Zero-Shot Composed Image Retrieval with Textual Inversion

*Alberto Baldrati[1,2] *Lorenzo Agnolucci[1] Marco Bertini[1] Alberto Del Bimbo[1]

[1] University of Florence - Media Integration and Communication Center (MICC)
[2] University of Pisa
Florence, Italy - Pisa, Italy
[name.surname]@unifi.it

Abstract

Composed Image Retrieval (CIR) aims to retrieve a target image based on a query composed of a reference image and a relative caption that describes the difference between the two images. The high effort and cost required for labeling datasets for CIR hamper the widespread usage of existing methods, as they rely on supervised learning. In this work, we propose a new task, Zero-Shot CIR (ZS-CIR), that aims to address CIR without requiring a labeled training dataset. Our approach, named zero-Shot composEd imAge Retrieval with textuaL invErsion (SEARLE), maps the visual features of the reference image into a pseudo-word token in CLIP token embedding space and integrates it with the relative caption. To support research on ZS-CIR, we introduce an open-domain benchmarking dataset named Composed Image Retrieval on Common Objects in context (CIRCO), which is the first dataset for CIR containing multiple ground truths for each query. The experiments show that SEARLE exhibits better performance than the baselines on the two main datasets for CIR tasks, FashionIQ and CIRR, and on the proposed CIRCO. The dataset, the code and the model are publicly available at https://github.com/miccunifi/SEARLE.

1. Introduction

Given a query composed of a reference image and a relative caption, Composed Image Retrieval (CIR) [24,35] aims to retrieve target images that are visually similar to the reference one but incorporate the changes specified in the relative caption. The bi-modality of the query provides users with more precise control over the characteristics of the desired image, as some features are more easily described with language, while others can be better expressed visually. Figure 3 shows some query examples.

* Equal contribution. Author ordering was determined by coin flip.

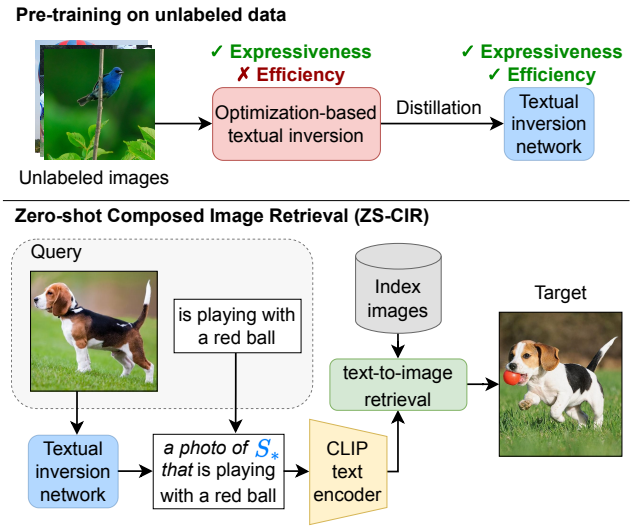

Figure 1. Workflow of our method. *Top*: in the pre-training phase, we generate pseudo-word tokens of unlabeled images with an optimization-based textual inversion and then distill their knowledge to a textual inversion network. *Bottom*: at inference time on ZS-CIR, we map the reference image to a pseudo-word S_* and concatenate it with the relative caption. Then, we use CLIP text encoder to perform text-to-image retrieval.

CIR datasets consist of triplets (I_r, T_r, I_t) composed of a reference image, a relative caption, and a target image, respectively. Creating a dataset for CIR is expensive as this type of data is not easily available on the internet, and generating it in an automated way is still very challenging. Thus, researchers must resort to manual labeling efforts. The manual process involves identifying pairs of reference and target images and writing a descriptive caption that captures the differences between them. This is a time-consuming and resource-intensive task, especially when creating large training sets. Current works tackling CIR [2, 3, 11, 22, 24] rely on supervision to learn how to combine the reference image and the relative caption. For instance, [2] proposes a fully-supervised two-stage approach that involves fine-

tuning CLIP text encoder and training a combiner network. While current approaches for CIR have shown promising results, their reliance on expensive manually-annotated datasets limits their scalability and broader use in domains different from that of the datasets used for their training.

To remove the necessity of expensive labeled training data we introduce a new task, Zero-Shot Composed Image Retrieval (ZS-CIR). In ZS-CIR, the aim is to design an approach that manages to combine the reference image and the relative caption without the need for supervised learning.

To tackle ZS-CIR, we propose an approach named zero-Shot composEd imAge Retrieval with textuaL invErsion (SEARLE) [1] that exploits the frozen pre-trained CLIP [25] vision-language model. Our method reduces CIR to standard text-to-image retrieval by mapping the reference image into a learned pseudo-word which is then concatenated with the relative caption. The pseudo-word corresponds to a pseudo-word token residing in CLIP token embedding space. We refer to this mapping process with *textual inversion*, following the terminology introduced in [13]. SEARLE involves pre-training a textual inversion network ϕ on an unlabeled image-only dataset. The training comprises two stages: an Optimization-based Textual Inversion (OTI) with a GPT-powered regularization loss to generate a set of pseudo-word tokens, and the distillation of their knowledge to ϕ. After the training, we obtain a network ϕ that is able to perform textual inversion with a single forward pass. At inference time, given a query (I_r, T_r), we use ϕ to predict the pseudo-word associated with I_r and concatenate it to T_r. Then, we leverage the CLIP common embedding space to carry out text-to-image retrieval. Figure 1 illustrates the workflow of the proposed approach.

