ELIP: Enhanced Visual-Language Foundation Models for Image Retrieval

Guanqi Zhan[1*], Yuanpei Liu[2*], Kai Han[2], Weidi Xie[1,3], Andrew Zisserman[1]
[1]*VGG, University of Oxford* [2]*The University of Hong Kong* [3]*Shanghai Jiao Tong University*
{guanqi,weidi,az}@robots.ox.ac.uk ypliu0@connect.hku.hk kaihanx@hku.hk

***Abstract*—The objective in this paper is to improve the performance of text-to-image retrieval. To this end, we introduce a new framework that can boost the performance of large-scale pre-trained vision-language models, so that they can be used for text-to-image re-ranking. The approach, Enhanced Language-Image Pre-training (ELIP), uses the text query, via a simple MLP mapping network, to predict a set of visual prompts to condition the ViT image encoding. ELIP can easily be applied to the commonly used CLIP, SigLIP and BLIP-2 networks. On the evaluation side, we set up two new out-of-distribution (OOD) benchmarks, *Occluded COCO* and *ImageNet-R*, to assess the zero-shot generalisation of the models to different domains. The results demonstrate that ELIP *significantly* boosts CLIP/SigLIP/SigLIP-2 text-to-image retrieval performance and outperforms BLIP-2 on several benchmarks, as well as providing an easy means to adapt to OOD datasets.**

***Index Terms*—image-text retrieval, vision-language models**

I. Introduction

This paper considers the problem of text-to-image retrieval, that aims to rank image instances based on their relevance to a text query. Effective retrieval generally includes two stages: the first stage provides an initial ranking in a fast and efficient manner, while the second *re-ranking* stage refines this ranking by re-computing the relevance scores between the text query and each of the top-ranked candidates with a more expensive model.

Recent advances in text-to-image retrieval have primarily focused on the first stage. Notable models, such as CLIP [1] and ALIGN [2], leverage contrastive learning [3] on large-scale image-text pairs to learn joint representations, demonstrating impressive generalization capabilities for cross-modal retrieval tasks.

Our primary contribution here focuses on the second stage of the retrieval pipeline, namely, the re-ranking. Specifically, our goal is to enhance the performance of off-the-shelf vision-language foundation models, so that they can be re-purposed for re-ranking the top-k candidates from the fast retrieval process. The approach we develop, termed *Enhanced Language-Image Pre-training* (**ELIP**), requires only a few trainable parameters, and the training can be conducted efficiently with 'student-friendly' resources and data. We demonstrate that ELIP can boost the performance of the pre-trained CLIP [1], SigLIP [4], SigLIP-2 [5], and BLIP-2 [6] for cross-modal retrieval.

*Equal contribution.

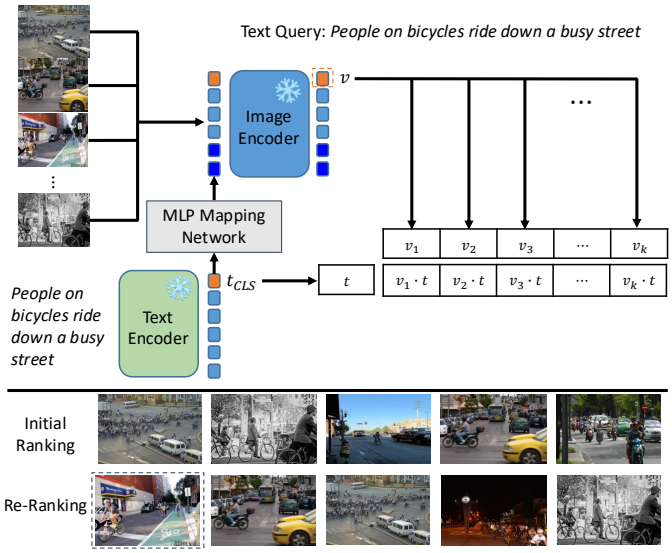

Fig. 1. **The ELIP architecture.** *Top*: We propose a novel architecture that can be applied to pre-trained and frozen vision-language foundation models, such as CLIP, SigLIP, SigLIP-2 and BLIP-2, to enhance their text-to-image retrieval performance. The *key idea* is to use the text query to define a set of visual prompt vectors that are incorporated into the image encoder to make it aware of the query when generating the embedding. An MLP maps from the text space to the visual space of the input to the ViT encoder. The architecture is lightweight, and our data curation strategies enable efficient and effective training with limited resources. *Bottom*: In this retrieval example from the COCO benchmark, the top-k (k=100) images are re-ranked by the ELIP model for the text query: 'People on bicycles ride down a busy street'. The ground truth image matching the query is not in the top-5 ranked images in the initial CLIP ranking, but is ranked top-1 (highlighted in the dashed box) by the re-ranking.

To achieve this goal, we first introduce a lightweight, text-guided visual prompting module. As illustrated in Figure 1, a query text is mapped to a set of visual prompt vectors [7], that are then concatenated with the `[CLS]` and patch embeddings of the image encoder. These augmented embeddings are then passed into the frozen vision encoder to recompute the image representation. The resulting image embedding is aware of the text conditioning and this enhances its performance in re-ranking.

