Before this paradigm, the existing image-text retrieval methods mainly focused on training algorithms using medium-sized datasets such as Flicker 30K and COCO. In contrast, the recent cross-modal image-text retrieval

[1] The code is publicly available at `https://github.com/NEC-N-SOGI/query-perturbation`

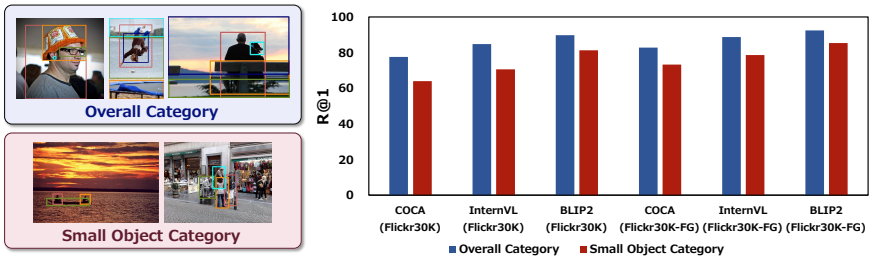


Fig. 2: Performance degradation induced by small objects.

using the pre-trained V&L models outperforms those legacy cross-modal image-text retrieval methods, achieving high zero-shot performance on diverse datasets and enabling open vocabulary retrieval. In particular, the recently proposed BLIP2 [35] has overwhelming performance in cross-modal image-text retrieval. However, it has recently been pointed out that there is a weakness in the localization for the V&L models such as CLIP, and several simple improvements have been proposed [45, 68]. As described later, this weakness has also been shared in cross-modal image-text retrieval. The proposed method is a novel framework that can overcome this weakness through cross-modal image-text retrieval while taking advantage of the potential capabilities of the existing V&L models.

3 Performance Degradation Induced by Small Objects

We discuss the performance degradation induced by small objects in a target image. We compared the overall performance of text-image retrieval (called overall category) on Flickr-30K [42] and Flickr-FG [8] with the performance on a dataset consisting of images with only small detected objects (called small object category). Specifically, we selected images for the small object category, where the ratio of the largest detected object rectangle's area to the entire image's area is less than 10%. Figure 2 shows the comparison results. Here, we used Recall@1 as the evaluation metric. It can be seen that as the area of the object detection rectangle becomes relatively smaller, the retrieval becomes more difficult, and the retrieval performance degrades.

The degradation is observed not only in Flickr-30K but also in Flickr-FG, where more detailed captions are annotated. Interestingly, we also find that this is a common drawback in the recent pre-trained V&L models [6, 35, 60]. This drawback is underestimated in the standard evaluation metric, Recall@K, because the number of images belonging to the small object category is small. For example, the number of images in the small object category accounts for about 1.5% of all images in Flickr-FG and Flickr-30K. We introduce an object-aware query perturbation to improve the poor performance of images with such small objects. Furthermore, we also discuss the effectiveness of the proposed method using an evaluation metric that accounts for this data bias regarding the object size in Sec 5.

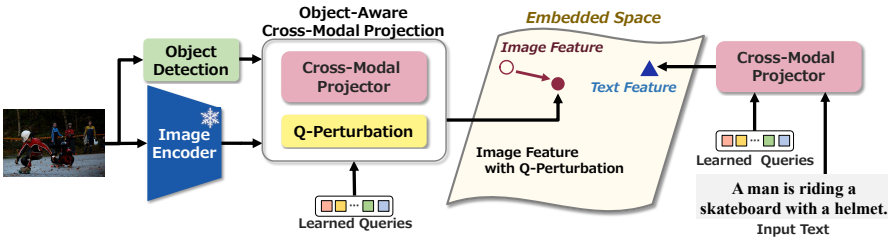


Fig. 3: Overview of the proposed framework. The proposed framework constructs an object-aware cross-modal projector by incorporating localization cues from object detection into the existing cross-modal projector.

4 Method

We describe the overview of our framework and key idea of the query perturbation (Q-Perturbation), then describe the details of the Q-Perturbations for the Q-Former module in BLIP2 [35]. Finally, we explain an extension of our Q-Perturbations to other V&L models [6, 60].

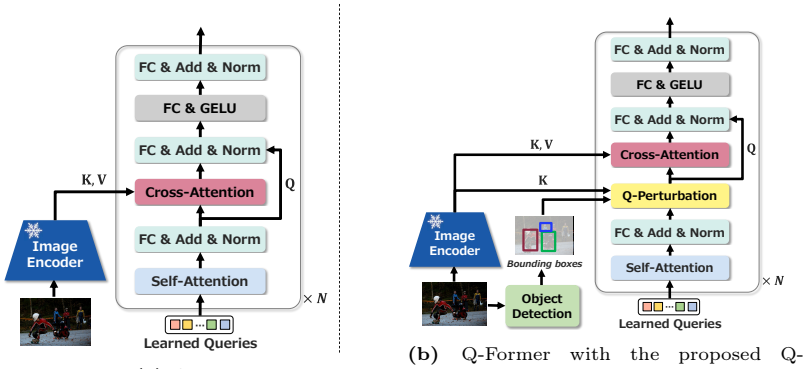
4.1 Overview

Our proposed method aims at improving the retrieval performance for an image containing small objects by extending the existing V&L models while inheriting the high expressiveness of these models. In general, V&L models contain a cross-modal projection module that aligns language features with image features. For example, BLIP2 introduces Q-Former architecture with a transformer structure as a cross-modal projector that combines image and text features.