Most available datasets for Composed Image Retrieval (CIR) focus on specific domains such as fashion [4, 14, 15, 36], birds [12], or synthetic objects [35]. To the best of our knowledge, the CIRR dataset [24] is the only one that considers natural images in an open domain. However, CIRR suffers from two main issues. First, the dataset contains several false negatives, which could lead to an inaccurate performance evaluation. Second, the queries often do not consider the visual content of the reference image, making the task addressable with standard text-to-image techniques. Furthermore, existing CIR datasets have only one annotated ground truth image for each query. To address these issues and support research on ZS-CIR, we introduce an open-domain benchmarking dataset named Composed Image Retrieval on Common Objects in context (CIRCO)[2], consisting of validation and test sets based on images from COCO [23]. Being a benchmarking dataset for ZS-CIR, the need for a large training set is removed, resulting in a significant reduction in labeling effort. To overcome the single ground truth limitation of existing CIR datasets, we propose a novel strategy that leverages SEARLE to ease the annotation process of multiple ground truths. As a result, CIRCO is the first CIR dataset with multiple annotated ground truths, enabling a more comprehensive evaluation of CIR models. We release only the validation set ground truths of CIRCO and host an evaluation server to allow researchers to obtain performance metrics on the test set[3].

The experiments show that our approach obtains substantial improvements (up to 7%) compared to the baselines on three different datasets: FashionIQ [36], CIRR [24] and the proposed CIRCO.

Recently, a concurrent work [31] has independently proposed the same task as ours. In Sec. 2 and Sec. 3 we provide a detailed comparison illustrating the numerous differences from our approach, while in Sec. 5 we show that our method outperforms this work on all the test datasets.

Our contributions can be summarized as follows:

- We propose a new task, Zero-Shot Composed Image Retrieval (ZS-CIR), to remove the need for high-effort labeled data for CIR;
- We propose a novel approach, named SEARLE, which employs a textual inversion network to tackle ZS-CIR by mapping images into pseudo-words. It involves two stages: an optimization-based textual inversion using a GPT-powered regularization loss and the training of the textual inversion network with a distillation loss;
- We introduce CIRCO, an open-domain benchmarking dataset for ZS-CIR with multiple annotated ground truths and reduced false negatives. To ease the annotation process we propose to leverage SEARLE;
- SEARLE obtains significant improvements over baselines and competing methods achieving SotA on three different datasets: FashionIQ, CIRR, and the proposed CIRCO.

2. Related Work

Composed Image Retrieval CIR belongs to the broader field of compositional learning, which has been extensively studied in various Vision and Language (V&L) tasks, such as visual question answering [1], image captioning [9, 18], and image synthesis [26, 30]. The goal of compositional learning is to generate joint-embedding features that capture relevant information from both text and visual domains.

The CIR task has been studied in various domains, such as fashion [4, 14, 15, 36], natural images [12, 24], and synthetic images [35]. It was first introduced in [35], where the authors propose to compose the image-text features using

[1] John Searle is an American philosopher who has studied the philosophy of language and how words are used to refer to specific objects.

[2] CIRCO is pronounced as /ʧirko/.

[3] https://circo.micc.unifi.it/

a residual-gating method that aims to integrate the multi-modal information. [33] proposes a training technique that integrates graph convolutional networks with existing composition methods. [22] presents two different neural network modules that consider image style and content separately. Recently, CLIP has been used to address CIR. [3] combines out-of-the-shelf image-text CLIP features using a combiner network, demonstrating their effectiveness. Later in [2], the authors add a task-oriented fine-tuning step of the CLIP text encoder, achieving state-of-the-art performance. All of these approaches are supervised and require training on a CIR dataset to effectively learn to combine the multimodal information. In contrast, our method does not involve supervision and uses an unlabeled dataset for training.

Textual Inversion In the field of text-to-image synthesis, mapping a group of images into a single pseudo-word has been proposed as a promising technique for generating highly personalized images [10, 13, 20, 28]. [13] presents an approach for performing textual inversion using the reconstruction loss of a latent diffusion model [26]. In addition to textual inversion [28] also fine-tunes a pre-trained text-to-image diffusion model.

Besides personalized text-to-image synthesis, textual inversion has also been applied to image retrieval tasks [8, 19, 31]. PALAVRA [8] tackles personalized image retrieval, in which each query is composed of multiple images depicting a shared specific subject, and the goal is to retrieve images of such subject based on an input text in natural language. PALAVRA comprises two stages: the pre-training of a mapping function and a subsequent optimization. It requires a labeled image-caption dataset for the pre-training and an input word concept for the optimization. On the contrary, we pre-train our textual inversion network on an unlabeled dataset and at inference time we do not need any additional inputs besides the reference image.

The most similar to our work is the concurrent Pic2Word [31], which tackles ZS-CIR. Pic2Word relies on a textual inversion network trained on the 3M images of CC3M [32] using only a cycle contrastive loss. Differently from this approach, we train our textual inversion network using only 3% of the data and employing a weighted sum of distillation and regularization losses. The distillation loss exploits the information provided by a set of pre-generated tokens obtained through an optimization-based textual inversion.