To assess the re-ranking performance of our proposed ELIP models, we experiment on the standard COCO [8] and Flickr30k [9] text-to-image retrieval benchmarks. As a further challenge, we also evaluate the generalisation of the ELIP-boosted models on out-of-distribution domains. To do so, we repurpose the Occluded COCO [10] and ImageNet-R [11] datasets to be used for text-to-image retrieval benchmarks.

In summary, we make three contributions: *First*, we propose a novel architecture to improve text-based image retrieval on large pre-trained vision-language models, including the most popular CLIP/SigLIP architectures and the state-of-the-art BLIP-2 architecture. *Second*, to evaluate the generalisation capability of text-to-image retrieval models to different out-of-distribution domains, we set up two new benchmarks of text-to-image retrieval, *Occluded COCO* and *ImageNet-R*. *Third*, and most significantly, we demonstrate that ELIP *substantially* improves the image retrieval performance of CLIP and SigLIP architectures, and outperforms the state-of-the-art BLIP-2 architecture. Furthermore, it provides an efficient method to adapt these architectures to OOD datasets, again giving a tremendous boost with CLIP, SigLIP, SigLIP-2 and BLIP-2. As an additional contribution, we show that the model can be trained efficiently with limited computing resources, and develop a 'student-friendly' *best practice*, involving global hard sample mining and batch selection. This training is detailed in the arXiv version of this paper [12].

II. Related Work

Text-to-Image Retrieval is a fundamental and much researched task in cross-modal learning [1], [13]–[49]. Large vision language models, such as CLIP [1], [50], ALIGN [2], BLIP-2 [6], SigLIP [4] and SigLIP-2 [5] that have powerful zero-shot capabilities have now become the de facto method for open-set text-based image retrieval. The most recent work [51] gives a slight improvement over BLIP-2 by incorporating the output of an object detector or annotations of detection bounding boxes. This succeeds in overcoming the failure cases where small but semantically important objects in an image are not properly understood by the model. We compare to this model and show superior performance.

CIR and Universal Retrieval. In composed image retrieval (CIR) [52]–[55], the query is specified by a composition of an image and text, with the text specifying how the image should be changed. For example, the query image may be of a dog lying down, and the query text may be 'playing with a ball'. This composed query defines the target image to be retrieved from the gallery. This differs from our task, where the query is specified only by text, and the text alone defines the target image to be retrieved from the gallery. A more general setting is 'universal retrieval' [56], [57] where the query can be a combination of image, text, and instruction; and the target can be image alone, text alone, or image and text.

Post-Retrieval Re-ranking. For single modality image retrieval, where the query is an image, there has been a series of works that have re-ranked the top-k images from an initial ranking via classical computer vision algorithms, such as 'query expansion', 'geometric verification', or a combination of the two [58]–[63], as well as via learning-based algorithms [64]–[67]. Re-ranking algorithms have been relatively less explored in text-to-image retrieval [68]–[70]. [71] introduced a method for computing the similarity score between an image and a text query by estimating the log-likelihood of the text conditioned on the image. While this approach has demonstrated strong performance, it remains computationally expensive both during training and inference, making it a *slow* process. Our paper also focuses on the re-ranking stage – developing a more powerful version of visual-language foundation models to give a better ranking of images that are hard to distinguish by the original retrieval model.

Multi-Modal Datasets. To obtain multi-modal foundation models with a strong capability of generalisation, it is important to train them on large-scale multi-modal datasets. Therefore, in recent years, there has been a significant increase in the number and scale of multimodal vision-language datasets that provide image-text pairs, such as COCO [8], SBU [72], Conceptual Captions [73], LAION [74], DataComp [75]. The increase in the size of multi-modal datasets enables the training of more powerful visual-language foundation models. More recently, DataCompDR [76] utilises prior knowledge from large-scale pre-trained image captioning models to generate synthetic captions for DataComp images, resulting in less noisy captions than the datasets collected from the web, such as the original DataComp dataset. In this paper, we have experimented with training the model using Conceptual Captions [73] and DataCompDR [76].

III. Preliminaries

Re-Ranking in Image Retrieval. Given an input query, the goal of a retrieval system is to rank all instances in a dataset $\Omega = \{I_1, \ldots, I_n\}$, based on their relevance to the query. In the case of text-to-image retrieval, the query is specified by text (T), and the ideal outcome is a set ($\hat{\Omega}$), with the relevant images being ranked higher than those that are not. In general, an effective retrieval system proceeds in two stages: the first stage provides an initial ranking in a fast and efficient manner, while the second stage—referred to as re-ranking–refines this ranking by recomputing the relevance scores between the text query and each of the top-k ranked candidates with a more powerful (and usually more expensive) ranking model. The k is selected such that in general there is a high recall for all the relevant images. In this paper, our novelty lies in the second stage, that aims to re-rank the top-k candidates from the first stage results.

