Zero-shot Composed Text-Image Retrieval

Yikun Liu[1,2]
yikunliu@sjtu.edu.cn
Jiangchao Yao[1,3]
Sunarker@sjtu.edu.cn
Ya Zhang[1,3]
ya_zhang@sjtu.edu.cn
Yanfeng Wang[1,3, †]
wangyanfeng622@sjtu.edu.cn
Weidi Xie[1,3, †]
weidi@sjtu.edu.cn

[1] Coop. Medianet Innovation Center,
Shanghai Jiao Tong University, China
[2] Beijing University of Posts and
Telecommunications, China
[3] Shanghai AI Laboratory

Abstract

In this paper, we consider the problem of composed image retrieval (CIR), it aims to train a model that can fuse multi-modal information, *e.g.*, text and images, to accurately retrieve images that match the query, extending the searching ability. We make the following contributions: (i) we initiate a scalable pipeline to automatically construct datasets for training CIR model, by simply exploiting a large-scale dataset of image-text pairs, *e.g.*, a subset of LAION-5B; (ii) we introduce a transformer-based adaptive aggregation model, **TransAgg**, which employs a simple yet efficient fusion mechanism, to adaptively combine information from diverse modalities; (iii) we conduct extensive ablation studies to investigate the usefulness of our proposed data construction procedure, and the effectiveness of core components in TransAgg; (iv) when evaluating on the publicly available benchmarks under the zero-shot scenario, *i.e.*, training on the automatically constructed datasets, then directly conduct inference on target downstream datasets, *e.g.*, CIRR and FashionIQ, our proposed approach either performs on par with or significantly outperforms the existing state-of-the-art (SOTA) models. Project page: https://code-kunkun.github.io/ZS-CIR/

1 Introduction

In the recent literature, vision-language models have made tremendous progress, by jointly training image and text representation on large-scale dataset collected from the Internet. For example, CLIP [24] and ALIGN [13] trained with simple noise contrastive estimation [22], have demonstrated surprisingly strong transferability and generalizability on zero-shot classification or cross-modal retrieval. In this paper, we consider the task of composed image retrieval (CIR), that aims to retrieve images by leveraging a combination of reference image and textual information that illustrates desired modifications. The model needs to use visual and language representation interchangeably, and discover target images that satisfy the user's expectation. In comparison to image-to-image or text-to-image retrieval, CIR captures richer semantics about the user's intention, and thus has the potential to enable more precise retrieval on images or e-commerce products.

Existing approaches [1, 6, 18, 27] for composed image retrieval typically train deep neural networks under fully supervised setting, which requires a dataset, consisting of sufficient {a reference image, a relative caption, and a target image} triplets. However, compared with collecting the text-image pairs, manually constructing such a triplet dataset is usually very expensive, that requires substantial human efforts, to thoroughly examine the reference image and target image and produce a text description to capturie their distinctions. Consequently, the practical datasets for training CIR models tend to be limited by scale.

In this paper, we initiate a scalable pipeline to automatically construct datasets for training CIR model, by exploiting the vast amount of image-caption data available on the Internet. Specifically, for one image-caption sample, we can revise its caption and use the resulting edited caption as a query to retrieve the target image with similar caption, where we adopt an off-the-shelf Sentence Transformer to compute similarity between sentences. Depending on the different approaches for revising captions, *i.e.*, using template or large language models (LLM), we obtain two different training datasets respectively. In addition, we introduce a transformer-based model, that employs a simple yet efficient fusion mechanism to adaptively combine information from diverse modalities. Once trained on the automatically constructed datasets, the model can be directly applied to target downstream CIR benchmarks without any finetuning, thus advocates zero-shot generalisation.

To summarise, we make the following contribution: (i) we propose a retrieval-based pipeline for automatically constructing dataset for training, with the easily-acquired image-caption data on Internet; (ii) we introduce a transformer-based aggregation model, termed as **TransAgg**, that employs a simple yet efficient modules to dynamically fuse information from different modalities. (iii) we train a model on the automatically constructed dataset, and directly evaluate on publicly available CIR benchmarks, thus resembling zero-shot composed image retrieval. In particular, we extensively evaluate the applicability of our constructed dataset, with different pre-trained backbones and fine-tuning types, and perform thorough ablation studies to validate the effectiveness of the transformer module and adaptive aggregation of our model; (iv) while comparing with existing approaches on two public benchmarks under zero-shot scenario, namely, CIRR and FashionIQ, our model performs on par or significant above the existing state-of-the-art (SOTA) models, and is sometimes comparable to fully supervised ones.