An overview of our proposed framework is shown in Fig. 3. The proposed framework is also an architecture that leverages cross-modal projectors such as Q-Former [35] and QLLaMA [6]. The input text, a retrieval query, is similar to a standard cross-modal retrieval, and feature representations of texts are generated using a text encoder. The proposed framework constructs an object-aware cross-modal projector by incorporating localization cues obtained from object detection into the existing cross-modal projector, in addition to image features obtained from existing image encoders. The key is how to incorporate the localization cues into existing cross-modal projection modules. To do this, the proposed method introduces an object-aware query perturbation, called *Q-Perturbation*, that adaptively adjusts the query according to the size of the detected objects and bounding boxes in the image.

4.2 Basic Idea: Object-Aware Query Perturbation

A standard approach to constructing a cross-modal projector in a transformer-based module is to introduce cross-attention. For example, in BLIP2, cross-attention is introduced in the Q-Former to integrate image features with queries, including learned queries and text tokens.



(a) Q-Former.

(b) Q-Former with the proposed Q-Perturbation.

Fig. 4: Overview of our object-aware mechanism with Q-Perturbation module. Our framework can incorporate the object localization cues from the bounding boxes into a cross-modal projection module, e.g., Q-Former, with minimal modification.

Our goal is to incorporate object localization cues from the bounding boxes into cross-modal projection modules with minimal modification while taking advantage of the highly expressive power of the existing V&L models. To this end, the following must be satisfied:
- **Inheritability**: The proposed method must be object-aware cross-modal projection without significantly destroying the weights and structures already learned, in order to maximize the potential of the existing V&L models.
- **Flexibility**: The proposed method must be scalable and flexible regarding the size and number of detected objects.

In the proposed method, as shown in Fig. 4, the Q-Perturbation is used to perturb the already obtained query to emphasize the object region features using object localization, i.e., bounding box. An object-aware cross-modal projection module can be implemented with minimal modifications by plugging the proposed Q-Perturbation module just before the cross-attention module. In the following, we first describe our Q-Perturbation module for the single object case and then extend it to multiple objects.

4.3 Q-Perturbation Module for Single Objects

The proposed Q-Perturbation consists of three components: 1) Object Key Pooling, 2) K-Subspace Construction, and 3) Query Enhancement as shown in Fig. 5. Let $\mathbf{Q} = \{\mathbf{q}_i\}$, $\mathbf{K} = \{\mathbf{k}_l\}$, and $\mathbf{V} = \{\mathbf{v}_l\}$ be the set of queries before the cross-attention, and the set of the keys and values from the image encoder, respectively. Here, i and l are subscripts to distinguish each token.

1) Object Key Pooling. First, we select image tokens that overlap the detected bounding box obtained by object detection, as shown in Fig. 5. Our pooling step is in the same manner as ROI pooling in two-stage object detection [44]. In the following, those selected image tokens are called object image tokens, and the

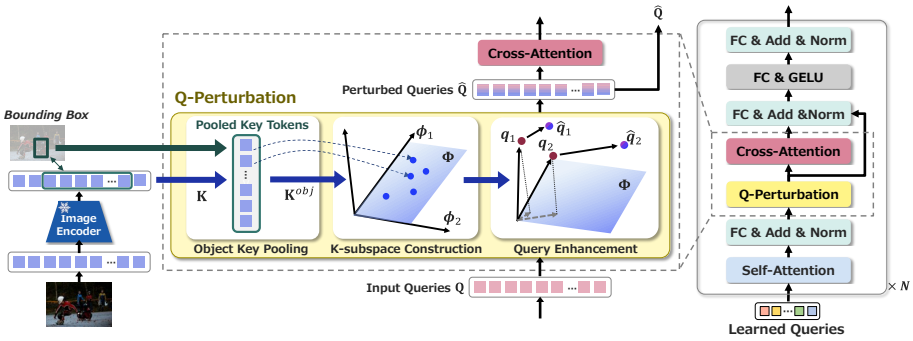


Fig. 5: Overview of our Q-Perturbation. The Q-Perturbation consists of three components: Object Key Pooling, K-subspace Construction, and Query Enhancement.

pooled set is denoted as $\mathbf{K}^{obj} = \{\mathbf{k}_j^{obj}\}$, where j is subscripts to distinguish each token. Note that, if there are multiple objects in an image, we perform the object key pooling for each detected object.

2) K-Subspace Construction. Next, the object-aware key subspace, called *K-subspace*, is generated from the pooled object image tokens $\mathbf{k}_j^{obj}$ for each object using Principal Component Analysis (for more detail, see supplementary material A.2). The K-subspace for each object is denoted by $\mathbf{\Phi} = [\phi_1, \phi_2, \cdots, \phi_p, \cdots]$, where ϕ_p is the p-th basis vector of the K-subspace. The K-subspace represents essential information about the corresponding object.

3) Query Enhancement. Finally, the set of the already obtained query $\mathbf{Q} = \{\mathbf{q}_i\}$ is enhanced by decomposing each query into the K-subspace $\mathbf{\Phi}$ and the complementary subspace using each basis vector ϕ_p of the K-subspace.

$$\mathbf{q}_i = \mathbf{q}_i^{\parallel} + \mathbf{q}_i^{\perp}, \quad \mathbf{q}_i^{\parallel} = \mathbf{\Phi}\mathbf{\Phi}^T\mathbf{q}_i, \quad \mathbf{q}_i^{\perp} = (\mathbf{I} - \mathbf{\Phi}\mathbf{\Phi}^T)\mathbf{q}_i, \tag{1}$$

here, $\mathbf{q}_i^{\parallel}$ and $\mathbf{q}_i^{\perp}$ are the decomposed queries based on the K-subspace $\mathbf{\Phi}$ and its complementary subspace.