Knowledge Distillation Knowledge distillation is a machine learning technique where a simple model (student) is trained to mimic the behavior of a more complex one (teacher) by learning from its predictions [17]. This approach has been successfully applied to several computer vision tasks such as image classification [17, 27] and object detection [6], achieving significant improvements in terms of model compression, speed, and performance. In our work, we refer to knowledge distillation as the process of transferring the knowledge acquired by a computationally expensive optimization method (teacher) to a light neural network (student). Specifically, we train a textual inversion network to mimic the output of an optimization-based textual inversion via a distillation loss. From another perspective, our light network can be interpreted as a surrogate model of the more resource-intensive optimization method.

3. Proposed Approach

Our approach relies on CLIP (Contrastive Language-Image Pre-training (CLIP) [25]), a vision and language model that learns to align images and corresponding text captions in a common embedding space using a large-scale dataset. CLIP comprises an image encoder ψ_I and a text encoder ψ_T. Given an image I, the image encoder extracts a feature representation $i = \psi_I(I) \in \mathbb{R}^d$, where d is the size of CLIP embedding space. For a given text caption T, each tokenized word is mapped to the token embedding space $\mathcal{W}$ using a word embedding layer E_w. The text encoder ψ_T is then applied to the token embeddings to produce the textual feature representation $t = \psi_T(E_w(T)) \in \mathbb{R}^d$. CLIP is trained to ensure that the same concepts expressed in an image or through text have similar feature representations.

Given a frozen pre-trained CLIP model, our approach, named SEARLE, aims to generate a representation of the reference image that can be used as input to the CLIP text encoder. We achieve this goal by mapping the visual features of the image into a new token embedding belonging to $\mathcal{W}$. We refer to this token embedding as *pseudo-word token*, since it is not associated with an actual word, but is rather a representation of the image features in the token embedding space. Our goal is twofold. First, we need to ensure that the pseudo-word token can accurately represent the content of the reference image. In other words, the text features of a basic prompt containing the pseudo-word need to be similar to the corresponding image features. Second, we need to make sure that such a pseudo-word can effectively integrate and communicate with the text of the relative caption. Theoretically, a single image could be mapped to multiple pseudo-word tokens. In this work, we use a single one since [13] shows that it is sufficient to encode the information of an image.

The first step of SEARLE involves the pre-training of a textual inversion network ϕ on an unlabeled image-only dataset. The training is accomplished in two stages. First, we employ an Optimization-based Textual Inversion (OTI) method to iteratively generate a set of pseudo-word tokens leveraging a GPT-powered regularization loss. Second, we train ϕ by distilling the knowledge embedded in the pre-generated pseudo-word tokens. The ϕ network takes as input the features of an image extracted with CLIP image en-

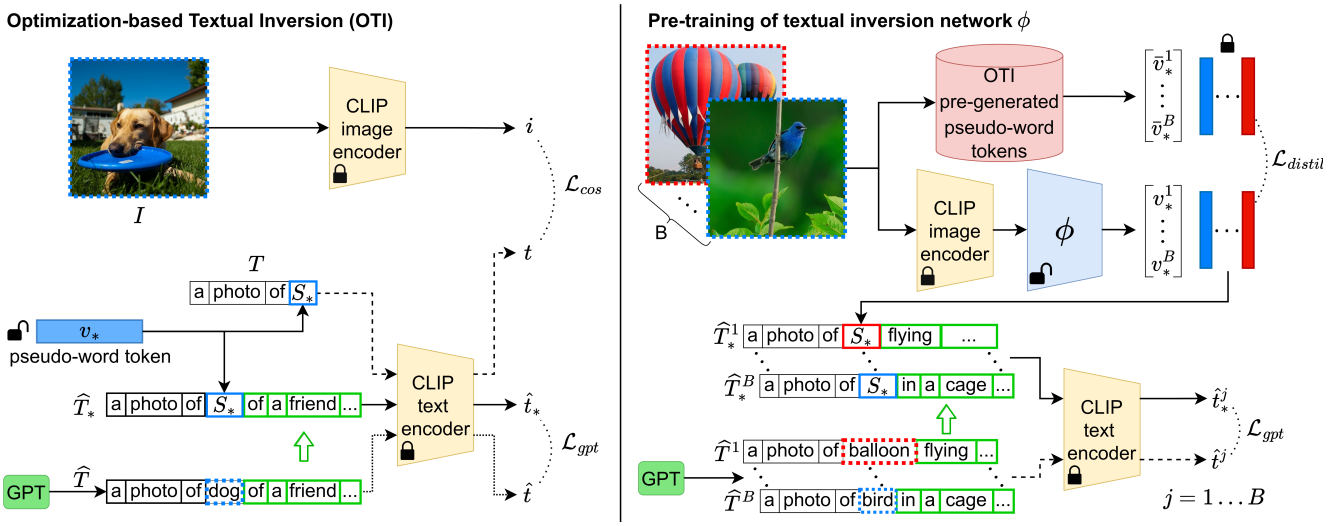

Figure 2. Overview of our approach. *Left*: we generate a pseudo-word token v_* from an image I with an iterative optimization-based textual inversion. We force v_* to represent the content of the image with a cosine loss $\mathcal{L}_{cos}$. We assign a concept word to I with a CLIP zero-shot classification and feed the prompt "a photo of {concept}" to GPT to continue the phrase, resulting in $\widehat{T}$. Let S_* be the pseudo-word associated with v_*, we build $\widehat{T}_*$ by replacing in $\widehat{T}$ the concept with S_*. $\widehat{T}$ and $\widehat{T}_*$ are then employed for a contextualized regularization with $\mathcal{L}_{gpt}$. *Right*: we train a textual inversion network ϕ on unlabeled images. Given a set of pseudo-word tokens pre-generated with OTI, we distill their knowledge to ϕ through a contrastive loss $\mathcal{L}_{distil}$. We regularize the output of ϕ with the same GPT-powered loss $\mathcal{L}_{gpt}$ employed in OTI. B represents the number of images in a batch.