Visual Prompt Tuning (VPT) [7] is a method of enhancing the ViT image encoder by inserting additional learnable prompts into the transformer layers. It enables efficient adaptation of ViT, requiring only the few parameters of the learnable prompts to be trained. VPT has two different variants – *VPT-Shallow* and *VPT-Deep*. *VPT-Shallow* only inserts the additional visual prompts into the first Transformer layer, whereas for *VPT-Deep*, prompts are introduced at every transformer layer's input space. We insert our generated set of visual prompt vectors into the first transformer layer of ViT, which is similar to *VPT-Shallow*.

IV. The ELIP Architecture

In this section, we describe the ELIP text-to-visual prompt mapping network, that can be efficiently applied to adapt the commonly used CLIP/SigLIP architectures as well as the

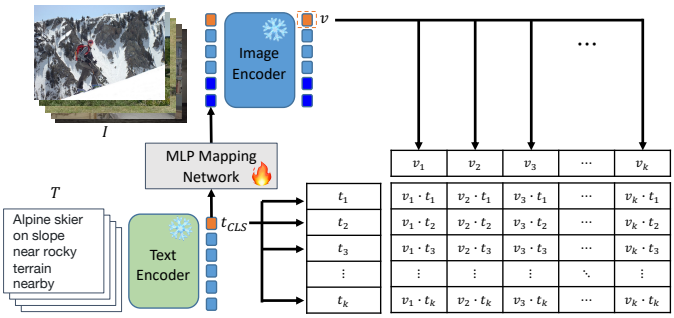

Fig. 2. **Architecture of ELIP-C / ELIP-S.** At training time, a batch of text-image pairs is fed into the architecture. The text feature is mapped to the visual embedding space as a set of prompt vectors via the MLP mapping network and then guides the encoding of the image feature. We use color coding for the [CLS] token, patch tokens, and generated visual tokens from text. The architecture is trained with InfoNCE loss (for ELIP-C) and Sigmoid loss (for ELIP-S/ELIP-S-2), to align the text feature with the corresponding re-computed image feature.

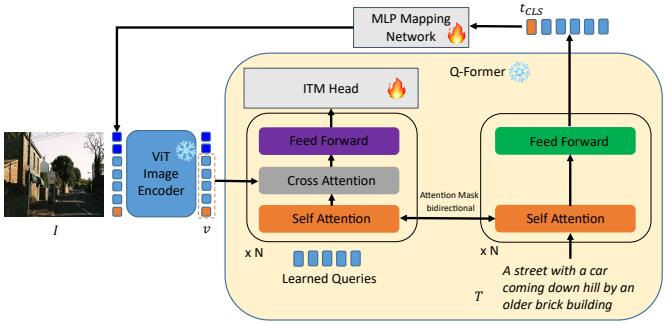

Fig. 3. **Architecture of ELIP-B.** Similar to the architecture on CLIP/SigLIP, the *MLP Mapping Network* maps the text feature to the visual embedding space. The only difference is that the text-guided image features are further fed into the Q-Former to cross-attend the input text and then passed through the Image-Text Matching (ITM) Head to predict whether the image and text match or not. As the input image features to the ITM head have been changed, we also fine-tune the ITM head, which is a lightweight MLP network. The network is fed pairs of text and positive/negative image features at training time and is trained with binary cross entropy loss.

more sophisticated BLIP-2 architectures for re-ranking. We first introduce the architecture of the network in Section IV-A, and the training/inference strategy in Sections IV-B and IV-C respectively. We refer to the network applied to CLIP as *ELIP-C*, applied to SigLIP/SigLIP-2 as *ELIP-S/ELIP-S-2*, and applied to BLIP-2 as *ELIP-B*.

A. Text-Guided MLP Mapping Network

Here, we propose a mapping network that projects the embedding of the text query into a set of prompt vectors within the visual embedding space. This set of prompt vectors is then incorporated as additional tokens into the first layer of the Vision Transformer (ViT) image encoder, used to re-compute the visual embeddings:

$$[t_p^1,\ \ldots,\ t_p^m,\ t_{\texttt{CLS}}] \ = \ \Phi_{\text{t}}(T)$$
$$v \ = \ \Phi_{\text{v}}([x_p^1,\ \ldots,\ x_p^n,\ x_{\texttt{CLS}};\ \psi_{\text{map}}(t_{\texttt{CLS}})])$$

where T denotes the query text, which is first encoded with a pre-trained, frozen text encoder ($\Phi_{\text{t}}(\cdot)$) into $m+1$ embeddings. The [CLS] token is further fed into a **trainable** mapping network to generate the prompt vectors, which are concatenated with the $n+1$ image embeddings ($[x_p^1, \ldots, x_p^n, x_{\texttt{CLS}}]$), and passed into the pre-trained, frozen visual encoder ($\Phi_{\text{v}}(\cdot)$). The MLP Mapping Network consists of 3 layers of linear layers with a GELU between every two linear layers. We expand the output dimension to be n times when we generate n tokens and then divide the generated vector into n tokens. The ELIP architecture is shown in Figure 2 and Figure 3.

