Bi-directional Training for Composed Image Retrieval via Text Prompt Learning

Zheyuan Liu[1] Weixuan Sun[1] Yicong Hong[1] Damien Teney[2,3] Stephen Gould[1]
[1]Australian National University
[2]Idiap Research Institute [3]Australian Institute for Machine Learning, University of Adelaide
{zheyuan.liu, weixuan.sun, stephen.gould}@anu.edu.au
mr.yiconghong@gmail.com, damien.teney@idiap.ch

Abstract

Composed image retrieval searches for a target image based on a multi-modal user query comprised of a reference image and modification text describing the desired changes. Existing approaches to solving this challenging task learn a mapping from the (reference image, modification text)-pair to an image embedding that is then matched against a large image corpus. One area that has not yet been explored is the reverse direction, which asks the question, what reference image when modified as described by the text would produce the given target image? In this work we propose a bi-directional training scheme that leverages such reversed queries and can be applied to existing composed image retrieval architectures with minimum changes, which improves the performance of the model. To encode the bi-directional query we prepend a learnable token to the modification text that designates the direction of the query and then finetune the parameters of the text embedding module. We make no other changes to the network architecture. Experiments on two standard datasets show that our novel approach achieves improved performance over a baseline BLIP-based model that itself already achieves competitive performance. Our code is released at https://github.com/Cuberick-Orion/Bi-Blip4CIR

1. Introduction

Composed image retrieval (CIR) [22, 37, 38] aims at retrieving images based on the user input of a reference image and a sentence stating certain desired changes. The retrieved target image must encompass the user-specified changes while remaining similar to the reference image on other aspects. Unlike conventional image [35] or text-based [21, 40] retrieval with one input modality, CIR is more precise, which is appreciated in domains such as image search or e-commercial product retrieval.

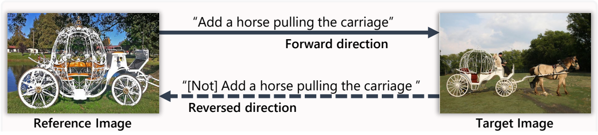

"Add a horse pulling the carriage"
Forward direction
"[Not] Add a horse pulling the carriage "
Reversed direction
Reference Image
Target Image

Figure 1. **Forward direction:** Existing methods on composed image retrieval focus on mapping the input ⟨reference image, modification text⟩-pair to the target image. **Reversed direction:** We propose to also exploit information in the mapping from the ⟨target image, reversed modification text⟩-pair to the reference image. Note that the reversed text is an illustration, which we do not have and have to infer from existing data.

Existing approaches [2, 6, 8, 37] on CIR mostly focus on learning the mapping of embeddings from the given ⟨reference image, text⟩-pair to a target image, where the multi-modal input pair is jointly embedded through various mechanisms including gating [37], multi-layer attention [6, 8] or convex summation [3]. However, few have considered the full potential of the information available in the training data. If we view conventional CIR methods as training in the forward direction from reference to target, while conditioning on the text; then it is easy to see that a reverse direction can be achieved from target to reference, provided that a suitable text reversal can be found (Figure 1). Harnessing information in such a bi-directional fashion benefits the training and increases the robustness of the model. We argue that this is particularly valuable given the limited dataset sizes [22, 38] and high annotation cost [29] of existing datasets for this task.

To this end, we propose bi-directional training for CIR to simultaneously learn to retrieve images (target or reference, respectively) from both the forward and reversed queries. As stated above, one major challenge is text reversal, that is, to create text that carries the opposite semantic meanings compared to the original. For human annotators, rewriting the text can be easy. However, in the absence of additional annotations, the naive approach of reversing text

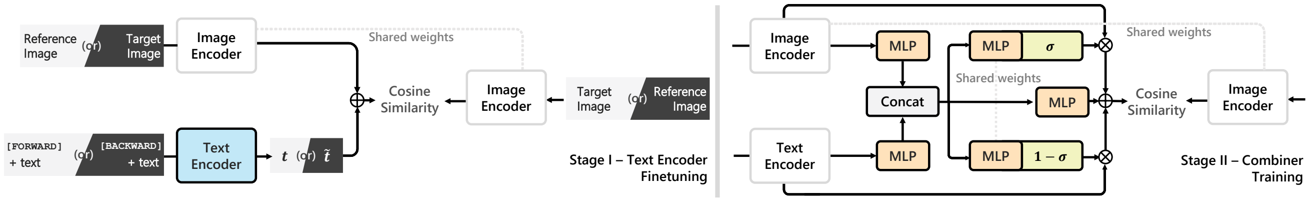


Figure 2. Our bi-directional training pipeline. Only modules with colors are being updated. Here, $\otimes$ denotes element-wise product and $\oplus$ denotes element-wise addition. Note the details of contrastive loss is not fully shown here (see Section 3.1), instead, we only illustrate with the positive target/reference images. **Left:** Text encoder finetuning, where we input queries of both directions (distinguished by light/dark colors), and infer $\tilde{t}$ through training, along with t. Note the bi-directional tokens prepended to the text. We omit the `[CLS]` token from the text encoder for brevity. **Right:** Combiner training detailing the architecture [4]. We train the model with bi-directional queries simultaneously, as in the previous stage.