coder and outputs the corresponding pseudo-word token in a single forward pass.

At inference time, CIR involves a query (I_r, T_r) representing the input reference image and relative caption, respectively. We predict the pseudo-word token v_* corresponding to the reference image as $v_* = \phi(I_r)$. Let S_* be the pseudo-word associated with v_*. To effectively integrate the visual information of I_r with T_r, we construct the template "a photo of S_* that {*relative caption*}" and extract its features using the CLIP text encoder. Notably, these text features comprise both textual and visual information, thus offering a multimodal representation of the reference image and its corresponding relative caption. Using the extracted text features, we perform a standard text-to-image retrieval by querying an image database. An overview of the workflow of our approach is illustrated in Fig. 1.

Conceptually, OTI and ϕ perform the same operation, *i.e.* a textual inversion that maps the visual features of an image into a pseudo-word token. Therefore, we could directly employ OTI at inference time without the need for ϕ. However, OTI requires a non-negligible amount of time to be carried out, while ϕ is significantly more efficient. Since OTI is proven to be powerful in generating effective pseudo-word tokens (see Sec. 5), we propose to distill their knowledge into a feed-forward network. Our goal is to retain the expressive power of OTI while achieving a negligible inference time. In the following, we refer to SEARLE when we employ ϕ to generate the pseudo-word token, and to SEARLE-OTI when we directly use OTI for inference.

3.1. Optimization-based Textual Inversion (OTI)

Given an image I, we adopt an optimization-based approach that performs the textual inversion by optimizing the pseudo-word token $v_* \in \mathcal{W}$ for a fixed amount of iterations. The left section of Fig. 2 shows an overview of OTI.

We start by randomly initializing the pseudo-word token v_* and associating the pseudo-word S_* to it. We build a template sentence T such as "a photo of S_*" and feed it to the CLIP text encoder ψ_T, obtaining $t = \psi_T(T)$. Similarly to [8], we randomly sample T from a pre-defined set of templates. Given an image I, we extract its features using the CLIP image encoder ψ_I, resulting in $i = \psi_I(I)$.

Since our goal is to obtain a pseudo-word token v_* that encapsulates the informative content of I, we rely on CLIP common embedding space and minimize the gap between the image and text features. To achieve our aim, we leverage a cosine CLIP-based loss:

$$\mathcal{L}_{cos} = 1 - \cos(i, t) \tag{1}$$

However, $\mathcal{L}_{cos}$ alone is insufficient to generate a pseudo-word that can interact with other words in CLIP dictionary. Indeed, similarly to [8], we observe that $\mathcal{L}_{cos}$ forces the pseudo-word token into sparse regions of CLIP token embedding space that are different from those observed during CLIP's training. An analogous effect has also been studied in GAN inversion works [34, 38]. Consequently, the pseudo-word token is unable to effectively communicate with other tokens. To overcome this issue, we propose a

novel regularization technique that constrains the pseudo-word token to reside on the CLIP token embedding manifold enhancing its reasoning capabilities (see Sec. 5.2 for more details). First, we perform a zero-shot classification of the image I relying on CLIP zero-shot capabilities. The vocabulary used to classify the images is taken from the ∼20K class names of the Open Images V7 dataset [21]. In particular, we assign the most similar k different class names to each image, where k is a hyperparameter. We will refer to the class names used to classify the images as *concepts*, *i.e*. we associate each image to k different concepts. Thanks to the zero-shot classification, different from [8], we do not require the concepts as input.

Once we have a pool of concepts associated with the image, we generate a phrase using a lightweight GPT [5] model. In each iteration of the optimization, we randomly sample one of the k concepts associated with the image I and feed the prompt "a photo of {concept}" to GPT. Being an autoregressive generative model, GPT is capable of continuing this prompt in a meaningful way. For instance, given the concept "dog", the GPT-generated phrase could be $\widehat{T}$ = "a photo of dog that was taken by his owner, who is a friend of mine". In practice, since the vocabulary is known a priori, we pre-generate all the GPT phrases for all the concepts in the vocabulary (see supplementary material for more details). Starting from $\widehat{T}$, we define $\widehat{T}_*$ by simply replacing the concept with the pseudo-word S_*, obtaining $\widehat{T}_*$ = "a photo of S_* that was taken…". We use the CLIP text encoder to extract the features of both phrases, ending up with $\hat{t} = \psi_T(\widehat{T})$ and $\hat{t}_* = \psi_T(\widehat{T}_*)$. Finally, we employ a cosine loss to minimize the gap between the features:

$$\mathcal{L}_{gpt} = 1 - \cos\left(\hat{t}, \hat{t}_*\right) \tag{2}$$

The idea behind this loss is to apply a contextualized regularization that pushes v_* toward the concept while taking into account a broader context. Indeed, the GPT-generated phrases are more elaborated and thus similar to the relative captions used in CIR, compared to a generic pre-defined prompt. This way we enhance the ability of v_* to interact with human-generated text such as the relative captions.