B. Training and Testing ELIP-C/ELIP-S

Text-Guided Contrastive Training. At training time, we compute the dot product between the [CLS] token embedding of the text query ($t_{\texttt{CLS}}$) and the re-computed image features guided by the query text, *i.e.*, $\{v_1, \ldots, v_b\}$ (b denotes the batch size). For ELIP-C, we train with the standard InfoNCE loss on the batches; For ELIP-S/ELIP-S-2, we train with pairwise Sigmoid loss. In the arXiv version of this paper [12], we provide more details on the batch selection scheme via global hard sample mining.

Re-Ranking at Inference Time. At inference time, for each text query, we first compute the similarity scores between the visual-language embedding, computed by the original CLIP/SigLIP model, to obtain an initial ranking of all images. We then select the top-k candidates for further re-ranking, where the visual features are re-computed by incorporating the prompted vectors from the mapping network. The final ranking is obtained via the dot product of the re-computed image features and the text feature.

C. Training and Testing ELIP-B

Figure 3 illustrates the application of our architecture on BLIP-2. The only difference with that described for CLIP-type models is that BLIP-2 re-ranking does not use a dual encoder; rather, the image and text encoders attend to each other. However, the purpose of our mapping network and its training are essentially unchanged.

Text-Guided Image-Text Matching Loss. At training time, we feed the text query (T) and the re-computed image features with the query text as prompts, *i.e.*,$\{v_+, v_-\}$ (v_+ denotes the positive image and v_- denotes the negative image), into the Q-Former, and then to an Image-Text Matching (ITM) Head to predict a score indicating whether the text and image match or not. The output of the ITM head is trained with binary cross entropy loss.

Inference Time Re-Ranking. For each text query, we first compute the similarity scores between the visual-language embedding, computed by the original BLIP-2 image and text encoders, to obtain an initial ranking of all images. We then select the top-k candidates for further re-ranking, where the visual features are re-computed by incorporating the prompted vectors from the mapping network. The final ranking is obtained via the sum of the initially computed similarity score and the score predicted by the ITM head based on the re-computed image features and text query.

V. Evaluation Datasets

Here we describe the details of the datasets we use for evaluation.

A. Standard Benchmarks

COCO [8], is a large-scale dataset for studying object detection, segmentation, and captioning. In terms of captioning, each image is annotated with 5 different captions. Previous works use the test split of 5,000 images and 25,010 captions for the evaluation of text-to-image retrieval.

The Flickr30k Dataset [9] contains images collected from Flickr, together with 5 reference sentences provided by human annotators. The test set for text-to-image retrieval consists of 1,000 images and 5,000 captions.

Evaluation Metrics. We adopt the standard metrics for assessing retrieval performance, namely, Recall@1, Recall@5 and Recall@10. Recall@k denotes the proportion of relevant images that are successfully retrieved within the top-k results for each text query.

B. Out-of-Distribution Benchmarks

To evaluate a model's capability for text-to-image retrieval in out-of-distribution (OOD) scenarios, we set up two new benchmarks for text-based image retrieval. Figure 4 shows examples from the Occluded COCO and ImageNet-R benchmarks.

Occluded COCO is curated with annotations from [10], with the method as described in [77], where the occlusion relationship is utilised to collect images containing occluded objects. This dataset aims to evaluate the model's performance on retrieving images with occluded target objects against images that do not contain the target object. It has 80 text queries and 5,000 images.

ImageNet-R is generated using annotations from [11] and aims to examine the model's performance for retrieval across various domains, for example, art, cartoons, deviantart, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video games. It has 200 text queries and 30,000 images.

Evaluation Metrics. Here, we use mAP as the evaluation metric. This is because there might be multiple positive images for each text query.

VI. Experiment

Training the Model. The recent visual-language foundation models are often trained on massive numbers (billions) of paired image-caption samples, with considerable computing resources. Here, we explore a 'resource efficient training' *best practice* for data curation that enables improving large-scale visual-language models with limited resources. Specifically, there are two major challenges to be addressed: (i) training with a large batch size is challenging, due to limitation on GPU memory; (ii) training on billions of samples is prohibitively expensive on computation cost. In the arXiv version of this paper [12], we describe a strategy for global hard sample mining to make the training more effective with a small batch size, and a procedure for selecting and curating an image-text training dataset with maximum information.