2 Related Work

Image Retrieval. Standard image retrieval includes both image-to-image retrieval and text-to-image retrieval. Existing research can be mainly divided into two categories. One uses dual tower structure [7, 14, 20, 23]. It relies on a good feature extractor to get features of text or image, and then uses cosine similarity for retrieval. The other one is to pass image-image or text-image pairs through a mutli-modal encoder to compute their similarity [4, 17, 21]. Despite the impressive progress, these retrieval models are unable to exploit the complemantary information in different modalities for constructing fine-grained queries.

Composed Image Retrieval. Composed Image Retrieval (CIR) considers the problem of retrieving images based on the reference images and relative captions. Till recently, majority research in CIR has concentrated on the fusion of multiple modalities to generate optimal multimodal representations. Specifically, TIRG [27] proposes to use residual modules and gating modules to fuse features. CIRPLANT [18] employed vision-and-language pre-trained (VLP) multi-layer transformers to fuse features that come from distinct modalities. CLIP4CIR [1] leverages CLIP [24] as feature extractor and follows a two-stage training procedure. In the first stage, CLIP [24] text encoder is fine-tuned, and a combiner is trained in the second stage, culminating in remarkable outcomes.

Concurrent Work. Several recent papers [9, 15, 25] also explore the idea of zero-shot composed image retrieval, specifically, Pic2Word [25] employs image-caption and unlabeled image datasets to train a mapping network that marks the image as a token, and performs cross-modal retrieval with CLIP [24]. CompoDiff [9] proposes a two-stage approach for training diffusion model to address the CIR problem and introduces the SynthTriplet18M dataset, comprising images synthesized via the prompt-to-prompt [11] model guided by corresponding captions. CASE [15] proposes to use BLIP [16] model to accomplish the CIR task through early fusion and utilzing the few-shot capability of GPT-3[3], and the VQA2.0[8] dataset to construct a dataset of almost 400K triplets in an semi-automatic manner. The target images are manually selected from the 24 visually nearest neighbors of referenece images. Unlike the aforementioned approach, our approach is fully automated and does not require any human intervention, based on retrieval from a large-scale corpus of real images.

3 Method

In this section, we start by formulating the problem of composed image retrieval in Sec. 3.1, then provide details of our proposed architecture in Sec. 3.2, lastly, in Sec. 3.3, we describe the two ideas for automatically constructing training set for CIR task, namely, Laion-CIR-Template and Laion-CIR-LLM.

3.1 Problem Scenario

We consider the problem of composed image retrieval, specifically, at training time, each sample can be represented as a triplet, *i.e.*, $\mathcal{D}_{\text{train}} = \left\{(I_r, I_t, t)\,|I_r \in \mathbb{R}^{H\times W\times 3}, I_t \in \mathbb{R}^{H\times W\times 3}\right\}$, Specifically, we train a model that takes the reference image (I_r) and relative caption (t) as input, and construct a composed query, that can retrieve one target image (I_t):

$$Q = \Phi_{\text{TransAgg}}(I_r, t) = \Phi_{\text{agg}}(\Phi_{\text{fuse}}(\Phi_{\text{visual}}(I_r),\ \Phi_{\text{text}}(t))) \tag{1}$$

Q refers to the composed query, that is to rank all images in a retrieval set $\Omega = \{I_i, i = 0, \cdots, m\}$ based on the relevance, *i.e.*, cosine similarity computed by between query and image embedding. For each composed query, the retrieval set is split into positive P_q and negative N_q sets, with the former consisting of instances that satisfy conditional editing on reference image. The trainable modules include: visual encoder (Φ_{visual}), text encoder (Φ_{text}), multi-modal fusion module (Φ_{fuse}), and an aggregation module (Φ_{agg}).

3.2 Composed Image Retrieval Model

Here, we start by introducing our proposed model for composed image retrieval, termed as TransAgg, and followed by its detailed training objective.

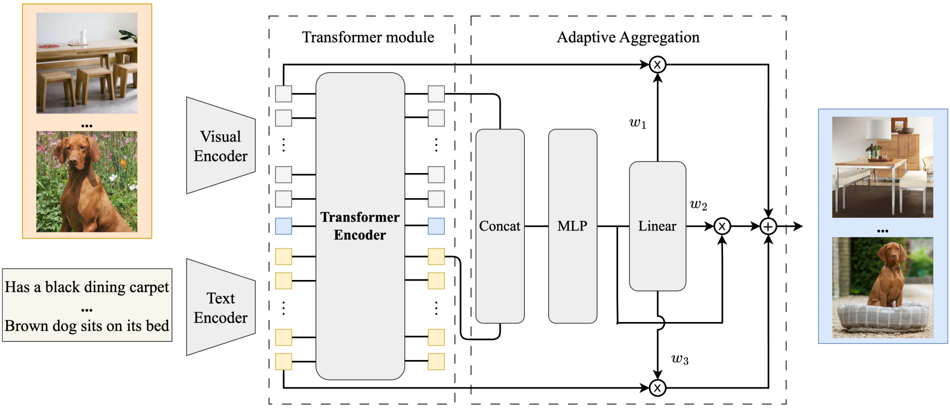


Figure 1: An overview of our proposed architecture, that consists of a visual encoder, a text encoder, a Transformer module and an adaptive aggregation module.