through handcrafted linguistic rules can be difficult. This is especially true for the CIRR dataset on generic images, as the sentences are of great variations and high complexity. One approach is to pass the text embeddings through a dedicated module (e.g., an MLP) when inversion is needed, while preserving the rest of the joint embedding model and train end-to-end. We argue a better alternative is to leverage the text encoders of vision-and-language pretrained (VLP) models, which are commonly used in CIR to extract text embeddings. The advantage of which is that such encoders are pretrained on massive corpus, and, hence, are powerful in capturing the nuanced semantics of natural language. Specifically, we take inspiration from recent work on few-shot text-guided image editing, i.e., text inversion [28] and propose to prepend the text with special learnable tokens signifying its directionality. We then leverage the state-of-the-art two-stage training pipeline [4] of the combiner architecture [3], where the first stage is to finetune the text encoder. We discover that through finetuning, the text encoder can associate the concept of directionality of the text with the special text tokens, and generate different embeddings from the same text input. This allows us to effectively treat queries of both directions equally after finetuning, and make no further changes to the subsequent joint embedding model (i.e., the combiner) or its training process when we include the reversed queries. In order to include the reversed queries in training, we involve a secondary loss term in the contrastive loss. To better take advantage of the reverse queries we employ a modified sampling strategy for negative samples, such that the loss is more coherent in the bi-directional training scheme, as will be detailed in Section 3.1.

In summary, we propose a bi-directional training scheme for composed image retrieval (CIR), which jointly trains on the forward queries, i.e., from reference to target images, as well as the reversed queries from target to reference. To obtain text embeddings of reversed semantics, we prepend the text with learnable tokens and finetune the text encoders. Additionally, we modify the contrastive loss on the reversed path. No further changes to the model architecture or training pipeline are needed, which makes our approach easily applicable to existing methods. We empirically show that our approach achieves improved performance on datasets of diverse domains over a BLIP-based baseline that has already achieved state-of-the-art performance.

2. Related Work

Composed Image Retrieval. The task of composed image retrieval (CIR) introduced by Vo et al. [37] aims at studying the composition of multi-modal features, where initially inputs of low complexity [12, 14] are considered. It is later adopted for fashion product [5, 10, 38], and recently further extend into generic images [22]. Most existing methods follow a fusion paradigm, where the features of the input reference image and text are jointly embedded and compared against features of all candidate target images for the closest match. Extensive research [2, 3, 6, 8, 22, 37] is done on the fusion mechanism of the network, with the recent state-of-the-art [3, 4] adopting a combiner architecture that performs a convex combination of the input modalities within the CLIP [27] feature manifold. We note that this is the first method that simultaneously achieves state-of-the-art on both fashion and generic image datasets. Meanwhile, in the contemporary unpublished work, Levy et al. [19] and Liu et al. [23] both adopt the BLIP [20] multimodal encoder that further improves the performance. Concurrently, others have explored splitting CIR into two-stage with coarse and fine searching [39], disentangling the image-image and image-text matching into dual branches [7], and expanding the task into zero-shot scenarios [29]. Among methods developed for CIR, we are mostly related to DCNet [15], where a correction module is used to model the difference between the reference and target images and match it to the text. In essence, it explores a different directionality of the training data than ours. Compared to DCNet, our method does not require an additional module, or joint loss that connects the said module to the main network. Instead, our bi-directional training treats samples of both directions in the same manner. We also avoid computing feature differences through subtractions, which can be hard to learn [1].

Cycle consistency. Conceptually, our bi-directional training resembles the cycle-consistency concept seen in vision-and-language (e.g., Robust VQA [31]) and generation (e.g., CycleGAN [42]) tasks, as we share the philosophy of manipulating model inputs and outputs to further exploit information in the training instances. However, such a concept bears different motivations and designs under various task setups. For VQA, it is implemented as a secondary question-answering stage with generated rephrases of the question, which improves the robustness of the model under linguistic variations. CycleGAN, however, utilizes cycle consistency with the absence of paired training instances between two domains. We note that CIR is different in that, three entities are included in the input and output. This setup requires unique designs, hence, making our method fundamentally different from previous work.

Text inversion. Recent work [9, 28] on few-shot text-guided image editing shows that it is possible to bind the appearance of a certain instance in an image with artificially injected special tokens within the text prompt through finetuning, so that the model can generate diverse samples containing said instance. A technique termed text inversion. Though not entirely equivalent, we take inspiration from the above work and adopt a similar idea in finetuning the text encoder. Our intention is to associate the concept of text directionality with special tokens, so that the model can recognize the need of reversing the semantics of the language.

3. Bi-directional Composed Image Retrieval

Given the embeddings of a query of ⟨reference image, modification text⟩-pair denoted as $q = \langle I_\mathrm{R}, t\rangle$, the objective is to locate a target image that best matches the query, whose embedding is denoted as I_T. Our goal is to also learn on the reversed query $\tilde{q} = \langle I_\mathrm{T}, \tilde{t}\rangle$ simultaneously, which maps from the target I_T to the I_R. Here, $\tilde{t}$ represents the text embedding that is semantically reversed. However, $\tilde{t}$ is not directly computable as the text associated with such reversed embeddings do not exist in the training data. To overcome this difficulty we propose to infer the semantically reversed text embedding $\tilde{t}$ from the original modification text using the text encoder.