Our Q-Perturbation generates a perturbed query $\hat{\mathbf{q}}_i$ that enhances the components belonging to the K-subspace, i.e., enhances object information retained in the original query. This enhancement by the Q-Perturbation is given by

$$\hat{\mathbf{q}}_i = \mathbf{q}_i + \alpha\mathbf{q}_i^{\parallel}, \tag{2}$$

where α is the parameter that control the perturbation magnitude.

This enhancement could be seen in that relevant queries to an object are automatically selected and then enhanced to have more object information, as the decomposition with the K-subspace has a mechanism to extract only object information retained in the original queries. Thus, Q-Perturbation enhances object awareness of V&L models without destroying the weights and structures of the V&L models, resulting in high inheritability.

4.4 Extension to Multiple Objects

So far, we have discussed Q-Perturbation on a single object. Our proposed Q-Perturbation can be easily extended to the case of multiple objects, i.e., it has high flexibility. For Q-Perturbation with multiple objects, the proposed method generates the object-aware K-subspace for each object, and then decomposition and enhancement for each query are performed based on these obtained K-subspaces.

Let $\mathbf{\Phi}_b$, $\mathbf{q}_{i,b}^{\parallel}$, and $\mathbf{q}_{i,b}^{\perp}$ be the K-subspace corresponding to b-th object, query components belonging to the b-th K-subspace, and query components orthogonal to the b-th K-subspace, respectively. Here, b is a subscript to distinguish each detected object in an image. Formally, Q-Perturbation for multiple K-subspaces can be expressed as follows.

$$\hat{\mathbf{q}}_i = \mathbf{q}_i + \alpha \sum_b w(S_b)\mathbf{q}_{i,b}^{\parallel} = \mathbf{q}_i + \alpha \sum_b w(S_b)\mathbf{\Phi}_b\mathbf{\Phi}_b^T\mathbf{q}_i, \tag{3}$$

where S_b and $w(S_b)$ are the area of the detected bounding box and the weight function for b-th detected object.

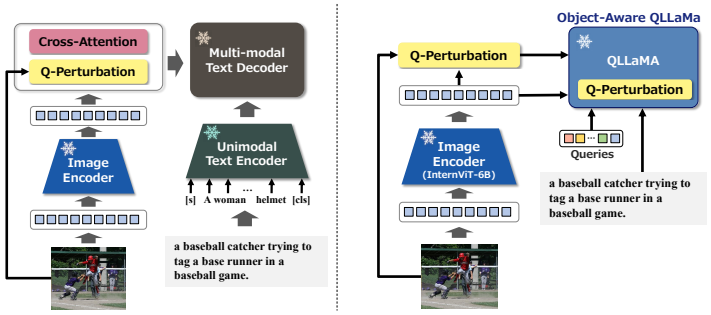
In this paper, for simplicity, the weight function is given by $w(\bar{S}_b) = \beta + \gamma\bar{S}_b$, where β, γ and $\bar{S}$ are the adjustment parameter and normalized area, respectively. The normalized area $\bar{S}_b = S_b/S_I$ is calculated by dividing the whole area of each bounding box by the area S_I of the corresponding image. Note that, in this paper, no exhaustive search was performed for β and γ, and either $\{0, \pm 1, \pm 0.5\}$ was used.

4.5 Beyond the Q-Perturbation Module for Q-Former

In previous sections, our Q-Perturbation has been described for the case of Q-Former-based model [35]. Finally, we discuss its extension to other V&L models and its potential for tasks other than cross-modal image-text retrieval.

Extension to other pre-trained V&L models. Many pre-trained V&L models for cross-modal image-text retrieval have been proposed. It is expected that many more pre-trained models will be proposed in the future. Our proposed Q-Perturbation module is a general and versatile approach to perturb query in cross-attention using the localization cues from object detection and the obtained key features. In this sense, the proposed Q-Perturbation is applicable to other existing V&L models such as COCA [60] and InternVL [6], as shown in Fig. 6. As discussed later, the proposed method with the existing V&L models allows for cross-modal image-text retrieval that is aware of smaller objects.

Other Tasks with Our Q-Perturbation. In general, the Q-Former proposed in BLIP2 is used for other tasks, such as image captioning by using LLMs in the latter additional stage. In this sense, our proposed Q-Perturbation can be used to re-present the existing captions more object-aware. This extension suggests new possibilities for using pre-trained V&L models based on human perception.



(a) COCA with the Q-Perturbation. **(b)** InternVL with the Q-Perturbation.

Fig. 6: Extension to other pre-trained V&L models. Our method is applicable to modern V&L models such as COCA [60] and InternVL [6].

5 Experiments

5.1 Settings

Datasets and Experimental Protocols. We use two widely used benchmark datasets, i.e., Flickr-30K [59] and MSCOCO [4, 39], and fine-grained extensions of the two datasets, i.e., Flickr-FG and COCO-FG [8]. We adapt the commonly used Karapathy split [27] for all the datasets.