3.2.1 Architecture

As illustrated in Figure 1, our proposed CIR model consists of three components: encoders to extract features from visual and textual inputs respectively, a Transformer module to capture the interaction between two modalities, and an adaptive aggregation module that combats modal redundancy and fuses the features together.

Visual and Text Encoders. We adopt pre-trained vision and language models as our encoders for different modalities given their impressive performance and flexibility to maintain the semantics. Formally, we denote the feature extraction via the following notations,

$$\mathcal{F}_{\mathrm{Vr}} = \Phi_{\mathrm{visual}}\left(I_r\right) \in \mathbb{R}^{|\mathcal{V}|\times d}, \qquad \mathcal{F}_{\mathrm{W}} = \Phi_{\mathrm{text}}\left(t\right) \in \mathbb{R}^{|\mathcal{W}|\times d} \tag{2}$$

where I_r denotes the reference image encoded by the visual encoder Φ_{visual}, and t refers to the relative caption encoded by the textual encoder Φ_{text}. In our experiments, we primarily use pretrained BLIP [16] or CLIP [24] as our visual and text encoders.

Transformer Fusion. Regarding the input of our Transformer module, in addition to $\mathcal{F}_{\mathrm{Vr}}$ and $\mathcal{F}_{\mathrm{W}}$, a learnable token embedding $\mathcal{F}_{\mathrm{sep}}$ is also integrated to discriminate the modalities. The feature interaction between visual and textual modality can be formulated as:

$$\left[\mathcal{F}'_{\mathrm{Vr}}, \mathcal{F}'_{\mathrm{sep}}, \mathcal{F}'_{\mathrm{W}}\right] = \Phi_{\mathrm{fuse}}\left(\left[\mathcal{F}_{\mathrm{Vr}}, \mathcal{F}_{\mathrm{sep}}, \mathcal{F}_{\mathrm{W}}\right]\right) \tag{3}$$

where $[\cdot,\cdot,\cdot]$ denotes the feature concatenation, $\Phi_{\mathrm{fuse}}(\cdot)$ is a two-layer Transformer module, and the input and output of each feature vector maintains the same shape. The visual and the textual features have been augmented through the feature interaction in the Transformer, resulting in the refined features $\mathcal{F}'_{\mathrm{Vr}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ and $\mathcal{F}'_{\mathrm{W}} \in \mathbb{R}^{|\mathcal{W}|\times d}$.

Adaptive Aggregation. Here, we take out the internal features corresponding to the image global patch and the text global token respectively, and concatenate them together to be transformed as the fusion features $\mathcal{F}_{\mathrm{U}} \in \mathbb{R}^d$ through an MLP module, we then apply a linear layer to project $\mathcal{F}_{\mathrm{U}}$ into weighting parameters (w_1, w_2, w_3) that act as multipliers for $\mathcal{F}^{\mathrm{G}}_{\mathrm{Vr}}$, $\mathcal{F}_{\mathrm{U}}$ and $\mathcal{F}^{\mathrm{G}}_{\mathrm{W}}$, where $\mathcal{F}^{\mathrm{G}}_{\mathrm{Vr}}$ indicates the global BLIP/CLIP visual features, $\mathcal{F}^{\mathrm{G}}_{\mathrm{W}}$ denotes the global BLIP/CLIP textual features. The final image-text representation Q is computed as:

$$Q = w_1 * \mathcal{F}^{\mathrm{G}}_{\mathrm{Vr}} + w_2 * \mathcal{F}_{\mathrm{U}} + w_3 * \mathcal{F}^{\mathrm{G}}_{\mathrm{W}} \tag{4}$$