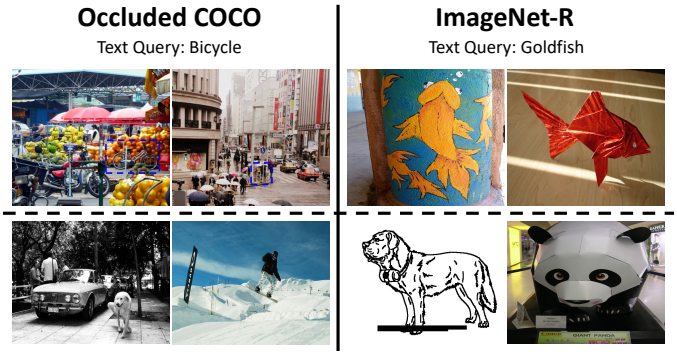

Fig. 4. **Examples of the out-of-distribution benchmarks.** Occluded COCO is on the left, and ImageNet-R is on the right. For both benchmarks, the positive images contain the object described by the text query while the negative images do not contain the object. We display positive images in the first row and negative images in the second row. For Occluded COCO, the target object in the image is occluded, making it more difficult to be retrieved. For example, for the text query *Bicycle* in Occluded COCO, positive images have an occluded bicycle (highlighted in dashed box) while negative images do not have a bicycle in it; for the text query *Goldfish* in ImageNet-R, positive images have goldfish while negative images do not have goldfish.

Implementation Details. Due to computational resource constraints, we train the ELIP-C model with a batch size of 40, the ELIP-S model with a batch size of 10, and the ELIP-B model with a batch size of 12. The initial learning rate is set to 1×10^{-3} for ELIP-C, ELIP-S, and ELIP-S-2, and 1×10^{-5} for ELIP-B. All models are trained on the DataCompDR dataset by default, with additional experiments conducted on the smaller CC3M dataset for ablation studies. Training is performed on two A6000 or A40 GPUs. For re-ranking, we select the top-k samples based on the dataset and model: for ELIP-C, k is set to 100 for COCO and Flickr, 500 for Occluded COCO, and 1000 for ImageNet-R; for ELIP-S and ELIP-S-2, k is set to 100 for COCO and Flickr, 500 for Occluded COCO, and 200 for ImageNet-R; for ELIP-B, k is set to 20 for COCO and Flickr, 100 for Occluded COCO, and 200 for ImageNet-R. The value of k is chosen to ensure high recall in the original ranking while maintaining fast inference. Compared to the original pre-training approach of CLIP, SigLIP and BLIP-2, our method significantly improves training efficiency in terms of reduced training time, GPU requirements, and batch size, with only a marginal increase in FLOPS introduced by the trainable MLP mapping network. Further details are provided in the arXiv version of this paper [12].

A. Results on COCO and Flickr Benchmarks

Ablation Study. In Table I, we evaluate the contributions of different components of the ELIP framework for CLIP. A comparison between Settings $\mathbb{A}$ and $\mathbb{B}$ highlights the effectiveness of the ELIP-C boost over the original CLIP. The comparison between Settings $\mathbb{B}$ and $\mathbb{C}$ demonstrates the importance of hard sample mining when training with a small batch size. Settings $\mathbb{C}$ and $\mathbb{D}$ show the benefit of training on larger datasets with

TABLE I

ABLATION STUDY ON ELIP-C FOR CHOICE OF TRAINING DATASET, HARD SAMPLE MINING, AND NUMBER OF PROMPT VECTORS GENERATED.

Setting	Architecture	Training Dataset	Hard Sample Mining	Multiple Prompts	COCO			Avg.	Flickr			Avg.
					R@1	R@5	R@10		R@1	R@5	R@10	
𝔸	CLIP	-			40.2	66.0	75.6	60.6	67.6	88.3	93.0	83.0
𝔹	ELIP-C	CC3M [73]			40.7	66.2	76.1	61.0	68.8	88.9	93.8	83.8
ℂ	ELIP-C	CC3M [73]	✓		41.8	67.5	77.5	62.3	69.5	89.7	94.1	84.4
𝔻	ELIP-C	DataCompDR [76]	✓		44.2	70.0	79.5	64.6	71.3	90.6	94.4	85.4
𝔼	ELIP-C	DataCompDR [76]	✓	✓	**45.6**	**71.1**	**80.4**	**65.7**	**72.3**	**90.6**	**94.7**	**85.9**

TABLE II

COMPARISON WITH RECENT STATE-OF-THE-ART METHODS. TOP: CLIP-BASED MODELS; MIDDLE: SIGLIP-BASED MODELS; BOTTOM: BLIP-2-BASED MODELS. ELIP-C/ELIP-S BRINGS A SIGNIFICANT **ZERO-SHOT** PERFORMANCE BOOST OF CLIP/SIGLIP ARCHITECTURES, AND ELIP-B OUTPERFORMS THE STATE-OF-THE-ART BLIP-2 MODEL. RESULTS FOR MODELS WITHOUT * ARE ZERO-SHOT, WHEREAS RESULTS FOR MODELS WITH * ARE ONLY ZERO-SHOT ON FLICKR, AS THE BLIP-2 MODEL HAS BEEN FINE-TUNED ON COCO, AND THE * MODELS ARE BASED ON BLIP-2. HOWEVER, OUR METHOD BRINGS AN IMPROVEMENT OVER BLIP-2 ON BOTH BENCHMARKS WHEN TRAINED ON DATACOMPDR.