The Flickr 30K dataset has 1,014 validation images and 1,000 test images. *The MSCOCO* dataset has 5,000 images for validation and test, respectively. Each image has five description texts. Therefore, there are 5,000 (=1,000 images $\times$ 5 texts) and 25,000 (=5,000 images $\times$ 5 texts) test image-text pairs for Flickr-30K and MSCOCO, respectively.

Flickr-FG and COCO-FG are extensions of the above two datasets. Flickr-FG and COCO-FG replaced original description texts with fine-grained descriptions. These datasets have five fine-grained texts for an image, as with Flickr-30K and MSCOCO. Therefore, there are 5,000 and 25,000 test image-text pairs, as well.

We conducted text-to-image (T2I) and image-to-text (I2T) tasks; the T2I task is to find the paired image of an input text, and the I2T task is vice versa.

Evaluation Metrics. We use the standard evaluation metric, Recall@K (R@K). R@K is a ratio of correct retrievals to all retrievals. Here, the correct retrieval is identified by whether the paired image or texts are in the top K retrieval results. Following the previous studies, K is set to 1, 5, 10.

We also use mean Recall@K (mR@K), which considers the size of objects in each image. mR@K is a harmonic mean of multiple R@Ks. As outlined in Fig. 2 and elaborated on later, the retrieval difficulty depends on the object size in the image. However, traditional R@K does not consider object size and cannot correctly assess this challenge. To alleviate this problem, we propose to use an object size-aware evaluation metric, mean R@K, which is a harmonic mean of R@Ks on subsets split by object size. Here, each R@K is calculated on a subset of

all text-image pairs, where each subset is determined by the largest normalized area $\bar{S}$ (please see Sec. 4.4) of detected objects in each image. In this paper, we generate ten subsets by splitting largest areas by every 10% and calculate a harmonic mean (mR@K) of the ten R@Ks calculated on the subsets.

Implementation Details. We use bounding boxes given by Flickr-Entities [42] and COCO-Entities [11]. Flickr-Entities and COCO-Entities have bounding boxes corresponding to each text. Flickr-Entities generates boxes manually by annotators, and COCO-Entities generates boxes semi-automatically, i.e., boxes are detected by Faster R-CNN [44] and matched nouns in each text by manually defined rules. We also use boxes that are detected automatically by CO-DINO [69]. Note that, object detection and image feature extraction can be carried out in advance for the T2I task, as they do not depend on a retrieval input text.

To split datasets and calculate mR@K, we use Flickr-Entities's bounding boxes for Flickr-30K and Flickr-FG. For COCO and COCO-FG, we use bounding boxes detected by CO-DINO instead of COCO-Entities, as COCO-Entities are built by the traditional detector, Faster R-CNN.

We applied the proposed Q-Perturbation to all cross-attentions in the Q-Former of BLIP2. We used Eva-CLIP [47] base BLIP2 finetuned on COCO validation data [33,35]. Following the previous study [35], we apply the re-ranking technique; we first select 64 candidates by image-text contrastive (ITC) and then re-ranked them by image-text matching (ITM).

We also evaluated our method combined with COCA and InternVL-G models [6,60]. We used the ViT-L/14 [14] based COCA model published by the OpenCLIP repository [21], and the InternVL-14B-224px model [6]. Our Q-Perturbation is applied to the last cross-attention for COCA and to all cross-attentions in the QLLaMA for InternVL-G (see the supplementary material A for more detail).

Q-Perturbation has three hyperparameters: 1) perturbation intensity α, 2) weight function $w(S_b)$, and 3) dimension of object-aware subspaces. These parameters were tuned by the grid-search algorithm using validation data with the mR@1. The intensity α was selected from 2, 4, 6, 8 and 10 for BLIP2 and 0.2, 0.4, 0.6, 0.8 and 1 for COCA and InternVL. The weight function was selected from five functions: constant value (=1), $\bar{S}_b$, $\bar{S}_b - 0.5$, $1 - \bar{S}_b$, $0.5 - \bar{S}_b$. The dimensions of subspaces were determined by using the contribution ratio. Thus, we tuned the threshold of the contribution ratio from 0.85, 0.9, 0.95, 0.99.

5.2 Comparative results

Results for small objects. Table 1 shows the evaluation results for small objects and the overall of each dataset. The proposed Q-Perturbation improves the retrieval performance of small objects, which in turn improves the overall performance. The conventional method would have had difficulty in considering small objects in the image. The results suggest that our method mitigates this difficulty by enhancing object awareness of the conventional methods.

Figure 7 shows examples of image retrieval results by BLIP2-ITC and BLIP2-ITC with Q-Perturbation. We can see that our method selects the correct image

Table 1: Result for the T2I retrieval task. Q-Pert. (E) and (D) are the results of the proposed method using bounding boxes produced by Flickr-/COCO-Entities (E), and detected by CO-DETR (D), respectively. ˜10% means R@1 calculated for data with only tiny objects, i.e., the object area is smaller than 10% to the corresponding image area. The best and second-best are highlighted in bold and underlined, respectively.