The final loss that we use for OTI is:

$$\mathcal{L}_{OTI} = \lambda_{cos}\mathcal{L}_{cos} + \lambda_{OTIgpt}\mathcal{L}_{gpt} \tag{3}$$

where λ_{cos} and λ_{OTIgpt} are the loss weights.

3.2. Textual Inversion Network ϕ Pre-training

We find OTI effective for obtaining pseudo-words that both encapsulate the visual information of the image and interact with actual words. However, being an iterative optimization-based method, it requires a non-negligible amount of time to be carried out. Therefore, we propose a method for training a textual inversion network ϕ capable of predicting the pseudo-word tokens in a single forward pass by distilling knowledge from a set of OTI pre-generated tokens. In other words, ϕ acts as a surrogate model of OTI, *i.e*. a faster and less computationally heavy approximation of it. An overview of the pre-training phase is illustrated in the right part of Fig. 2.

Our aim is to obtain a single model that can invert images of any domain and that does not need labeled data for training. In particular, we consider an MLP-based textual inversion network ϕ with three linear layers, each followed by a GELU [16] activation function and a dropout layer.

Given an unlabeled pre-training dataset $\mathcal{D}$, we start by applying OTI to each image. While this operation may be time-consuming, it is a one-time requirement, making it tolerable. We end up with a set of pseudo-word tokens $\overline{\mathcal{V}}_* = \{\bar{v}_*^j\}_{j=1}^N$, where N is the number of images of $\mathcal{D}$. Our aim is to distill to ϕ the knowledge acquired by OTI and embedded in $\overline{\mathcal{V}}_*$. Starting from an image $I \in \mathcal{D}$, we extract its features using the CLIP visual encoder obtaining $i = \psi_I(I)$. We employ ϕ to predict the pseudo-word token $v_* = \phi(i)$. We minimize the distance between the predicted pseudo-word token v_* and its corresponding pre-generated token $\bar{v}_* \in \overline{\mathcal{V}}_*$ and, at the same time, maximize the discriminability of each token. To this end, we employ a symmetric contrastive loss inspired by SimCLR [7, 8] as follows:

$$\begin{aligned}\mathcal{L}_{distil} = &\frac{1}{B}\sum_{k=1}^{B} -\log\frac{e^{(c(\bar{v}_*^k, v_*^k)/\tau)}}{\sum_{j=1}^{B} e^{(c(\bar{v}_*^k, v_*^j)/\tau)} + \sum_{j\neq k} e^{(c(v_*^k, v_*^j)/\tau)}} \\ &-\log\frac{e^{(c(v_*^k, \bar{v}_*^k)/\tau)}}{\sum_{j=1}^{B} e^{(c(v_*^k, \bar{v}_*^j)/\tau)} + \sum_{j\neq k} e^{(c(\bar{v}_*^k, \bar{v}_*^j)/\tau)}}\end{aligned} \tag{4}$$

Here, $c(\cdot)$ denotes the cosine similarity, B is the number of images in a batch, and τ is a temperature hyperparameter.

To regularize the training of ϕ we follow the same technique described in Sec. 3.1 and rely on $\mathcal{L}_{gpt}$.

The final loss used to update the weights of ϕ is

$$\mathcal{L}_{\phi} = \lambda_{distil}\mathcal{L}_{distil} + \lambda_{\phi gpt}\mathcal{L}_{gpt} \tag{5}$$

where λ_{distil} and $\lambda_{\phi gpt}$ are the loss weights.

The training of our textual inversion network ϕ is fully unsupervised as we do not rely on any labeled data. Indeed, we utilize raw images, different from [8], which also needs captions. In particular, we employ the unlabeled test split of the ImageNet1K [29] dataset as $\mathcal{D}$ to pre-train ϕ. It contains 100K images without any given label. In comparison to PALAVRA [8] and Pic2Word [31], our method utilizes significantly fewer data, approximately the 10%, and the 3%, respectively. We chose this dataset as it includes real-world images with a high variety of subjects. We believe that other similar datasets could serve our purpose.

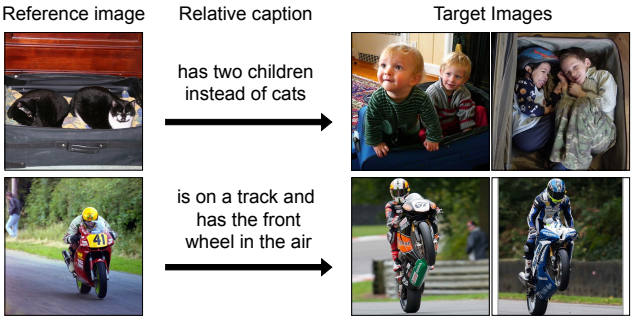

Figure 3. Examples of CIR queries and ground truths in CIRCO.