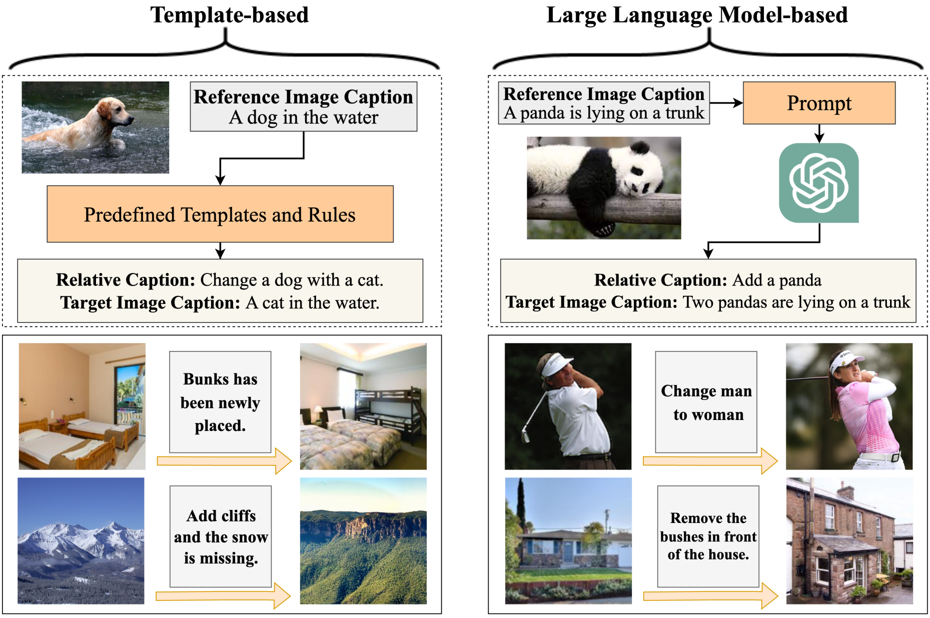

Template-based
Reference Image Caption
A dog in the water
Predefined Templates and Rules
Relative Caption: Change a dog with a cat.
Target Image Caption: A cat in the water.
Large Language Model-based
Reference Image Caption
A panda is lying on a trunk
Prompt
Relative Caption: Add a panda
Target Image Caption: Two pandas are lying on a trunk
Bunks has been newly placed.
Add cliffs and the snow is missing.
Change man to woman
Remove the bushes in front of the house.

Figure 2: An overview of our proposed dataset construction procedure, based on sentence template (left), or large language models (right).

3.2.2 The Training Objective

For model training, we follow previous work and use the batch-based classification (BBC) loss [27]. Given a batch size of B, the i-th query pair (I_r^i, t^i) should be close to its positive target I_t^i and far away from the negative instances, which can be formulated as

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\log\left[\frac{\exp\left[\kappa\left(Q^i,\mathcal{F}^i_{\mathrm{Vt}}\right)/\tau\right]}{\sum_{j=1}^{B}\exp\left[\kappa\left(Q^i,\mathcal{F}^j_{\mathrm{Vt}}\right)/\tau\right]}\right] \tag{5}$$

where $\tau = 0.01$ refers to the temperature parameter, and $\kappa(\cdot,\cdot)$ denotes the cosine similarity, Q^i is computed by Eq. (4) and $\mathcal{F}^i_{\mathrm{Vt}} = \Phi_{\mathrm{visual}}(I_t^i)$ is the representation of the target image of that query. In practise, to effectively train a model for composed image retrieval, a significant amount of triplet data is often required, unfortunately, collecting and annotating CIR datasets can be time-consuming and costly. In the following section, we describe an automatic pipeline for constructing dataset suitable for CIR training.

3.3 Dataset Construction

In order to train the CIR model, we need to construct a dataset with triplet samples, *i.e.*, reference image, relative caption, target image. Specifically, we start from the **Laion-COCO**[1] that contains a massive number of image-caption pairs, and then edit the captions with sentence templates or large-language models (Sec. 3.3.1), to retrieve the target images (Sec. 3.3.2), as shown in Figure 2. The details are discussed in the following sections.

[1] https://laion.ai/blog/laion-coco/

3.3.1 Generating Relative Caption

Generation Based on Language Templates. Here, we aim to generate the relative caption based on predefined templates and rules. Specifically, we take inspiration from [18], and consider eight types of semantic operations, namely *cardinality, addition, negation, direct addressing, compare&change, comparative statement, statement with conjunction and viewpoint*. For these operations, it is straightforward to define diverse rules to edit the original caption of Laion-COCO images. Taking the type *compare&change* as an example, we first extract the noun phrases from the captions with a part-of-speech (POS) tagger, provided by Spacy [12]. Then, we define the template as: "replace {entity A} with {entity B}", where entity A is replaced with other similar noun phrases, measured with the Sentence-Transformers similarity score, *i.e.*, we replace the original noun phrase with an alternative noun phrase with similarity ranging from 0.5 to 0.7 measured by all-MiniLM-L6-v2[2]. To this end, we acquire the edited image caption, which will be later used to retrieve the target image. For more implementation details, please refer to our supplementary materials.