	Flickr-30K			Flickr-FG			COCO-5K			COCO-FG		
	˜10%	R@1	mR@1	˜10%	R@1	mR@1	˜10%	R@1	mR@1	˜10%	R@1	mR@1
BLIP2	<u>81.33</u>	89.76	88.75	<u>85.33</u>	92.50	91.79	59.87	68.25	71.09	63.93	72.68	75.71
w/Q-Pert.(E)	**84.00**	<u>89.82</u>	<u>89.10</u>	**86.67**	<u>92.56</u>	<u>91.98</u>	**60.17**	<u>68.34</u>	<u>71.14</u>	<u>64.00</u>	<u>72.75</u>	<u>75.77</u>
w/Q-Pert.(D)	**84.00**	**89.86**	**89.11**	**86.67**	**92.66**	**92.06**	<u>60.09</u>	**68.35**	**71.15**	**64.06**	**72.78**	**75.78**

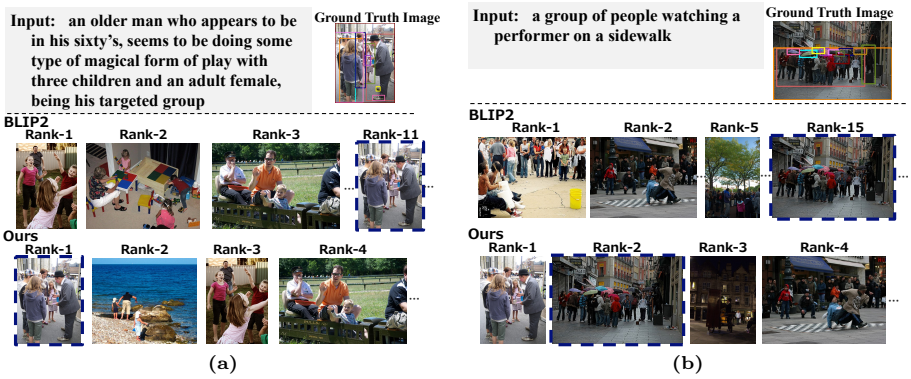


Fig. 7: Examples of image retrieval results by BLIP2 [35] and our method.

at a higher rank than the original BLIP2. This is mainly due to the advantage of our object-aware mechanism; the proposed method efficiently utilizes information from small objects, such as "an older man", and "a performer" for Fig. 7 (a) (b), respectively.

Table 2 shows comparative results with additional evaluation metrics. We can confirm the effectiveness of our method again, as our method shows competitive results. Furthermore, our method stably improves performance even with noisy bounding boxes obtained by a detector automatically. This property is helpful in applying our method to real-world applications.

As we discussed in Sec. 4.5, our Q-Perturbation is applicable to other tasks, such as image captioning, as our method is plugged into a pre-trained V&L model. We carried out image captioning to visualize the effect of Q-Perturbation, as shown in Fig. 8. It can be seen that output captions become object-aware compared with the original BLIP2s' results by emphasizing the corresponding objects, such as "glass", with our Q-Perturbation. This object-aware property of our method helps in improving the retrieval performance of a V&L model.

Overall results. We then discuss performance comparisons between our method and various cross-modal image-text retrieval methods.

Table 2: Results on four datasets for T2I and I2T retrieval tasks. Q-Pert. (E) and (D) are the results using bounding boxes produced by Flickr-/COCO-Entities (E), and detected by CO-DINO (D), respectively. The best and second-best are highlighted in bold and underlined, respectively.

	T2I						I2T					
	R@1	R@5	R@10	mR@1	mR@5	mR@10	R@1	R@5	R@10	mR@1	mR@5	mR@10
Flickr-30K												
BLIP2	89.76	98.18	98.96	88.75	97.85	98.66	97.60	100.00	100.00	97.23	100.00	100.00
w/Q-Pert. (E)	89.82	**98.20**	99.04	89.10	**97.89**	98.60	**98.00**	100.00	100.00	**97.55**	100.00	100.00
w/Q-Pert. (D)	**89.86**	**98.20**	**99.06**	**89.11**	97.86	**98.76**	97.70	100.00	100.00	97.29	100.00	100.00
COCO 5K												
BLIP2	68.25	87.74	**92.67**	71.09	89.82	**94.12**	85.32	**96.94**	**98.38**	87.43	**97.78**	98.86
w/Q-Pert. (E)	68.34	**87.76**	92.63	71.14	**89.84**	94.11	**85.52**	96.90	**98.38**	87.52	97.74	**98.87**
w/Q-Pert. (D)	**68.35**	87.72	92.65	**71.15**	89.79	94.11	85.44	96.88	98.36	**87.69**	97.71	98.85
Flickr-FG												
BLIP2	92.50	99.00	99.46	91.79	98.83	99.45	98.50	100.00	100.00	98.74	100.00	100.00
w/Q-Pert. (E)	92.56	99.00	**99.56**	91.98	**98.95**	**99.53**	**98.6**	100.00	100.00	**98.87**	100.00	100.00
w/Q-Pert. (D)	**92.66**	**99.06**	99.48	**92.06**	98.88	99.47	**98.6**	100.00	100.00	98.85	100.00	100.00
COCO-FG												
BLIP2	72.68	90.22	94.24	75.71	92.17	95.52	87.58	97.64	98.94	89.83	98.42	99.32
w/Q-Pert. (E)	72.75	**90.23**	94.22	75.77	92.17	95.49	**87.82**	97.64	98.92	**89.98**	98.44	**99.37**
w/Q-Pert. (D)	**72.78**	90.22	**94.30**	**75.78**	92.17	**95.58**	87.66	**97.70**	**98.98**	89.87	**98.45**	**99.37**

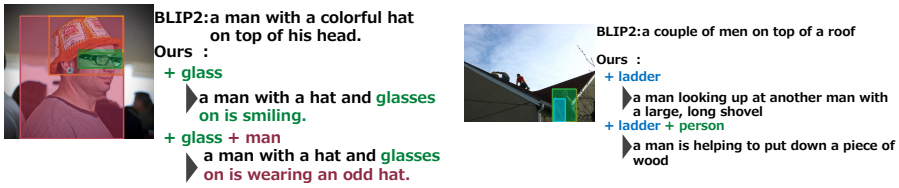


Fig. 8: Image captioning results. "+ word" means enhancing information of the region corresponding to the word.