Our method is based on a recent architecture by Baldrati et al. [4], which is a two-stage design as illustrated in Figure 2. Though, as an augmentation scheme other models are also applicable, provided that they utilize tokenized text embeddings that is common in VLP models. In the following subsections, we first propose our main idea on leveraging the text prompts for inferring the reverse query (Section 3.1). We then detail the model and training pipeline adapted from [3, 4] (Section 3.2). Finally, we discuss the

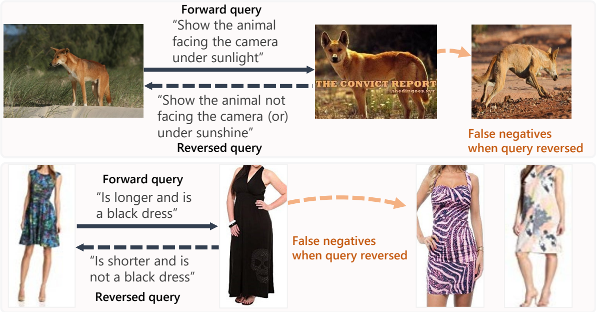

Forward query
"Show the animal facing the camera under sunlight"
"Show the animal not facing the camera (or) under sunshine"
Reversed query
THE CONVICT REPORT
thedingoes.xyz
False negatives when query reversed
Forward query
"Is longer and is a black dress"
"Is shorter and is not a black dress"
Reversed query
False negatives when query reversed

Figure 3. False negatives can exist in the reversed direction. A one-to-one text in the forward direction can become one-to-many when its semantics are reversed. Since we lack human labels for the reversed queries, only the original reference image is deemed positive. We point out that this issue is related to both the semantics of the text, as well as the image corpus. Hence, the prevalence of false negatives might vary for individual samples in different datasets. Note that the reversed text here is provided to illustrate the case, we do not have such annotations in the datasets.

false negatives in the reversed queries in Section 3.3, which impacts our inference strategy.

3.1. Bi-directional Training

Text Prompt Learning. As shown in Figure 2 (left), we leverage the first-stage text encoder finetuning to infer $\tilde{t}$ alongside t, such that it can produce text embedding of either direction for a given text. Specifically, we prepend special learnable tokens to the modification text sentences. The idea is to bind the concept of query directionality to such tokens through learning.

Here one option is to make no changes to the text of the forward queries, and only inject a token when the text needs to be reversed. The implication is that the forward queries shall be trained in their de-facto forms (see Section 3.2), while only making necessary changes to the additional augmentations included. However, this introduces an asymmetry in the treatment of the forward and reversed queries. Our intuition is that the balanced approach will force the model to better recognize the purpose of the injected tokens and distinguish between the forward and reversed modes.

We choose `[FORWARD]` and `[BACKWARD]` as the learnable tokens for the forward and reversed queries respectively in all our experiments. Together with the `[CLS]` token from the text encoder, which is pretrained to summarize the text, a tokenized text sequence $\mathtt{t}$ of $\{\mathtt{t}_1 \cdots \mathtt{t}_n\}$ in the forward query is then processed into $\{\mathtt{[CLS]}, \mathtt{[FORWARD]}, \mathtt{t}_1 \cdots \mathtt{t}_n\}$, which is passed to the text encoder for embedding. Likewise for the reversed case with $\{\mathtt{t}_1 \cdots \mathtt{t}_n\}$ unmodified.

We note that prompt design [18, 28] could potentially be of significant value to the end results. However, our focus is on the bi-directional training scheme, hence, we choose to rely on simple, generic tokens and make no specific adjustments to suit each dataset.

Modifications to the Training Pipeline. A favorable characteristic of our approach is the minimum changes made to the existing training pipeline. Here, we leverage the state-of-the-art two-stage training scheme (details of which are in Section 3.2), as illustrated in Figure 2. We note that, after finetuning the text encoder to infer $\tilde{t}$, the reversed queries $\langle q = \langle I_{\mathrm{T}}, \tilde{t} \rangle, I_{\mathrm{R}} \rangle$ can be constructed by a simple exchange of image orders. The entire process is computed on-the-fly with low additional cost. We then treat all queries equally regardless of their directionality, which allows us to train the second-stage combiner with bi-directional samples simultaneously without changes to its architecture.

Negative Sampling in Contrastive Loss. We follow previous work and use the batch-based classification (BBC) loss [37]. Given a batch size of B, with the embeddings of the i-th query pair $\langle I_{\mathrm{R}}^i, t^i \rangle$, its corresponding positive target I_{T}^i, the forward query loss is computed as:

$$\mathcal{L}_{\mathrm{F}} = -\frac{1}{B} \sum_{i=1}^{B} \log \left[\frac{\exp\left[\lambda \cdot \kappa\left(f(I_{\mathrm{R}}^i, t^i), I_{\mathrm{T}}^i\right)\right]}{\sum_{j=1}^{B} \exp\left[\lambda \cdot \kappa\left(f(I_{\mathrm{R}}^i, t^i), I_{\mathrm{T}}^j\right)\right]} \right], \quad (1)$$

where $f(\cdot, \cdot)$ denotes the combination function, $\kappa(\cdot, \cdot)$ is the similarity kernel implemented as cosine similarity and λ is the temperature parameter. We follow Baldrati et al. [4] and set λ to 100 in all our experiments. As the denominator shows, we normalize over all other matches within a batch in training, which includes both the positive I_{T}^i and all other target images in the batch as negatives I_{T}^j for $j \neq i$.

When training on the reversed queries that maps query $q = \langle I_{\mathrm{T}}, \tilde{t} \rangle$ to reference image I_{R}, multiple options exist in sampling the negatives to form a contrastive loss. Here, we propose to formulate the loss as:

$$\mathcal{L}_{\mathrm{B}} = -\frac{1}{B} \sum_{i=1}^{B} \log \left[\frac{\exp\left[\lambda \cdot \kappa\left(f(I_{\mathrm{T}}^i, \tilde{t}^i), I_{\mathrm{R}}^i\right)\right]}{\sum_{j=1}^{B} \exp\left[\lambda \cdot \kappa\left(f(I_{\mathrm{T}}^j, \tilde{t}^i), I_{\mathrm{R}}^i\right)\right]} \right], \quad (2)$$

where we have chosen to sample the negatives among candidate target images, i.e., I_{T}^j, as in Equation 1. Our intuition is to unify losses for the forward and reversed queries so that they are learned to contrast against the same group of negatives. We empirically confirm that such a loss obtains better performance compared to, e.g., sampling negatives among the I_{R}^j (see Section 4.3).