Generation Based on Large Language Model. Given the image caption for reference image, we prompt ChatGPT (gpt-3.5-turbo) to simultaneously generate relative caption and caption of target image, with the **following prompt**: *I have an image. Carefully generate an informative instruction to edit this image content and generate a description of the edited image. I will put my image content beginning with "Image Content:". The instruction you generate should begin with "Instruction:". The edited description you generate should begin with "Edited Description:". The Instruction you generate can cover various semantic aspects, including cardinality, addition, negation, direct addressing, compare&change, comparative, conjunction, spatial relations&background, viewpoint. The edited description need to be as simple as possible. The instruction does not need to explicitly indicate which type it is. Avoid adding imaginary things. "Image Content: {}". Each time generate one instruction and one edited description only.*

3.3.2 Target Image Retrieval

With the target image captions generated by the template-based or LLM-based approach, we use a sentence transformer model to extract features from the caption, and then we perform a text-only retrieval between the target image caption and the captions of the images in the Laion-COCO pool using cosine similarity. The images with their corresponding captions to have similarity scores above the given threshold are kept as candidate target images, resulting in a scalable pipeline for constructing triplet samples, with reference image, relative caption, and target image.

4 Experiment

In this section, we first describe the experiment setups and implementation details (Sec. 4.1), then followed by ablation studies to investigate the applicability of our method and the effectiveness of the core components in our TransAgg model (Sec. 4.2), lastly, we present comparison results to the recent approaches (Sec. 4.3). **Note that**, there has been several concurrent work on composed image retrieval [2, 9, 15, 25], here, we try to compare with

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

them as fairly as we can, however, there still remain differences on some small experimental details, such as visual and text encoder, embedding dimensions, batch size, *etc*.

4.1 Experimental Setups

Training Datasets. We construct the training sets by using the data collection pipeline outlined in Section 3.3, resulting Laion-CIR-Template and Laion-CIR-LLM, depending on the adopted approaches. Both datasets contain around 16K triplets. We also combined two approaches and construct a 32K dataset, named Laion-CIR-Combined.

Evaluation Datasets. We evaluate our model on two public benchmarks, namely, CIRR [18] and FashionIQ [28]. *CIRR* comprises approximately 36K triplets that are sampled from generic images obtained from NLVR^2 [26]. To mitigate the false negative cases, the author conduct two benchmarks to demonstrate fine-grained retrieval. The first one involves a general search using the entire validation corpus as the target search space. The second focuses on a subset of six images similar to the query image, based on pre-trained ResNet152 [10] feature distance. *FashionIQ* focuses on the fashion domain and is divided into three sub categories, *Dress, Shirt* and *Toptee*. It contains more than 30k triplets. The reference and target images are matched based on similarities in their titles, and each triplet is accompanied by two annotations that are manually generated by human annotators. **Note that**, in this paper, we consider zero-shot evaluation, that is to say, we only train on our automatically constructed training set, and directly evaluate on the target benchmarks.

Evaluation Metrics. We adopt the standard metric in retrieval, *i.e.*, Recall@K, which denotes the percentage of target images being included in the top-*K* list. For CIRR, we also report Recall$_{\text{Subset}}$@K metric, which considers only the images within the subset of the query.

Implementation Details. Our framework is implemented with PyTorch. We adopt the same image pre-processing scheme as in CLIP4CIR [1], and realize the transformer-based fusion module of 2 layers with 8 heads. Regarding the training schedule, AdamW optimizer with a cosine decay is applied. The learning rate of the visual and text encoder parameters is initialized to 1e-6, while that of the remaining parameters are initialized to 1e-4. For visual and text encoders, we use pre-trained BLIP [16] w/ViT-B, ViT-B/32 CLIP [24] and ViT-L/14 CLIP [24]. The language model used in the process of Laion-CIR-Template dataset construction is all-MiniLM-L6-v2.

4.2 Ablation Study

In this section, we evaluate on FashionIQ and CIRR benchmarks, to investigate the effectiveness of our proposed dataset construction procedure, compare different pre-trained visual backbones, and ablation studies on the transformer-based fusion, adaptive aggregation.

Pretrained Backbone and Finetuning. We train our TransAgg model on Laion-CIR-Template, and explore various backbones and fine-tuning types. As shown in Table 1, it can be observed that using BLIP [16] model as the visual and text encoder yield the best performance, and fine-tuning more parameters leads better results in most cases. In the following experiments, we choose to use BLIP [16] model as our visual and text encoder.