4. CIRCO dataset

We recall that CIR datasets consist of triplets (I_r, T_r, I_t) composed of a reference image, relative caption, and target image (*i.e*. the ground truth), respectively.

Existing datasets contain several false negatives, *i.e*. images that could be potential ground truths for the query but are not labeled as such. Indeed, since each query triplet contains only a target, all the other images are considered negatives. In addition, most datasets revolve around specialized domains such as fashion [4, 14, 15, 36], birds [12], or synthetic objects [35]. To the best of our knowledge, CIRR [24] is the only dataset based on real-life images in an open domain. During the data collection process, CIRR builds sets of 6 visually similar images in an automated way. Then, the queries are created such that the reference and the target images belong to the same set and so as to avoid the presence of false negatives within the set. The flaw with this approach is that it does not guarantee the absence of false negatives in the whole dataset. Furthermore, despite the visual similarity, the difference between images belonging to the same set can possibly be not easily describable with a relative caption and require an absolute description. This reduces the importance of the visual information of the reference image and makes the retrieval addressable with standard text-to-image methods (see Sec. 5.1 for more details).

To address these issues, we introduce an open-domain benchmarking dataset named Composed Image Retrieval on Common Objects in context (CIRCO). It is based on real-world images from COCO 2017 unlabeled set [23] and is the first dataset for CIR with multiple ground truths. Contrary to CIRR, we start from a single pair of visually similar images and write a relative caption. In addition, we provide an auxiliary annotation with the shared characteristics of the reference and target images to clarify ambiguities. Then, we propose to employ our approach to retrieve the top 100 images according to the query and combine them with the top 50 images most visually similar to the target one. Finally, we select the images that are valid matches for the query. We estimate that this approach allows us to reduce the percentage of missing ground truths in the dataset to less than 10%. CIRCO comprises a total of 1020 queries, randomly divided into 220 and 800 for the validation and test set, respectively, with an average of 4.53 ground truths per query. We use all the 120K images of COCO as the index set, thus providing significantly more distractors than the 2K images of CIRR test set. Figure 3 shows some query examples. More details on the CIRCO dataset and a more comprehensive comparison with CIRR are provided in the supplementary material.

To mitigate the problem of false negatives, most works evaluate the performance using Recall@K, with K set to quite large values (*e.g*. 10, 50 [36]), thus making a fine-grained analysis of the models difficult. CIRR addresses the issue by employing also $\text{Recall}_{\text{Subset}}@K$, which considers only the images in the same set of the reference and target ones. Thanks to our multiple ground truths, we can rely on a more fine-grained metric such as mean Average Precision (mAP), which takes into account also the ranks in which the ground truths are retrieved. In particular, we use mAP@K, with K ranging from small to quite large values.

5. Experimental Results

We test our approach following the standard evaluation protocol [2, 24] on three datasets: FashionIQ [36], CIRR [24] and the proposed CIRCO. In particular, we employ the three categories of FashionIQ validation split and the test sets of CIRR and CIRCO. We introduce two variants of our approach: SEARLE, based on CLIP ViT-B/32, and SEARLE-XL, using CLIP ViT-L/14 as the backbone. In the following, we refer to ViT-B/32 and ViT-L/14 as B/32 and L/14, respectively. For the sake of space, we provide the implementation details and the qualitative results in the supplementary material.

5.1. Quantitative Results

We compare our approach with several zero-shot baselines and competing methods, including: 1) *Text-only*: the similarity is computed using only the CLIP features of the relative caption; 2) *Image-only*: retrieves the most similar images to the reference one; 3) *Image + Text*: the CLIP features of the reference image and the relative caption are summed together; *4) Captioning*: we substitute the pseudo-word token with the caption of the reference image generated with a pre-trained captioning model [37][4] 5) *PALAVRA* [8]: a textual inversion-based two-stage approach with a pre-trained mapping function and a subsequent optimization of the pseudo-word token; 6) *Pic2Word* [31]: forward-only method employing a pre-trained textual inversion network.

For PALAVRA, we use CLIP B/32 as the backbone following the original paper, while for Pic2Word, we report

[4]https://huggingface.co/laion/CoCa-ViT-B-32-laion2B-s13B-b90k

Backbone	Method	Shirt		Dress		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
B/32	Image-only	6.92	14.23	4.46	12.19	6.32	13.77	5.90	13.37
	Text-only	19.87	34.99	15.42	35.05	20.81	40.49	18.70	36.84
	Image + Text	13.44	26.25	13.83	30.88	17.08	31.67	14.78	29.60
	Captioning	17.47	30.96	9.02	23.65	15.45	31.26	13.98	28.62
	PALAVRA [8]	21.49	37.05	17.25	35.94	20.55	38.76	19.76	37.25
	SEARLE-OTI	**25.37**	41.32	17.85	**39.91**	24.12	45.79	22.44	42.34
	SEARLE	24.44	**41.61**	**18.54**	39.51	**25.70**	**46.46**	**22.89**	**42.53**
L/14	Pic2Word† [31]	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	SEARLE-XL-OTI	**30.37**	**47.49**	**21.57**	**44.47**	**30.90**	**51.76**	**27.61**	**47.90**
	SEARLE-XL	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23

Table 1. Quantitative results on FashionIQ validation set. Best and second-best scores are highlighted in bold and underlined, respectively. † indicates results from the original paper.