The loss for bi-directional training is then computed as the weighted sum of forward and reversed terms,

$$\mathcal{L} = \mathcal{L}_{\mathrm{F}} + \alpha \mathcal{L}_{\mathrm{B}}, \quad (3)$$

where α is a hyperparameter to balance the magnitudes of the two loss terms. We refer readers to the supp. mat. for details on determining this parameter and analysis.

3.2. Model Architecture and Training Pipeline

We base our bi-directional training on a recent state-of-the-art baseline obtained using the combiner architecture [3, 4] with BLIP [20] vision-and-language pretrained (VLP) model, termed BLIP4CIR. We follow Baldrati et al. [4] and adopt a two-stage training scheme as follows.

Text Encoder Finetuning. As shown in Figure 2 (left), we first finetune the text encoder. The architecture is relatively light with the multi-modal combination done through an element-wise addition. The output is compared against embeddings of candidate targets through cosine similarity. Note that the image encoder is kept frozen as it is prohibitively expensive to finetune.

The intuition of this finetuning stage is to reduce the domain gap between the pretraining tasks and the downstream task, CIR. Conceptually, the element-wise addition is used to encourage the output t as a displacement vector between I_{R} and I_{T} in the image manifold. Once the text encoder is finetuned, we freeze it and train a combiner module that replaces the element-wise addition mentioned above, which involves more sophisticated joint embedding functions.

Combiner Training. In the second-stage training, combiner replaces element-wise addition as the joint embedding function, with architecture depicted in Figure 2 (right). We refer readers to [3] for details. Overall, it can be viewed as a three-branch summation, with two branches implemented as a convex combination of the two input modalities, and a third branch of a learned image-text mixture. During combiner training, both the image and text encoders are frozen.

BLIP Embeddings. We propose to change the CLIP [27] image and text encoders in both stages to BLIP [20]. The motivation is two-fold. First, BLIP has been demonstrated to be a powerful VLP network, and through early experiments, we found its text encoder to be better than CLIP's. Second, we notice that BLIP is trained on text of potentially higher complexity that better matches the annotations in CIR datasets. In contrast, while the training data of CLIP is proprietary, the qualitative examples demonstrated by Radford et al. [27] are less sophisticated in general. We hypothesize that the BLIP text encoder is of stronger reasoning ability, and is, therefore, better suited to CIR. This is especially true for bi-directional training, as the text encoder needs to reverse the semantics of a given sentence. We empirically confirm that BLIP provides a much stronger baseline over the stock CLIP encoders. The change from pre-trained CLIP to BLIP encoders is straightforward, as they are of similar transformer [36] structures. Note that we do not involve the cross-attention fusion layer in BLIP that jointly embeds image and text modalities. Instead, we treat the image and text encoders as separate modules, as shown in Figure 2. This is to align with the usage of CLIP so that we can preserve the stock training pipeline.

3.3. False Negatives in Reversed Queries

A systematic issue for our bi-direction training scheme is the occurrences of false negatives in the reversed queries. As illustrated in Figure 3, a one-to-one modification text in the forward query could become one-to-many when its semantics are reversed. For instance, a text of "*change to a white cat*" will be reversed to "*change **from** a white cat*", which now leads to all cats (gray, ginger, fawn, etc.) that are not white — yet, only the original reference image is labeled as the positive. We especially note that the prevalence of false negatives in the reversed path is related to the semantics of the text as well as the image corpus (e.g., following the example above, the number cat images in the corpus shall affect the number of false negatives). Hence, the scale of such an issue can vary for individual samples across datasets. This renders global label smoothing techniques [33, 34] ineffective in our testings, as it could inadvertently affect queries with few false negatives.

Granted, false negatives also exist in the forward path [22, 38], due to the prohibitive high cost in exhaustively labeling all candidates for each query. But they are less common, as human annotations are often sufficiently specific. Plus, existing metrics are designed to mitigate the issue by using Recall@K of large K values.

We argue that nevertheless, training on the reversed queries still benefits the model, as demonstrated by our experiments of improved performance. However, special attention shall be paid in balancing the magnitudes of the loss terms (Equation 3), as the loss of the reversed path is generally higher. We also note that the prevalence of false negatives in the reversed direction suggests that validating on such reversed queries will lead to inferior results (see supp. mat. for details).

Inference Strategy. Following the above, we design our inference strategy mirroring existing work [4, 37], which only performs on the forward queries. For each query, we rank the similarities of the combined (I_R, t) embedding with all candidate I'_T and pick the highest as the prediction.

4. Experiments and Discussions

Datasets and Metrics. We follow Baldrati et al. [4] and test on two datasets of different domains.

Fashion-IQ [38] focuses on fashion products of three subtypes, *Dress, Shirt* and *Toptee*. In total it contains over 30k triplets sampled from 77k images. Each triplet includes two human-generated annotations. We follow previous work and report Recall@K with K = 10 and 50, while comparing the overall model performance with $(\text{R@}10 + \text{R@}50)/2$, as advised by Wu et al. [38]. The choice of K accounts for the potential false negatives in the forward queries. All results are on the validation set, as the test set is not publicly available.