Model	Year	COCO			Average	Flickr			Average
		Recall@1	Recall@5	Recall@10		Recall@1	Recall@5	Recall@10	
CLIP [1], [50]	2021	40.16	65.95	75.62	60.58	67.56	88.34	93.00	82.97
ELIP-C(Ours)	-	**45.61**	**71.08**	**80.43**	**65.71**	**72.30**	**90.62**	**94.68**	**85.87**
SigLIP [4]	2023	54.21	76.78	84.24	71.74	82.96	96.10	98.04	92.37
ELIP-S(Ours)	-	**61.03**	**82.62**	**88.70**	**77.45**	**87.62**	**98.16**	**99.16**	**94.98**
SigLIP-2 [5]	2025	56.87	78.79	85.49	73.72	83.94	96.62	98.20	92.92
ELIP-S-2(Ours)	-	**62.91**	**83.86**	**89.70**	**78.82**	**87.74**	**97.96**	**98.94**	**94.88**
*BLIP-2** [6]	2023	68.25	87.72	92.63	82.87	89.74	98.18	98.94	95.62
Q-Pert.(E)* [51]	2024	68.34	87.76	92.63	82.91	89.82	98.20	99.04	95.69
Q-Pert.(D)* [51]	2024	68.35	87.72	92.65	82.91	89.86	98.20	99.06	95.71
*ELIP-B(Ours)**	-	**68.41**	**87.88**	**92.78**	**83.02**	**90.08**	**98.34**	**99.22**	**95.88**

less noisy captions. Finally, the comparison between Settings 𝔻 and 𝔼 reveals that generating multiple visual prompts (*e.g.*, 10 prompts in this study) is more beneficial than generating a single prompt. Further ablation studies on the number of generated prompts are detailed in the arXiv version of this paper [12].

Comparison with State-of-the-Art. As shown in Table II, we compare our models (ELIP-C, ELIP-S, ELIP-S-2, and ELIP-B) with prior state-of-the-art methods. When trained on DataCompDR12M, our method demonstrates zero-shot performance improvements for CLIP, SigLIP, SigLIP-2, and BLIP-2 on the COCO and Flickr benchmarks. Notably, ELIP-B outperforms the most recent work [51], establishing a new state-of-the-art for text-to-image retrieval on the BLIP-2 backbone. Furthermore, our ELIP-S, when applied to SigLIP and SigLIP-2, achieves performance comparable to BLIP-2. We have also compared ELIP with several baseline methods for re-ranking in the arXiv version of this paper [12].

Recall Top-k Curves. Figure 5 (right) presents the Recall@Top-k curves for the original CLIP model and our ELIP-C on the COCO benchmark. The curves are generated by plotting the Recall values across various Top-k thresholds. Notably, there is a significant performance gap between the two models, demonstrating that ELIP-C re-ranking consistently improves text-to-image retrieval performance across different k values.

Qualitative Results. Figure 6 provides a qualitative comparison between the initial rankings produced by the CLIP model and the re-ranked results obtained with ELIP-C on the COCO (left) and Flickr (right) benchmarks. In both cases, ELIP-C significantly improves the rankings by elevating the ground truth image (highlighted with a dashed box) to rank 1. Additional qualitative results are provided in the arXiv version of this paper [12].

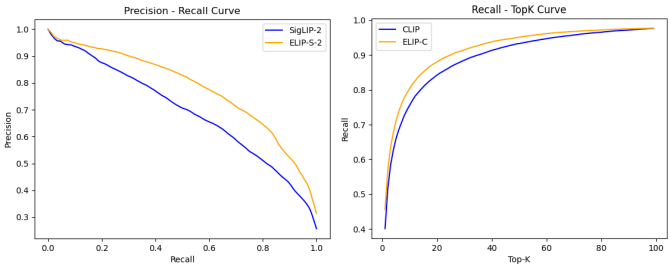

Fig. 5. **Before/after comparisons.** Left: Precision-Recall curves for Occluded COCO retrieval, comparing SigLIP-2 initial rankings to the re-rankings given by ELIP-S-2. Right: Recall Top-k curves for COCO retrieval, comparing CLIP initial rankings to the re-rankings given by ELIP-C.

Visualisation of Attention Map. Figure 7 visualises the cross-attention maps of the `[CLS]` token on patch tokens for both CLIP and ELIP-C on COCO. When the image matches the text query, our generated visual prompt vectors effectively enhance the selection of image features relevant to the query. This improvement can be attributed to ELIP-C's early fusion approach, which integrates text features at the beginning of the image encoder, enabling the model to produce image embeddings more closely aligned with the query text.

Fig. 6. **Qualitative comparison between CLIP initial ranking and ELIP-C re-ranking.** COCO: Columns 1–2; Flickr: Columns 3–4. The ground truth image for each query is highlighted with a dashed box, with the top-3 retrieved images shown. For the COCO query "A large wooden pole with a green street sign hanging from it", CLIP ranks a non-wooden pole as top-1, while ELIP-C correctly re-ranks the large wooden pole to top-1. For the Flickr query "A man wearing bathing trunks is parasailing in the water", CLIP ranks a wakeboarding person as top-1, whereas ELIP-C accurately re-ranks the parasailing man wearing bathing trunks to top-1.