- **Baselines:** We compared our method with 1) object-aware models; SCAN [32], IMRAM [7], SHAN [24], NAAF [64], and 2) V&L pre-trained models; CLIP [43], ALBEF [37], UNITER [5], BEIT-3 [52], COCA [60], InternVL [6], BLIP2 [35].
- **Results:** Table 3 shows comparative results with various conventional methods. Our method archives competitive results compared with state-of-the-art methods. These results suggest that our Q-Perturbation enhances object awareness of the pre-trained V&L model while inheriting its impressive performance (for more comparison with simple baselines, see supplementary material B).

5.3 Results with Other State-of-The-Art V&L Models

Our proposed Q-Perturbation can be plugged into any model, including cross-attention layers. To confirm the versatility of our method, we apply Q-Perturbation to two V&L models, COCA and InternVL. Table 4 shows comparative results.

Table 3: Comparative results with baselines. The best and second-best are highlighted in bold and underlined, respectively.

	Flickr-30K						COCO 5K					
	T2I			I2T			T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [32]	48.60	77.70	85.20	67.40	90.30	95.80	38.60	69.30	80.40	50.40	82.20	90.00
IMRAM [7]	53.90	79.40	87.20	74.10	93.00	96.60	39.70	69.10	79.80	53.70	83.20	91.00
SHAN [24]	55.30	81.30	88.40	74.60	93.50	96.90	-	-	-	-	-	-
VSE [16]	76.10	94.50	97.10	88.70	98.90	99.80	52.70	80.20	-	68.10	90.20	-
NAAF [64]	61.00	85.30	90.60	81.90	96.10	98.30	42.50	70.90	81.40	58.90	85.20	92.00
CLIP [43]	68.70	90.60	95.20	88.00	98.70	99.40	-	-	-	-	-	-
ALBEF [37]	82.80	96.30	98.10	94.10	99.50	99.70	60.70	84.30	90.50	77.60	94.30	97.20
UNITER [5]	68.70	89.20	93.90	83.60	95.70	97.70	52.90	79.90	88.00	65.70	88.60	93.80
BEIT-3 [52]	81.50	95.60	97.80	94.90	99.90	**100.00**	67.20	87.70	**92.80**	84.80	96.50	98.30
COCA [60]	77.64	94.00	96.48	90.90	98.20	99.20	-	-	-	-	-	-
InternVL-G [6]	84.78	97.02	98.38	94.90	99.90	**100.00**	-	-	-	-	-	-
BLIP2 [35]	89.76	98.18	98.96	97.60	**100.00**	**100.00**	68.25	87.74	92.67	85.32	**96.94**	**98.38**
w/Q-Pert.(E)	89.82	**98.20**	99.04	**98.00**	**100.00**	**100.00**	68.34	**87.76**	92.63	**85.52**	96.90	**98.38**
w/Q-Pert.(D)	**89.86**	**98.20**	**99.06**	97.70	**100.00**	**100.00**	**68.35**	87.72	92.65	85.44	96.88	98.36

Table 4: Results with the method of Q-perturbation with two V&Ls.

	Flickr-30K						Flickr-FG					
	T2I			I2T			T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
COCA	77.64	**94.00**	96.48	90.90	98.20	99.20	82.84	96.10	**97.90**	93.00	99.10	99.70
w/Q-Pert.	**77.68**	93.98	**96.50**	**91.10**	98.20	99.20	**82.88**	**96.12**	97.86	**93.20**	99.10	99.70
InternVL-G	84.78	97.02	98.38	94.90	99.90	100.00	89.04	98.22	99.14	97.10	99.90	100.00
w/Q-Pert.	**84.82**	97.02	**98.42**	**95.40**	99.90	100.00	**89.10**	**98.26**	99.14	**97.20**	99.90	100.00
	mR@1	mR@5	mR@10	mR@1	mR@5	mR@10	mR@1	mR@5	mR@10	mR@1	mR@5	mR@10
COCA	75.86	92.76	95.43	88.89	96.95	99.34	81.78	94.92	**97.47**	93.08	98.57	99.77
w/Q-Pert.	**75.91**	**92.77**	**95.45**	**89.03**	96.95	99.34	**81.83**	**94.95**	97.44	**93.16**	98.57	99.77
InternVL-G	83.12	96.67	98.05	94.35	99.89	100.00	87.68	97.98	99.07	97.39	99.83	100.00
w/Q-Pert.	**83.15**	96.67	**98.10**	**94.75**	99.89	100.00	**87.72**	**98.02**	99.07	**97.46**	99.83	100.00

We can see that Q-Perturbation is highly versatile as our method improves retrieval performance. COCA w/Q-pert. has a slight improvement. This is because the impact of the proposed method is small, as a cross-attention layer is only placed at the end of the vision encoder. Input features to QLLaMA, which is the cross-modal projector of InternVL, or cross-attentions may focus on the global context, as QLLaMA is followed by the large vision encoder (6B parameters ViT model). Even in such a difficult situation, the proposed method could enhance object-awareness and improve performance.