CIRR [22] includes around 36k triplets sampled from 21k generic images sourced from NLVR^2 [32]. The human annotations are of higher complexity compared to Fashion-IQ. The dataset is designed to overcome the issue of false negatives, as it is prohibitive to exhaustively label every candidate target for each input query. Specifically, Liu et al. [22] group images by subsets of six and draw reference-target pairs from them. When annotating a given pair, annotators are instructed to avoid creating false negatives within the subset from which the pair is drawn. To this end, the evaluation protocol for CIRR is designed to be a combination of standard Recall@K with K =1, 5, 10, 50 and Recall$_\text{Subset}$@K with K =1, 2, 3, where Recall$_\text{Subset}$@K only considers candidates from the same subset as the pair. Following Liu et al. [22], we assess the overall model performance with $(\text{R@}5 + \text{R}_\text{Subset}\text{@}1)/2$. Results in the main table are reported on the test set, the ground truths of which are not available. Instead, we obtain results through the official evaluation server[1]. Results of ablation studies are reported on the validation set.

Implementation Details. We adopt the default image pre-processing scheme and model configuration as in [4], except when BLIP encoders [20] require a different setting in dimensions. This includes the input image resolutions (384 × 384) and the combiner input feature dimension (256 from BLIP encoder outputs). For finetuning the text encoder, we follow BLIP downstream task settings and optimize with AdamW [24] for 15 epochs, with a learning rate of 5×10^{-5}, a weight decay of 0.05 and a cosine learning rate schedule. We increase the learning rate of the last linear projection layer to 5×10^{-3} to speed up the convergence. For training the combiner, we adopt the original settings detailed in [3] and train for 200 epochs. The batch size of baseline experiments (i.e., without bi-directional training) follows [4]. The batch size of all bi-directional training experiments is reduced by half due to the GPU memory limit.

All experiments are trained with mixed-precision [25] in PyTorch with one NVIDIA A100 80G. We base our implementation on the official codebases released by Baldrati et al. [4] [2] and Li et al. [20] [3].

4.1. Results on Fashion-IQ

Table 1 compares our approach with existing state-of-the-art methods on Fashion-IQ. We note that our BLIP-based baseline model (row 18) outperforms all previous ap-

[1] https://cirr.cecs.anu.edu.au/

[2] https://github.com/ABaldrati/CLIP4Cir

[3] https://github.com/salesforce/BLIP

		Dress		**Shirt**		**Toptee**		**Average**		**Avg.**
	Methods	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	**Metric**
1	MRN [16]	12.32	32.18	15.88	34.33	18.11	36.33	15.44	34.28	24.86
2	FiLM [26]	14.23	33.34	15.04	34.09	17.30	37.68	15.52	35.04	25.28
3	TIRG [37]	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.40
4	Relationship [30]	15.44	38.08	18.33	38.63	21.10	44.77	18.29	40.49	29.39
5	CIRPLANT [22]	14.38	34.66	13.64	33.56	16.44	38.34	14.82	35.52	25.17
6	CIRPLANT w/OSCAR [22]	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
7	VAL w/GloVe [6]	22.53	44.00	22.38	44.15	27.53	51.68	24.15	46.61	35.40
8	CurlingNet [39]	24.44	47.69	18.59	40.57	25.19	49.66	22.74	45.97	34.36
9	DCNet [15]	28.95	56.07	23.95	47.30	30.44	58.29	27.78	53.89	40.84
10	CoSMo [17]	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
11	MAAF [8]	23.8	48.6	21.3	44.2	27.9	53.6	24.3	48.8	36.6
12	ARTEMIS [7]	25.68	51.25	28.59	55.06	21.57	44.13	25.25	50.08	37.67
13	SAC w/BERT [13]	26.52	51.01	28.02	51.86	32.70	61.23	29.08	54.70	41.89
14	AMC [41]	31.73	59.25	30.67	59.08	36.21	66.06	32.87	61.64	47.25
15	CLIP4CIR [4]	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	50.03
16	CASE† [19]	**47.77**	**69.36**	**48.48**	**70.23**	**50.18**	**72.24**	**48.79**	**70.68**	**59.74**
17	BLIP4CIR (first-stage)	37.13	62.67	35.92	60.40	43.60	68.28	38.88	63.78	51.33
18	BLIP4CIR	40.65	66.34	40.38	64.13	**46.86**	69.91	42.63	66.79	54.71
19	BLIP4CIR+Bi (first-stage)	36.94	63.71	37.49	60.06	43.60	67.77	39.34	63.85	51.60
20	BLIP4CIR+Bi	**42.09**	**67.33**	**41.76**	**64.28**	46.61	**70.32**	**43.49**	**67.31**	**55.40**

Table 1. Comparison on Fashion-IQ validation set, we follow Wu et al. [38] to report Avg. Metric as *(R@10+R@50)/2*. † Contemporary work to ours, included for completeness. Best (resp. second-best) numbers are in bold-black (resp. blue), this excludes intermediate first-stage text encoder fine-tuning results marked in gray (rows 17, 19). BLIP4CIR denotes the baseline using BLIP encoders. +Bi denotes the bi-directional training. For CLIP4CIR [4], we report their best-performing model that uses the two-stage training with RN50x4 as backbone. Rows 1-2 are cited from [38]. Methods leveraging additional data and auxiliary tasks (e.g., [11]) are not included.