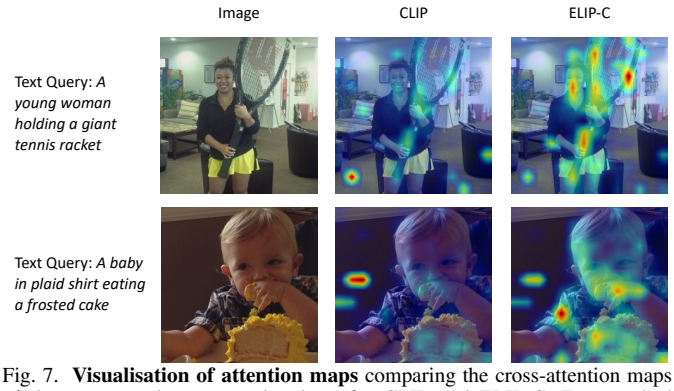

Fig. 7. **Visualisation of attention maps** comparing the cross-attention maps of the `[CLS]` token on patch tokens for CLIP and ELIP-C. For matched queries, ELIP-C enhances attention on image features relevant to the text. For example: (Row 1) ELIP-C focuses more on the giant tennis racket and the young woman for the query "A young woman holding a giant tennis racket"; (Row 2) ELIP-C highlights the cake, baby, and shirt for the query "A baby in plaid shirt eating a frosted cake". Differences are minimal when the image does not match the query (examples provided in the arXiv version of this paper [12]).

The visualisations provide strong evidence supporting this hypothesis.

B. Results on OOD Benchmarks

The results on the out-of-distribution (OOD) benchmarks are presented in Table III. ELIP achieves notable *zero-shot* improvements across all models on the OOD benchmarks, Occluded COCO and ImageNet-R, highlighting the strong generalization capabilities of the ELIP models. The performance can be improved further by fine-tuning the mapping network on suitable datasets (the image and text encoders

TABLE III
MAP RESULTS ON OOD DATASETS. ELIP PROMPTING ACHIEVES NOTABLE ZERO-SHOT IMPROVEMENTS FOR CLIP, SIGLIP SERIES, AND BLIP-2. THESE GAINS ARE FURTHER AMPLIFIED THROUGH FINE-TUNING ON RELEVANT DATASETS. FOR EXAMPLE, TO ADAPT TO THE OCCLUDED COCO, THE ELIP MODEL IS FINE-TUNED ON COCO. SIMILARLY, FINE-TUNING ON IMAGENET ADAPTS ELIP TO IMAGENET-R. THESE RESULTS DEMONSTRATE ELIP'S CAPABILITY FOR EFFICIENTLY ADAPTING THE MODELS TO NEW DATASETS.

Model	Occluded COCO	ImageNet-R	Average
CLIP	47.47	76.01	61.74
ELIP-C (zero-shot)	48.89	76.81	62.85
ELIP-C (fine-tuned)	**59.88**	**81.44**	**70.66**
SigLIP	61.74	92.11	76.93
ELIP-S (zero-shot)	64.58	92.42	78.50
ELIP-S (fine-tuned)	**71.99**	**92.86**	**82.43**
SigLIP-2	66.40	92.66	79.53
ELIP-S-2 (zero-shot)	67.42	92.74	80.08
ELIP-S-2 (fine-tuned)	**76.10**	**94.00**	**85.05**
BLIP-2	62.73	82.31	72.52
ELIP-B (zero-shot)	63.40	82.99	73.20
ELIP-B (fine-tuned)	**70.49**	**83.68**	**77.09**

are frozen). Since it is not feasible to fine-tune on Occluded COCO (very few data samples) and ImageNet-R (evaluation only), for Occluded COCO retrieval, we fine-tune on the original COCO dataset, and for ImageNet-R retrieval, we fine-tune on ImageNet. As can be seen in Table III by this fine-tuning the performance of all the models is significantly boosted further. This demonstrates that fine-tuning ELIP enables efficient adaptation of the models to new datasets. The significant difference ELIP makes is also illustrated in Figure 5 (left). Please refer to the arXiv version of this paper [12] on the fine-tuning.

VII. CONCLUSION

In this paper, we introduced *Enhance Language-Image Pre-training (ELIP)*, a method to improve visual-language foundation models for text-to-image retrieval. ELIP is a simple plug-and-play modification to pre-trained visual-language foundation models that significantly improves their zero-shot performance. Furthermore, the mapping network can be fine-tuned to efficiently adapt these models to OOD datasets, leading to still further improvements. We have also demonstrated, by visualising the attention maps, that ELIP enables the image encoder to attend to more relevant details. Future work could apply ideas similar to ELIP to enhance generative Multimodal Large Language Models by introducing more effective text-guided visual attention and encoding for both decoder-only [78] and cross-attention-based [79] architectures. Please refer to the arXiv version of this paper [12] for more details and future updates.