Effectness of Individual Modules. We conduct ablation studies on transformer fusion and adaptive aggregation, as well as the different ways for constructing dataset, *i.e.*, Laion-CIR-Template, and Laion-CIR-LLM. As shown in Table 2, we can make the following observations: (i) template-based sentence editing is more effective for dataset construction, *e.g.*,

		CIRR				**FashionIQ**		
Backbone	Fine-tuning	R@1	R@5	R_{Subset}@1	Average	R@10	R@50	Average
CLIP-B/32	✘	24.46	53.61	57.81	55.71	23.91	44.68	34.30
	only text enc.	27.08	57.21	62.70	59.96	25.67	46.43	35.65
	both	29.30	60.48	63.57	62.03	25.15	46.10	35.63
CLIP-L/14	✘	25.04	53.98	55.33	54.66	28.57	48.29	38.43
	only text enc.	27.90	58.27	60.48	59.38	30.61	50.38	40.50
	both	33.04	64.39	63.37	63.88	32.63	53.65	43.14
BLIP	✘	34.89	64.75	66.34	65.55	26.95	46.10	36.53
	only text enc.	**38.10**	**68.42**	**70.34**	**69.38**	32.07	53.26	42.67
	both	37.18	67.21	69.34	68.28	**34.64**	**55.72**	**45.18**

Table 1: Generalization for different backbones and fine-tuning types on CIRR and FashionIQ. For CIRR, the average column denotes $(\text{Recall@5} + \text{Recall}_{\text{Subset}}\text{@1})/2$. For FashionIQ, we report the average Recall@10 and 50 of all three categories. Best (resp. second-best) numbers are in red (resp. blue). Refer the reader to supplementary material for more detailed comparison.

					Shirt		**Dress**		**TopTee**		**Average**	
Model	Dataset Const.	Finetune	Fusion	Aggregation	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
A1	Template	✘	✔	✘	25.22	42.89	19.88	40.16	26.77	46.81	23.96	43.29
A2	Template	✘	✘	✔	27.43	45.24	22.21	41.40	29.07	51.66	26.24	46.10
A3	Template	✘	✔	✔	28.07	45.63	21.67	41.89	31.11	50.79	26.95	46.10
B1	Template	text enc.	✔	✘	32.19	52.80	27.37	49.28	35.08	55.84	31.55	52.64
B2	Template	text enc.	✘	✔	32.43	51.42	28.56	49.73	35.03	56.20	32.01	52.45
B3	Template	text enc.	✔	✔	32.83	52.31	27.67	49.38	35.70	58.08	32.07	53.26
C1	Template	both	✔	✘	32.78	52.55	29.65	50.22	35.90	57.27	32.78	53.35
C2	Template	both	✘	✔	34.64	54.66	29.85	51.71	38.35	59.41	34.28	55.26
C3	Template	both	✔	✔	34.84	53.93	31.28	52.75	37.79	60.48	34.64	55.72
D1	LLM	✘	✔	✘	18.74	34.45	16.41	33.57	20.50	37.43	18.55	35.15
D2	LLM	✘	✘	✔	28.21	47.60	25.88	47.05	32.99	54.56	29.03	49.74
D3	LLM	✘	✔	✔	31.89	48.72	25.53	46.80	32.99	54.11	30.14	49.88
E1	LLM	text enc.	✔	✘	31.55	49.76	26.23	48.29	33.86	53.70	30.55	50.58
E2	LLM	text enc.	✘	✔	32.63	52.06	28.51	49.73	35.95	57.01	32.36	52.93
E3	LLM	text enc.	✔	✔	32.92	52.16	28.56	49.58	36.82	58.59	32.77	53.44
F1	LLM	both	✔	✘	28.85	47.99	26.62	48.14	31.06	52.01	28.84	49.38
F2	LLM	both	✘	✔	32.04	50.74	30.39	50.87	34.93	55.79	32.45	52.47
F3	LLM	both	✔	✔	34.64	53.58	30.84	51.22	37.99	59.15	34.49	54.65

Table 2: Ablation study on FashionIQ. No Fusion means we remove the transformer fusion module, and no Aggregation means we replace adaptive aggregation with a static aggregation utilizing three learnable weight parameters.

model C3 vs. F3; (ii) adaptive aggregation has a greater impact than transformer fusion, *e.g.*, model D1 vs. D2; (iii) finetuning both the text encoder and visual encoder gives better performance, similar to the observations in Table 1, *e.g.*, model B3 vs. C3. Overall, our results demonstrate positive effects of our module, regardless of the fine-tuning type.