		Recall@K				**Recall$_{\text{Subset}}$@K**			**Avg.**
	Methods	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	**Metric**
1	TIRG [37]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
2	TIRG+LastConv [37]	11.04	35.68	51.27	83.29	23.82	45.65	64.55	29.75
3	MAAF [8]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
4	MAAF+BERT [8]	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
5	MAAF−IT [8]	9.90	32.86	48.83	80.27	21.17	42.04	60.91	27.02
6	MAAF−RP [8]	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.37
7	CIRPLANT [22]	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
8	CIRPLANT w/OSCAR [22]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
9	ARTEMIS [7]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
10	CLIP4CIR [4]	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
11	CASE† [19]	**48.00**	**79.11**	**87.25**	**97.57**	**75.88**	**90.58**	**96.00**	**77.50**
12	BLIP4CIR (first-stage)	35.18	67.11	79.18	94.70	68.71	86.65	94.51	67.90
13	BLIP4CIR	**40.17**	71.81	83.18	95.69	**72.34**	**88.70**	95.23	72.07
14	BLIP4CIR+Bi (first-stage)	35.30	67.42	79.88	94.58	68.55	86.46	94.75	67.99
15	BLIP4CIR+Bi	40.15	**73.08**	**83.88**	**96.27**	72.10	88.27	**95.93**	**72.59**

Table 2. Comparison on CIRR test set. We follow Liu et al. [22] and report the Avg. Metric as *(Recall@5+Recall$_{\text{Subset}}$@1)/2*. † Contemporary work to ours, included for completeness. Note that CASE can be additionally pre-trained on the large-scale LaSCo dataset [19]. Here, we include the results without such pre-training for a fair comparison. Best (resp. second-best) numbers are in bold-black (resp. blue), this excludes intermediate first-stage text encoder fine-tuning results marked in gray (rows 12, 14). BLIP4CIR denotes the baseline using BLIP encoders. +Bi denotes the bi-directional training. For CLIP4CIR [4], we report their best-performing model that uses the two-stage training with RN50x4 as backbone. Rows 1-8 are cited from [22].

proaches by a large margin, albeit exceeded by the contemporary work CASE (row 16), also using BLIP. Impressively, through the first-stage text encoder finetuning (row 17), the performance already surpasses the previous CLIP-based best-performing model (row 15), demonstrating the high quality of BLIP embeddings.

With the addition of bi-directional training, the overall performance is further improved consistently throughout the two stages (rows 17 *vs.* 19; rows 18 *vs.* 20). On the final results obtained in the second stage in row 20, we gain notable improvements on categories of *Dress* and *Shirt*, while retaining approximately a similar performance on *Toptee* compared to the BLIP baseline in row 18. We conjecture that each sub-class might benefit from the bi-directional training differently, due to the quality of the reversed queries considering the issue of the false negatives and the specific image corpus (see Section 3.3).

Our method is outperformed by the contemporary work CASE [19]. We conjecture the main reason to be that the pre-trained BLIP [20] multimodal encoder adopted by CASE is more powerful than the combiner, which is also observed in its concurrent work [23] with further perfor-

				Fashion-IQ			CIRR			
	Methods	Neg-Sampling	Bi-Token	R@10	R@50	Average	R@1	R@5	R$_{\text{Subset}}$@1	Average
1	BLIP4CIR	—	—	42.63	66.79	54.71	41.11	74.89	72.66	73.78
2	BLIP4CIR+Bi	○		42.43	67.23	54.83	40.61	74.46	**73.36**	73.91
3	BLIP4CIR+Bi		○	43.30	66.77	55.03	40.28	73.95	72.57	73.26
4	BLIP4CIR+Bi	FULLY-CONFIGURED		**43.49**	**67.31**	**55.40**	**42.36**	**75.46**	72.90	**74.18**

Table 3. Ablation studies on Fashion-IQ and CIRR. Best numbers are in bold. Results reported on validation sets. For Fashion-IQ, we report the average Recall@10 and 50 of all three categories. For CIRR, the Average column denotes *(Recall@5+Recall*$_{\text{Subset}}$*@1)/2*, as in Table 2. *Bi-Token* suggests adding learnable tokens to queries of both directions (i.e., bi-directional). *Neg-Sampling* represents our negative sampling scheme in the reversed contrastive loss. Here ○ denotes the item we are ablating (i.e., where we remove this particular item from the fully-configured model and assess the outcome).

mance increase. In comparison, we only leverage the BLIP visual and text encoders for feature extraction. Notably, Levy et al. [19] have introduced a similar bi-directional data augmentation scheme as ours on CASE that shows a benefit, effectively demonstrating the generalizability of this idea.

4.2. Results on CIRR

Table 2 compares the performance of state-of-the-art methods with our approach on CIRR test set. We note that BLIP embeddings (row 13) brings a consistent performance increase compared to the previous CLIP-based model (row 10), as in Fashion-IQ. Our bi-directional training scheme (row 15) brings further increase to the performance of the strong baseline. Admittedly, the performance increase is inconsistent across metrics, particularly for Recall$_{\text{Subset}}$. We conjecture two reasons that could account for this. First, the high complexity of the text in CIRR would render the learning of the reversed semantics harder. Second, as demonstrated in Figure 4 (left, blue), we notice that the combiner gains very little, if none, on Recall$_{\text{Subset}}$ throughout training, and that the fluctuations in performance are high among epochs. Given that Recall$_{\text{Subset}}$ is both more challenging by design and of high granularity, as it only considers five candidates from the same image subset, we point out the possibility that the state-of-the-art combiner architecture, though powerful, may fail in this metric. To compare, we overlay the validation curve obtained using our bi-directional training scheme, as in Figure 4 (left, orange). We note that globally, our performance on Recall$_{\text{Subset}}$@1 is on par with the baseline. However, since we assess the performance via both the Recall@5 and Recall$_{\text{Subset}}$@1, the reported performance on the test set (Table 2 row 15) is not necessarily optimal on each individual metric. We stress that our bi-directional training, as a data augmentation scheme, does *not* aim at improving the existing model architecture. It is therefore unsurprising that our result exhibits a similar learning behavior on Recall$_{\text{Subset}}$ as the combiner baseline.