Acknowledgements. This research is supported by EPSRC Programme Grant VisualAI EP/T028572/1, a Royal Society Research Professorship RP\R1\191132, a China Oxford Scholarship and the Hong Kong Research Grants Council – General Research Fund (Grant No.: 17211024).

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[3] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[4] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023.

[5] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[7] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[9] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[10] H. Lee and J. Park, "Instance-wise occlusion and depth orders in natural scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[12] G. Zhan, Y. Liu, K. Han, W. Xie, and A. Zisserman, "Elip: Enhanced visual-language foundation models for image retrieval," *arXiv preprint arXiv:2502.15682*, 2025.

[13] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[14] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[18] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for vision and vision-language tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Transactions on Machine Learning Research (TMLR)*, 2022.

[20] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[21] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] Y. Chen, Z. Ma, Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, W. Hu, X. Qie, and J. Wu, "Vilem: Visual-language error modeling for image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[23] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[25] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord, "Deep semantic-visual embedding with localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal lstm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Z. Ji, H. Wang, J. Han, and Y. Pang, "Saliency-guided attention network for image-sentence matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[29] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[30] D. Kim, N. Kim, and S. Kwak, "Improving cross-modal retrieval with set of diverse embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[32] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proceedings of the 27th ACM International Conference on Multimedia (ACMMM)*, 2019.

[33] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] C. Thomas and A. Kovashka, "Preserving semantic neighborhoods for robust cross-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[36] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[37] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[38] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[39] Z. Wang, Z. Gao, K. Guo, Y. Yang, X. Wang, and H. T. Shen, "Multilateral semantic relations modeling for image text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[40] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[41] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[42] S. Yan, L. Yu, and Y. Xie, "Discrete-continuous action space policy gradient-based attention for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[43] S. Zeng, C. Liu, J. Zhou, Y. Chen, A. Jiang, and H. Li, "Learning hierarchical semantic correspondences for cross-modal image-text retrieval," in *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 2022.

[44] H. Zhang, Z. Mao, K. Zhang, and Y. Zhang, "Show your faith: Cross-modal confidence-aware network for image-text matching," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[45] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[46] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[47] X. Zheng, R. Ji, X. Sun, B. Zhang, Y. Wu, and F. Huang, "Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[48] E. Vendrow, O. Pantazis, A. Shepard, G. Brostow, K. Jones, O. Mac Aodha, S. Beery, and G. Van Horn, "Inquire: A natural world text-to-image retrieval benchmark," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[49] G. Kordopatis-Zilos, V. Stojnić, A. Manko, P. Suma, N.-A. Ypsilantis, N. Efthymiadis, Z. Laskar, J. Matas, O. Chum, and G. Tolias, "Ilias: Instance-level image retrieval at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[50] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021, if you use this software, please cite it as below. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[51] N. Sogi, T. Shibata, and M. Terao, "Object-aware query perturbation for cross-modal image-text retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[52] Y. Liu, J. Yao, Y. Wang, and W. Xie, "Zero-shot composed text-image retrieval," in *British Machine Vision Conference (BMVC)*, 2023.

[53] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," *arXiv preprint arXiv:2303.11916*, 2023.

[54] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[55] L. Ventura, A. Yang, C. Schmid, and G. Varol, "Covr: Learning composed video retrieval from web video captions," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[56] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen, "Uniir: Training and benchmarking universal multimodal information retrievers," in *Proceedings of the European conference on computer vision (ECCV)*, 2024.

[57] Y. Liu, P. Chen, J. Cai, X. Jiang, Y. Hu, J. Yao, Y. Wang, and W. Xie, "Lamra: Large multimodal model as your advanced retrieval assistant," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[58] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

[59] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[60] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[61] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2007.

[62] G. Tolias and H. Jégou, "Visual query expansion with or without geometry: refining local descriptors by feature aggregation," *Pattern recognition*, 2014.

[63] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[64] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[65] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[66] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[67] Y. Bhalgat, J. F. Henriques, and A. Zisserman, "A light touch approach to teaching transformers multi-view geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[68] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Text-to-image gan-based scene retrieval and re-ranking considering word importance," *IEEE Access*, 2019.

[69] L. Qu, M. Liu, W. Wang, Z. Zheng, L. Nie, and T.-S. Chua, "Learnable pillar-based re-ranking for image-text retrieval," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[70] Z. Long, X. Ge, R. McCreadie, and J. M. Jose, "Cfir: Fast and effective long-text to image retrieval for large corpora," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[71] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2021.

[72] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[73] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[74] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[75] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, "Datacomp: In search of the next generation of multimodal datasets," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[76] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel, "Mobileclip: Fast image-text models through multi-modal reinforced training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[77] G. Zhan, W. Xie, and A. Zisserman, "A tri-layer plugin to improve occluded detection," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2022.

[78] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[79] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.