4.3 Comparison with State-of-the-art

We train our model on the combination of both constructed datasets, and compare with various zero-shot composed image retrieval methods on CIRR and FashionIQ. As shown in Table 3, on CIRR dataset, our proposed model achieves state-of-the-art results in all metrics except for Recall@50. While on the FashionIQ dataset, our proposed TransAgg model trained on the automatically constructed dataset also falls among the top2 best models, perform-

			CIRR				FashionIQ		
Method	Zero-shot	# Training triplets	R@1	R@5	R@50	R_{Subset}@1	R@10	R@50	Average
Pic2Word [25]$_{\text{CVPR'2023}}$	✔	-	23.90	51.70	87.80	-	24.70	43.70	34.20
PALAVRA [5]$_{\text{ECCV'2022}}$	✔	-	16.62	43.49	83.95	41.61	19.76	37.25	28.51
SEARLE-XL-OTI [2]$_{\text{arXiv'2023}}$	✔	-	24.87	52.31	88.58	53.80	27.61	47.90	37.76
CompoDiff w/T5-XL [9]$_{\text{arXiv'2023}}$	✔	18m	19.37	53.81	90.85	28.96	**37.36**	50.85	44.11
CASE Pre-LaSCo.Ca. [15]$_{\text{arXiv'2023}}$	✔	360k	35.40	65.78	**94.63**	64.29	-	-	-
TransAgg (Laion-CIR-Template)	✔	16k	**38.10**	68.42	93.51	**70.34**	32.07	53.26	42.67
TransAgg (Laion-CIR-LLM)	✔	16k	36.71	67.83	93.86	66.03	32.77	53.44	43.11
TransAgg (Laion-CIR-Combined)	✔	32k	37.87	**68.88**	93.86	69.79	34.36	**55.13**	**44.75**
CLRPLANT w/OSCAR [18]$_{\text{ICCV'2021}}$	✘	-	19.55	52.55	92.38	39.20	18.87	41.53	30.20
ARTEMIS [6]$_{\text{ICLR'2022}}$	✘	-	16.96	46.10	87.73	39.99	26.05	50.29	38.17
CLIP4CIR [1]$_{\text{CVPRW'2022}}$	✘	-	38.53	69.98	95.93	68.19	38.32	61.74	50.03
BLIP4CIR+Bi [19]$_{\text{arXiv'2023}}$	✘	-	40.15	73.08	96.27	72.10	43.49	67.31	55.40
CASE [15]$_{\text{arXiv'2023}}$	✘	-	48.00	79.11	97.57	75.88	48.79	70.68	59.74

Table 3: Comparasion on CIRR test set and FashionIQ validation set. The best and second-best numbers are shown in red and blue respectively. For more detailed comparison, we refer the reader to the supplementary material.

ing competitively with the concurrent work, namely CompoDiff [9]. **Note that**, CompoDiff has been trained on over 18M triplet samples, while ours only need to train on 16k/32k, significantly more efficient than CompoDiff.

4.4 Failure Cases of Dataset Construction

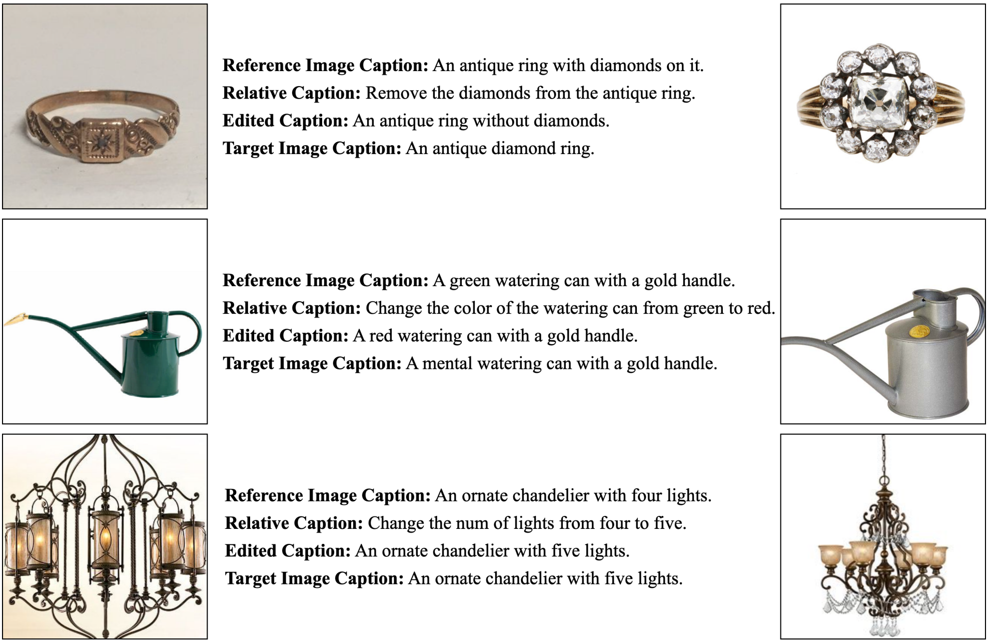
There remains limitation on our dataset construction pipeline, for instance, as shown in the 1st and 2nd row of Figure 3, while using sentence transformers for computing sentence similarity, it may not well capture the crucial information between sentences, resulting in the failure to retrieve the correct target image. Additionally, we use the Laion-COCO as our data corpus, with captions generated automatically, thus can be inaccurate.