In contrast, on Recall, our method gains noticeable improvements when $K = 5$ and onwards in Table 2 (row 15 *vs.* 13). We demonstrate that the performance improvement is consistently observed during training, as in Figure 4 (right, orange *vs.* blue), suggesting valuable information exists in the reversed queries that benefits the learning. Regarding the similar performance on Recall@1 when compared with the baseline, we point to the fact that Recall@1 could potentially be impacted by false negatives in the forward queries [22]. Given that CIRR is designed to assess Recall@K with $K = 5$, we conclude that our bi-directional training is generally beneficial to this task.

4.3. Ablation Studies

We conduct ablation studies on the two design choices of our method detailed in Section 3.1, as shown in Table 3.

Negative Sampling in Reversed Contrastive Loss. In Table 3 row 2, we compare the negative sampling scheme in the reversed path. Here, we report results obtained by contrasting against I^j_{R} as opposed to I^j_{T} in Equation 2. We note that the proposed sampling scheme is vital to the bi-directional training on both datasets. Specifically, we find that if adopting the negative sampling scheme on I^j_{R}, the model gains very little from the bi-directional training, with Recall@K either dropping below the baseline (row 1) or merely slightly improving over it, let alone performing on par with the fully-configured model (row 4). We conjecture that a negative sampling scheme on I^j_{R} in the reverse queries causes a misalignment between the forward and reversed training paths, as the model is trained to contrast against I^j_{T} in the forward queries. To compare, our proposed sampling technique is more coherent and yields better performance. We note that the Recall$_{\text{Subset}}$@1 for CIRR presents as an outlier in this case. In this particular case, we propose to assess the performance difference more on the global Recall metrics, and refer to our discussions on the granularity of Recall$_{\text{Subset}}$ as well as the general learning behaviour of the baseline method in Section 4.2.

Bi-directional Text Tokens. Table 3 row 3 demonstrates the effectiveness of adding learnable tokens in *both* directions of the text rather than in the reversed direction alone. Interestingly, we discover that compared to Fashion-IQ, CIRR is more benefited from such a technique. Without

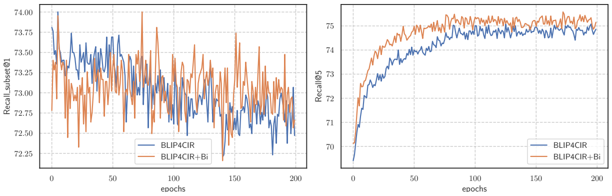


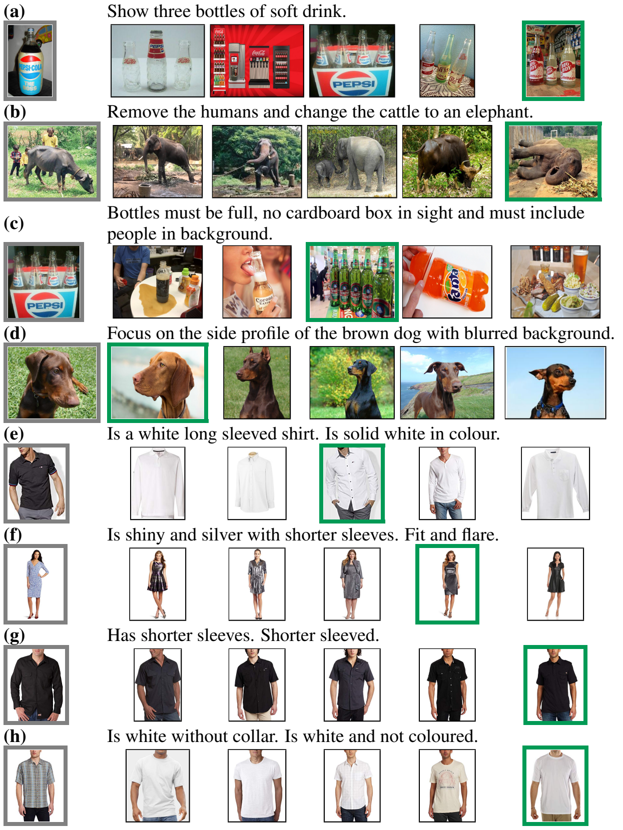
Figure 4. Validation performance on **(left)** $\text{Recall}_{\text{Subset}}$@1 and **(right)** Recall@5 trained for 200 epochs. Results obtained on the CIRR validation set. **BLIP4CIR (blue):** baseline model with combiner architecture. **BLIP4CIR+Bi (orange):** combiner with our bi-directional training scheme. We observe that the combiner architecture gains very little, if none on the more challenging $\text{Recall}_{\text{Subset}}$ metric with high fluctuations among epochs. And that our method, as an augmentation scheme, inherits such a property.

it, the performance, in particular on Recall@5, drops significantly (row 3 *vs.* 4). The reason is thought to be the complexity of the text inputs. In Fashion-IQ, the text is often short and carries simple meanings (Figure 5 e-h), such as "longer sleeves". In such cases, prepending a token solely in the reversed direction might still result in a proper reversion in semantics — as the model could learn to associate the token with a general negation of the context (i.e., "not"). However, the same cannot be said for CIRR, where text tends to be more complicated, as shown in Figure 5 (a-d). In such scenarios, our more balanced approach can better assist the model in identifying the directionality and associating it with said tokens.

4.4. Qualitative Examples

Figure 5 illustrates the qualitative examples of the retrieved results on both datasets. We specifically demonstrate successful cases where positive targets are highly ranked (d, e, f), as well as failure cases (b, g, h). We especially point to (d), where our method succeeds in reasoning over text with sophisticated intentions and retrieving the target, which demonstrates the quality of the BLIP embeddings as well as the power of our method.