Backbone	Method	Recall@K				Recall$_{\text{Subset}}$@K		
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
B/32	Image-only	6.89	22.99	33.68	59.23	21.04	41.04	60.31
	Text-only	21.81	45.22	57.42	81.01	**62.24**	**81.13**	**90.70**
	Image + Text	11.71	35.06	48.94	77.49	32.77	56.89	74.96
	Captioning	12.46	35.04	47.71	77.35	42.94	65.49	80.36
	PALAVRA [8]	16.62	43.49	58.51	83.95	41.61	65.30	80.94
	SEARLE-OTI	**24.27**	53.25	66.10	88.84	54.10	75.81	87.33
	SEARLE	24.00	**53.42**	**66.82**	**89.78**	54.89	76.60	88.19
L/14	Pic2Word† [31]	23.90	51.70	65.30	87.80	–	–	–
	SEARLE-XL-OTI	**24.87**	52.31	**66.29**	88.58	**53.80**	74.31	86.94
	SEARLE-XL	24.24	**52.48**	**66.29**	**88.84**	53.76	**75.01**	**88.19**

Table 2. Quantitative results on CIRR test set. Best and second-best scores are highlighted in bold and underlined, respectively. † indicates results from the original paper. – denotes results not reported in the original paper.

the results provided by the authors when available. Considering our approach, we provide the results of both SEARLE and SEARLE-OTI.

FashionIQ Table 1 shows the results on FashionIQ. With the B/32 backbone, SEARLE achieves comparable performance with SEARLE-OTI. It is worth noting that SEARLE provides a significant efficiency gain without compromising performance. Our approach outperforms the baselines in both versions. In particular, the enhancement over Captioning underscores that the pseudo-word token embeds more information than the actual words comprising the generated caption. Considering the L/14 backbone, SEARLE-XL significantly improves over Pic2Word, up to a 7% gain in the Recall@50 for the Dress category. We recall that the two approaches are directly comparable as they both rely on a single forward of a pre-trained network with no subsequent optimization, but our model is trained with 3% of the data. However, we notice a gap with the performance obtained by SEARLE-XL-OTI. We suppose it is due to the very narrow domain of FashionIQ, which is quite different from the natural images of the pre-training dataset we use for training ϕ. To support our theory, we trained a version of ϕ using the FashionIQ training set as the pre-training dataset, obtaining an average R@10 and R@50 of 27.95 and 49.24 in the validation set, respectively. These results are comparable with SEARLE-XL-OTI, confirming our hypothesis. More details are provided in the supplementary material.

CIRR In Tab. 2 we report the results for CIRR test set. The Text-only baseline obtains the best performance on Recall$_{\text{Subset}}$ and outperforms Image-only and Image+Text in the global metrics. These results highlight a major flaw in CIRR: the relative captions are often not actually relative in practice. Specifically, we find that the reference image may not provide useful information for retrieval, and may even have a detrimental effect, as also observed in [31]. This is especially true when considering only the subset of images that comprises the reference and target ones, as the visual information is very similar. Indeed, the Recall$_{\text{Subset}}$

results of Image-only correspond to random guessing as the retrieval in the subset involves only five images.

In this dataset, we notice that for both backbones the results obtained by our approach with OTI and ϕ are comparable, showing the effectiveness of our distillation process. It is worth noticing that there is no performance gap between the B/32 and L/14 versions, and in some cases, the B/32 even outperforms the L/14. We improve over PALAVRA and Pic2Word when using the same backbones. Unfortunately, we can not compare our performance on $\text{Recall}_{\text{Subset}}$ with Pic2Word as the authors do not report their results.

CIRCO Table 3 shows the results on CIRCO test set. First, we can notice how, contrary to FashionIQ and CIRR, Image+Text achieves better results than Image-only and Text-only. This shows how CIRCO comprises queries in which the reference image and the relative caption are equally important to retrieve the target images. Second, SEARLE significantly improves over all the baseline methods, even outperforming Pic2Word, which employs a larger backbone. SEARLE-XL would achieve the best results, but we do not consider them completely fair as it was employed to retrieve the images that the multiple ground truths were selected from. Still, we report them for completeness and as a baseline for future works that will use our dataset for testing.

		mAP@K			
Backbone	Method	$K=5$	$K=10$	$K=25$	$K=50$
B/32	Image-only	1.34	1.60	2.12	2.41
	Text-only	2.56	2.67	2.98	3.18
	Image + Text	2.65	3.25	4.14	4.54
	Captioning	5.48	5.77	6.44	6.85
	PALAVRA [8]	4.61	5.32	6.33	6.80
	SEARLE-OTI	<u>7.14</u>	<u>7.83</u>	<u>8.99</u>	<u>9.60</u>
	SEARLE	**9.35**	**9.94**	**11.13**	**11.84**
L/14	Pic2Word [31]	8.72	9.51	10.64	11.29
	SEARLE-XL-OTI	<u>10.18</u>	<u>11.03</u>	<u>12.72</u>	<u>13.67</u>
	SEARLE-XL	**11.68**	**12.73**	**14.33**	**15.12**

Table 3. Quantitative results on CIRCO test set. Best and second-best scores are highlighted in bold and underlined, respectively.