4.5 Qualitative Results for CIR

In Figure 4, we show qualitative results on composed image retrieval, which has only been trained on the automatically constructed dataset, without finetuning on the downstream datasets. Each row includes reference image, relative caption and the top five retrieved images, where the ground truth is marked with a red box. The results demonstrate the effectiveness of our proposed method in successfully retrieving the target image. For instance, as shown in the last row, the model must be able to maintain the semantic category of the animal in the reference image, and then add a blue sky in order to retrieve the target image.

5 Conclusion

In this paper, we propose a retrieval-based pipeline for automatic CIR dataset construction, using the easily-acquired image-caption data on Internet. Specifically, we obtain two different CIR datasets based on templates and large language model. Furthermore, we propose TransAgg, a transformer-based adaptive aggregation model that can effectively integrate information across different modalities. Extensive experiments show that our method performs on par or significant above the existing state-of-the-art (SOTA) models on two public benchmarks and our zero-shot result is sometimes comparable to fully supervised ones.

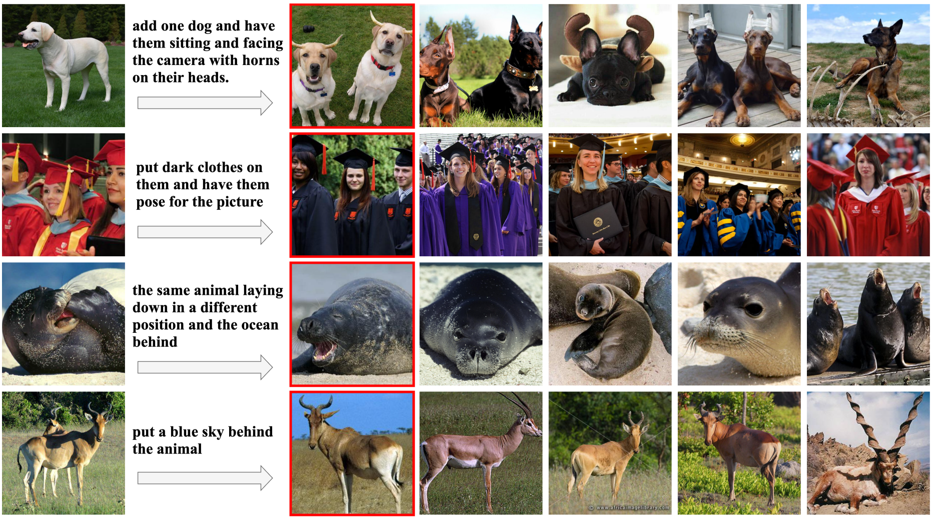


Reference Image Caption: An antique ring with diamonds on it.
Relative Caption: Remove the diamonds from the antique ring.
Edited Caption: An antique ring without diamonds.
Target Image Caption: An antique diamond ring.

Reference Image Caption: A green watering can with a gold handle.
Relative Caption: Change the color of the watering can from green to red.
Edited Caption: A red watering can with a gold handle.
Target Image Caption: A mental watering can with a gold handle.

Reference Image Caption: An ornate chandelier with four lights.
Relative Caption: Change the num of lights from four to five.
Edited Caption: An ornate chandelier with five lights.
Target Image Caption: An ornate chandelier with five lights.

Figure 3: Failure cases of dataset construction. The edited caption and target image caption in the first row have a high similarity score, but their semantic meanings are significantly different. In the second row, we intend to retrieve a red watering can, but a mental watering can is mistakenly retrieved instead. In the third row, the numerical values in both reference image caption and target image caption are incorrect.



add one dog and have them sitting and facing the camera with horns on their heads.

put dark clothes on them and have them pose for the picture

the same animal laying down in a different position and the ocean behind

put a blue sky behind the animal

Figure 4: Qualitative results on CIRR. From left to right are the reference image, relative caption and the top five retrieved images. The ground truth is marked with a red box.

Acknowledgement. This work is supported by National Key RD Program of China (No. 2022ZD0161400). We thank Zechuan Fang and Wenhao Lu for proof-reading.

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR Workshops*, 2022.

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv preprint arXiv:2303.15247*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 2021.

[5] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *ECCV*, 2022.

[6] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *CVPR*, 2019.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[9] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[14] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.

[15] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

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

[18] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021.

[19] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. *arXiv preprint arXiv:2303.16604*, 2023.

[20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[21] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *CVPR*, 2021.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[25] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023.

[26] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[27] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019.

[28] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021.