As discussed in Section 3.3, we expect to encounter the issue of false negatives when reversing the queries. We show that for generic, short descriptions commonly found in Fashion-IQ (e-h), the reversed text can be widely imprecise, thus, leading to a great many possible candidates. The issue is worsened as images within Fashion-IQ can be of high similarity (e.g., g) — partially due to the natural low variability of cloth images. We show that this issue is less noticeable in CIRR, as the images are often diverse, containing multiple entities and/or rich differences in details. In addition, text in CIRR is often complex, which carries semantics that can remain specific when reversed. However, as discussed in Section 4.3, the high complexity of text in CIRR also poses a challenge in the model training, as the text embeddings can be hard to semantically reverse.


(a) Show three bottles of soft drink.

(b) Remove the humans and change the cattle to an elephant.

(c) Bottles must be full, no cardboard box in sight and must include people in background.

(d) Focus on the side profile of the brown dog with blurred background.

(e) Is a white long sleeved shirt. Is solid white in colour.

(f) Is shiny and silver with shorter sleeves. Fit and flare.

(g) Has shorter sleeves. Shorter sleeved.

(h) Is white without collar. Is white and not coloured.

Figure 5. Qualitative examples obtained with our method. In each example, gray box (leftmost) denotes the reference image, green box denotes the positive target, modification text is provided above the images. We show the top-5 candidates in ranking, except for when the positive target is ranked beyond top-5, in which case we remove the fifth-ranked candidate and append the positive at the end. This includes examples (b), (g) and (h).

5. Conclusion

In this work, we propose a bi-directional training scheme for composed image retrieval that additionally exploits information from the mapping of the (target image, modification text)-pair to the reference image. To tackle the challenge of inferring the reversed semantics of the text with the absence of additional annotations, we leverage the text encoder and prepend learnable tokens to the text inputs. Through finetuning, the text encoder binds the concept of text directionality to said tokens and can produce text embeddings for queries of either direction. We also involve a secondary contrastive loss with a modified sampling strategy for negative samples. Our approach is simple to try on any contrastive-based CIR model and can yield significant performance improvement at almost no cost. We empirically demonstrate that our bi-directional training scheme yields improved performance over a BLIP-based model that has already achieved competitive performance.

References

[1] A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, and A. M. Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[2] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber. Compositional learning of image-text query for image retrieval. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 2

[3] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5

[4] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2, 3, 4, 5, 6

[5] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010. 2

[6] Y. Chen, S. Gong, and L. Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 6

[7] G. Delmas, R. S. de Rezende, G. Csurka, and D. Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2022. 2, 6

[8] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modality-agnostic attention fusion for visual search with text feedback, 2020, *arXiv preprint* arXiv:2007.00145 [cs.CV]. 1, 2, 6

[9] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 3

[10] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. 2

[11] X. Han, X. Zhu, L. Yu, L. Zhang, Y.-Z. Song, and T. Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[12] P. Isola, J. J. Lim, and E. H. Adelson. Discovering states and transformations in image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[13] S. Jandial, P. Badjatiya, P. Chawla, A. Chopra, M. Sarkar, and B. Krishnamurthy. Sac: Semantic attention composition for text-conditioned image retrieval. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 6

[14] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[15] J. Kim, Y. Yu, H. Kim, and G. Kim. Dual compositional learning in interactive image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2, 6

[16] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in neural information processing systems*, 2016. 6

[17] S.-M. Lee, D. Kim, and B. Han. Cosmo: Content-style modulation for image retrieval with text feedback. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[18] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 3

[19] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski. Data roaming and early fusion for composed image retrieval, 2023, *arXiv preprint* arXiv:2303.09429 [cs.CV]. 2, 6, 7

[20] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 2, 4, 5, 6

[21] W. Li, L. Duan, D. Xu, and I. W. Tsang. Text-based image retrieval using progressive multi-instance learning. In *IEEE International Conference on Computer Vision*, 2011. 1

[22] Z. Liu, C. Rodriguez, D. Teney, and S. Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *IEEE International Conference on Computer Vision*, 2021. 1, 2, 5, 6, 7

[23] Z. Liu, W. Sun, D. Teney, and S. Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder, 2023, *arXiv preprint* arXiv:2305.16304 [cs.CV]. 2, 6

[24] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[25] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu. Mixed precision training, 2018, *arXiv preprint* arXiv:1710.03740 [cs.AI]. 5

[26] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 6

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 2, 4

[28] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[29] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[30] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, 2017. 6

[31] M. Shah, X. Chen, M. Rohrbach, and D. Parikh. Cycle-consistency for robust visual question answering. In *2019 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[32] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 5

[33] W. Sun, J. Zhang, J. Wang, Z. Liu, Y. Zhong, T. Feng, Y. Guo, Y. Zhang, and N. Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[35] S. Tong and E. Chang. Support Vector Machine active learning for image retrieval. In *Proceedings of the Ninth ACM International Conference on Multimedia*, 2001. 1

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin. Attention is all you need. In *International Conference on Neural Information Processing Systems*, 2017. 4

[37] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 4, 5, 6

[38] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5, 6

[39] Y. Yu, S. Lee, Y. Choi, and G. Kim. Curlingnet: Compositional learning between images and text for fashion iq data, 2020, *arXiv preprint* arXiv:2003.12299 [cs.CV]. 2, 6

[40] C. Zhang, J. Y. Chai, and R. Jin. User term feedback in interactive text-based image retrieval. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005. 1

[41] H. Zhu, Y. Wei, Y. Zhao, C. Zhang, and S. Huang. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2023. 6

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3