5.4 Sensitivity on hyperparameters

We analyze the sensitivity of the performance to the hyperparameters, including the weight function, scale factor, and dimension of subspaces, i.e., threshold of contribution ratio for PCA. In this experiment, we use bounding boxes by Flickr-Entities. Table 5 shows the evaluation results with varying the hyperparameters. The proposed method has a high performance by adequately selecting

Table 5: Sensitivity on hyperparameters, i.e., the weight function, scale factor, and dimension of subspace.

(a) Flickr-30K datasets. mR@1.

	scale				
	2	4	6	8	10
$\bar{S}_b$	**89.76**	89.56	89.68	89.72	89.66
1-$\bar{S}_b$	89.74	89.52	89.68	**89.78**	89.68
$\bar{S}_b$-0.5	89.72	89.56	89.76	89.50	89.52
0.5-$\bar{S}_b$	89.72	**89.84**	89.66	89.72	**89.70**
constant	89.66	89.74	**89.82**	89.72	89.44

(b) Flickr-30K datasets. mR@1.

	Contribution ratio				
	0.80	0.85	0.90	0.95	0.99
$\bar{S}_b$	89.64	89.68	89.76	89.62	89.66
1-$\bar{S}_b$	**89.94**	**90.00**	89.62	**89.82**	89.68
$\bar{S}_b$-0.5	89.84	89.58	89.54	89.68	89.52
0.5-$\bar{S}_b$	89.80	89.86	89.74	89.74	**89.70**
constant	89.56	89.68	**89.82**	89.64	89.44

the parameters, although the proposed method has low sensitivities to the hyperparameters. In this paper, we selected the hyperparameters from the manually set values using validation data. It would be an excellent direction to learn the hyperparameters in the future if learning data were available.

6 Limitations

The key idea behind our method is to enhance object information in a V&L model's cross-attention features following an image encoder. It would, therefore, be problematic if object information had disappeared entirely in the image encoder.

This paper uses all bounding boxes in each image. It is an excellent future direction to filter bounding boxes or adjust the scale α according to an input text to improve retrieval performance further. Note that this approach increases computational cost at the retrieval stage, as we need to extract bounding boxes and process a neural network, including cross-attentions, after receiving the input. This direction would be a trade-off between scalability and retrieval performance.

7 Conclusions

In this paper, we proposed an object-aware query perturbation for cross-modal image-text retrieval. The key is to use query perturbation to focus on small target objects by enhancing the query weights with keys corresponding to object regions in the cross-attention module. The proposed method is applicable to various V&L models based on cross-modal projection, including COCA, BLIP2, and InternVL. Comprehensive experiments on four public datasets demonstrate the effectiveness of the proposed method.

References

1. Cao, Min and Li, Shiping and Li, Juntao and Nie, Liqiang and Zhang, Min: Image-Text Retrieval: A Survey on Recent Research and Development. In: IJCAI. pp. 5410–5417 (2022)

2. Chen, F., Shao, J., Zhang, Y., Xu, X., Shen, H.T.: Interclass-relativity-adaptive metric learning for cross-modal matching and beyond. IEEE TMM **23**, 3073–3084 (2020)
3. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: CVPR. pp. 15789–15798 (2021)
4. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server (2015)
5. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023)
7. Chen, Hui and Ding, Guiguang and Liu, Xudong and Lin, Zijia and Liu, Ji and Han, Jungong: IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In: CVPR. pp. 12652–12660 (2020)
8. Chen, Weijing and Yao, Linli and Jin, Qin: Rethinking Benchmarks for Cross-modal Image-text Retrieval. In: SIGIR. pp. 1241–1251 (2023)
9. Chen, Yuxin and Ma, Zongyang and Zhang, Ziqi and Qi, Zhongang and Yuan, Chunfeng and Shan, Ying and Li, Bing and Hu, Weiming and Qie, Xiaohu and Wu, Jianping: ViLEM: Visual-Language Error Modeling for Image-Text Retrieval. In: CVPR. pp. 11018–11027 (2023)
10. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: CVPR. pp. 8415–8424 (2021)
11. Cornia, M., Baraldi, L., Cucchiara, R.: Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In: CVPR (2019)
12. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: AAAI. vol. 35, pp. 1218–1226 (2021)
13. Dong, Xinfeng and Han, Longfei and Zhang, Dingwen and Liu, Li and Han, Junwei and Zhang, Huaxiang: Giving Text More Imagination Space for Image-text Matching. In: ACM MM. pp. 6359–6368 (2023)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR (2021)
15. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization. In: CVPR. pp. 3984–3993 (2018)
16. Faghri, Fartash and Fleet, David J. and Kiros, Jamie Ryan and Fidler, Sanja: VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In: BMVC. p. (2017)
17. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: CVPR. pp. 14105–14115 (2022)
18. Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: CVPR. pp. 7181–7189 (2018)
19. Guo, Zixin and Wang, Tzu-Jui Julius and Pehlivan, Selen and Radman, Abduljalil and Laaksonen, Jorma: PiTL: Cross-modal Retrieval with Weakly-supervised Vision-language Pre-training via Prompting. In: SIGIR. pp. 2261–2265 (2023)
20. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal lstm. In: CVPR. pp. 2310–2318 (2017)

21. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (Jul 2021), https://doi.org/10.5281/zenodo.5143773
22. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. IJCAI (2021)
23. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. In: CVPR. pp. 5754–5763 (2019)
24. Ji, Zhong and Chen, Kexin and Wang, Haoran: Step-Wise Hierarchical Alignment Network for Image-Text Matching. In: IJCAI. pp. 765–771 (2021)
25. Jia, Chao and Yang, Yinfei and Xia, Ye and Chen, Yi-Ting and Parekh, Zarana and Pham, Hieu and Le, Quoc and Sung, Yun-Hsuan and Li, Zhen and Duerig, Tom: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: ICML. pp. 4904–4916 (2021)
26. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: CVPR. pp. 2787–2797 (2023)
27. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137 (2015)
28. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping **27** (2014)
29. Kaur, Parminder and Pannu, Husanbir Singh and Malhi, Avleen Kaur: Comparative Analysis on Cross-Modal Information Retrieval: A Review. Computer Science Review **39**, 100336 (2021)
30. Kim, Dongwon and Kim, Namyup and Kwak, Suha: Improving Cross-Modal Retrieval With Set of Diverse Embeddings. In: CVPR. pp. 23422–23431 (2023)
31. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
32. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV. pp. 201–216 (2018)
33. Li, D., Li, J., Le, H., Wang, G., Savarese, S., Hoi, S.C.: LAVIS: A one-stop library for language-vision intelligence. In: ACL. pp. 31–41 (Jul 2023)
34. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: ICCV. pp. 4654–4662 (2019)
35. Li, Junnan and Li, Dongxu and Savarese, Silvio and Hoi, Steven: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: ICML. pp. 19730–19742 (2023)
36. Li, Junnan and Li, Dongxu and Xiong, Caiming and Hoi, Steven: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML. pp. 12888–12900 (2022)
37. Li, Junnan and Selvaraju, Ramprasaath and Gotmare, Akhilesh and Joty, Shafiq and Xiong, Caiming and Hoi, Steven Chu Hong: Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In: NeurIPS. pp. 9694–9705 (2021)
38. Liao, Lizi and Ma, Yunshan and He, Xiangnan and Hong, Richang and Chua, Tat-Seng: Knowledge-Aware Multimodal Dialogue Systems. In: ACM MM. pp. 801–809 (2018)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
40. Liu, C., Mao, Z., Liu, A.A., Zhang, T., Wang, B., Zhang, Y.: Focus your attention: A bidirectional focal attention network for image-text matching. In: ACM MM. pp. 3–11 (2019)

41. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: CVPR. pp. 10921–10930 (2020)
42. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. IJCV **123**(1), 74–93 (2017)
43. Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya: Learning Transferable Visual Models From Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28** (2015)
45. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. arXiv preprint arXiv:2304.06712 (2023)
46. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: CVPR. pp. 1979–1988 (2019)
47. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
48. Thomas, C., Kovashka, A.: Preserving semantic neighborhoods for robust cross-modal retrieval. In: ECCV. pp. 317–335. Springer (2020)
49. Upadhyay, Uddeshya and Karthik, Shyamgopal and Mancini, Massimiliano and Akata, Zeynep: ProbVLM: Probabilistic Adapter for Frozen Vison-Language Models. In: ICCV. pp. 1899–1910 (2023)
50. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: ECCV. pp. 18–34. Springer (2020)
51. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE TPAMI **41**(2), 394–407 (2018)
52. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. CVPR pp. 19175–19186 (2023)
53. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval. In: ICCV. pp. 5764–5773 (2019)
54. Wang, Zheng and Gao, Zhenwei and Guo, Kangshuai and Yang, Yang and Wang, Xiaoming and Shen, Heng Tao: Multilateral Semantic Relations Modeling for Image Text Retrieval. In: CVPR. pp. 2830–2839 (2023)
55. Wei, J., Xu, X., Wang, Z., Wang, G.: Meta self-paced learning for cross-modal matching. In: ACM MM. pp. 3835–3843 (2021)
56. Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., Shen, H.T.: Universal weighting metric learning for cross-modal matching. In: CVPR. pp. 13005–13014 (2020)
57. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: CVPR. pp. 10941–10950 (2020)
58. Yan, S., Yu, L., Xie, Y.: Discrete-continuous action space policy gradient-based attention for image-text matching. In: CVPR. pp. 8096–8105 (2021)
59. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
60. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. TMLR (2022)
61. Yuan, J., Chen, T., Li, B., Xue, X.: Compositional scene representation learning via reconstruction: A survey. IEEE TPAMI (2023)

62. Zeng, Sheng and Liu, Changhong and Zhou, Jun and Chen, Yong and Jiang, Aiwen and Li, Hanxi: Learning Hierarchical Semantic Correspondences for Cross-Modal Image-Text Retrieval. In: ACM MM. pp. 239–248 (2022)
63. Zhang, H., Mao, Z., Zhang, K., Zhang, Y.: Show your faith: Cross-modal confidence-aware network for image-text matching. In: AAAI. vol. 36, pp. 3262–3270 (2022)
64. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: CVPR. pp. 15661–15670 (2022)
65. Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z.: Context-aware attention network for image-text retrieval. In: CVPR. pp. 3536–3545 (2020)
66. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: ECCV. pp. 686–701 (2018)
67. Zheng, X., Ji, R., Sun, X., Zhang, B., Wu, Y., Huang, F.: Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In: AAAI. vol. 33, pp. 9291–9298 (2019)
68. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16793–16803 (2022)
69. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: ICCV. pp. 6748–6758 (2023)