5.2. Ablation Studies

We conduct ablation studies to evaluate the individual contributions of the components in our approach. To avoid confounding effects, we assess the two main components of the proposed method separately. Specifically, we evaluate the textual inversion network ϕ while keeping fixed the set of OTI pre-generated tokens obtained with the method described in Sec. 3.1. As ϕ distills their knowledge, we assume that the more informative they are (*i.e*. the better OTI performs), the better the results obtained by ϕ will be.

		CIRR			FashionIQ	
Abl.	Method	R@1	R@5	R@10	R@10	R@50
OTI	w/o GPT reg	<u>21.63</u>	<u>50.51</u>	<u>64.07</u>	<u>21.34</u>	39.70
	random reg	21.09	50.42	63.84	20.90	<u>40.24</u>
	w/o reg	19.30	46.81	59.96	17.86	35.99
	SEARLE-OTI	**23.54**	**53.93**	**67.69**	**22.44**	**42.34**
ϕ	cos distil	<u>23.75</u>	<u>53.19</u>	<u>67.21</u>	21.59	39.41
	w/o distil	22.24	50.06	62.71	19.41	38.39
	w/o reg	22.41	53.00	66.90	<u>22.83</u>	<u>42.02</u>
	SEARLE	**25.09**	**55.18**	**68.79**	**22.89**	**42.53**

Table 4. Ablation studies on CIRR and FashionIQ validation sets. For FashionIQ, we consider the average recall. Best and second-best scores are highlighted in bold and underlined, respectively.

We perform the ablation studies on CIRR and FashionIQ validation sets and report the results for the main evaluation metrics. In particular, for FashionIQ we report the average scores. For simplicity, we consider only the version of our approach with the B/32 backbone.

Optimization-based textual inversion (OTI) We ablate the regularization loss used during the optimization process: 1) *w/o GPT reg*: we regularize with a prompt comprising only the concept word, without the GPT-generated suffix; 2) *random reg*: we additionally substitute the concept word with a random word; 3) *w/o reg*: we completely remove the regularization loss.

The upper section of Tab. 4 shows the results. As a different loss corresponds to a different speed of convergence, for each ablation experiment we use a tailored number of optimization iterations and report the best performance. We notice that some kind of regularization is essential to make the pseudo-word tokens reside on the CLIP token embedding manifold and communicate effectively with CLIP vocabulary tokens. Specifically, the proposed GPT-based regularization loss allows the pseudo-word tokens to interact with text resembling human-written text, thus enhancing their communication with the relative caption and the performance of retrieval. This is especially true for CIRR, as the relative captions are more elaborated and have a more varied vocabulary.

Textual inversion network ϕ We ablate the losses used during the pre-training: 1) *cos distil*: we employ a cosine distillation loss instead of a contrastive one; 2) *w/o distil*: we replace $\mathcal{L}_{distil}$ with the cycle contrastive loss employed by [8], which directly considers the image and text features; 3) *w/o reg*: we remove the $\mathcal{L}_{gpt}$ regularization loss.

We report the results in the lower part of Tab. 4. The contrastive version of the distillation loss proves to be more effective than the cosine one. Compared to the cycle con-

trastive loss, our distillation-based loss achieves significantly superior performance, showing how learning from OTI pre-generated tokens is more fruitful than from raw images. Finally, although the pre-generated pseudo-word tokens are already regularized, we notice that our GPT-based regularization loss is still beneficial for training ϕ.

6. Conclusion

In this paper we introduce a new task, Zero-Shot Composed Image Retrieval (ZS-CIR), that aims to address CIR without the need for an expensive labeled training dataset. Our approach, named SEARLE, involves the pre-training of a textual inversion network that leverages a distillation loss to retain the expressive power of an optimization-based method while achieving a significant efficiency gain. We also present a new open-domain benchmarking dataset for CIR, Composed Image Retrieval on Common Objects in context (CIRCO). CIRCO is the first dataset for CIR that contains multiple ground truths for each query. Both versions of our approach achieve superior performance over baselines and competing methods on popular datasets such as CIRR and FashionIQ and on the proposed CIRCO.

In future work, we plan to investigate the potential of our method in the personalized image generation task. In particular, we believe that the proposed GPT-based regularization loss could improve the ability of a generative model to consider input text for synthesizing personalized objects.

Acknowledgments This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4959–4968, 2022.

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining CLIP-based features. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21466–21474, 2022.

[4] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 663–676. Springer, 2010.

[5] 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 *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[6] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[8] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2022.

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587, 2020.

[10] Giannis Daras and Alex Dimakis. Multiresolution textual inversion. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[11] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity. In *Proc. of International Conference on Learning Representations (ICLR)*, 2022.

[12] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Proc. of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 708–717, 2019.

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proc. of International Conference on Learning Representations (ICLR)*, 2023.

[14] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[15] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1463–1471, 2017.

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[18] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989, 2022.

[19] Bruno Korbar and Andrew Zisserman. Personalised clip or: how to find your vacation videos. In *Proc. of British Machine Vision Association (BMVA)*, 2022.

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020.

[22] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, 2021.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[24] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, 2021.

[25] 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 *Proc. of International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[33] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. Rtic: Residual learning for text and image composition using graph convolutional network. *arXiv preprint arXiv:2104.03015*, 2021.

[34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[35] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6439–6448, 2019.

[36] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317, 2021.

[37] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, Aug 2022, 2022.

[38] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 592–608. Springer, 2020.