Zero-Shot Composed Image Retrieval with Textual Inversion

Supplementary Material

*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

S1. Implementation Details

For the Optimization-based Textual Inversion (OTI), we perform 350 iterations with a learning rate of $2e{-}2$. We set the loss weights λ_{cos} and λ_{OTIgpt} in Eq. (3) to 1 and 0.5, respectively. We adopt an exponential moving average with 0.99 decay. Regarding the training of the textual inversion network ϕ, we train for 100 and 50 epochs for SEARLE and SEARLE-XL, respectively, with a learning rate of $1e{-}4$ and a batch size of 256. We set the loss weights λ_{distil} and $\lambda_{\phi gpt}$ in Eq. (5) to 1 and 0.75, respectively. The temperature τ in Eq. (4) is set to 0.25. For both OTI and ϕ, we employ the AdamW optimizer [7] with weight decay 0.01. During OTI we set the number of concept words k associated with each image to 15, while during the training of ϕ to 150. We tune each hyperparameter individually with a grid search on the CIRR validation set. With a single A100 GPU, OTI for SEARLE-XL takes $\sim$30 seconds for a single image and $\sim$1 second per image with batch size 256. The training of ϕ for SEARLE-XL takes 6 hours in total on a single A100 GPU. Throughout all the experiments, we use the pre-processing technique proposed in [1]. We use Mixed-precision [8] to save memory and increase computational efficiency. For retrieval, we normalize both the query and index set features to have a unit L_2-norm.

To generate the phrases used for the regularization with $\mathcal{L}_{gpt}$, we employ the GPT-Neo-2.7B model with 2.7 billion parameters developed by EleutherAI. For each of the 20,932 class names of the Open Images V7 dataset [4], we generate 256 phrases a priori with a temperature of 0.5, constraining the length to a maximum of 35 tokens. The whole process takes approximately 12 hours to complete on a single NVIDIA A100 GPU. We need to perform this operation only once, making the time requirements tolerable.

Since the FashionIQ dataset provides two relative captions for each triplet, during inference, we concatenate them using the conjunction "and". To ensure our approach remains unaffected by the order of concatenation, we employ both potential concatenation orders and afterward average the resulting features.

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

Layer	Module
Input	nn.Linear(d, $d*4$)
GELU	nn.GELU
Dropout	nn.Dropout(0.5)
Hidden	nn.Linear($d*4$, $d*4$)
GELU	nn.GELU
Dropout	nn.Dropout(0.5)
Output	nn.Linear($d*4$, d_w)

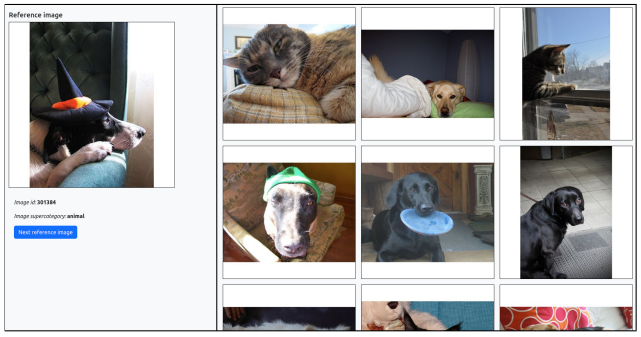
Table S1: Pytorch-style description of the textual inversion network ϕ. d and d_w represent the dimension of the CLIP feature space and token embedding space $\mathcal{W}$, respectively.

S1.1. ϕ Architecture

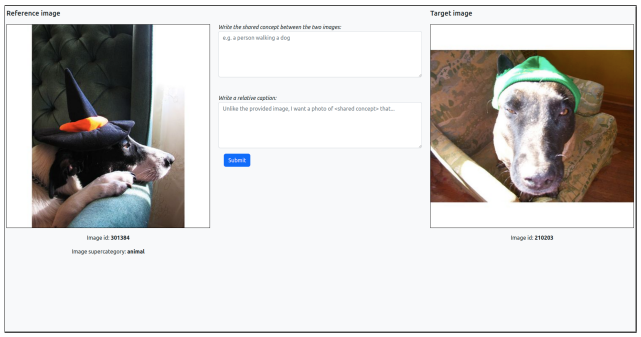
Table S1 illustrates the details of the architecture of the textual inversion network ϕ. For the B/32 backbone, the dimension of the CLIP feature space and token embedding space $\mathcal{W}$, respectively d and d_w, are both equal to 512. For the L/14 backbone d and d_w both equal 768.

S2. CIRCO Dataset

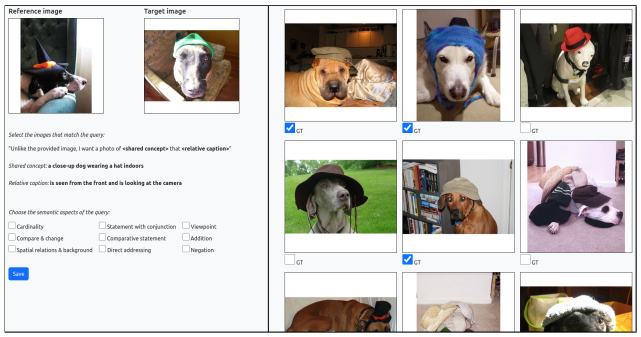
In this section, we provide details about the annotation process of the proposed CIRCO dataset, which consists of two phases. In the first phase, we build the triplets composed of a reference image, a relative caption, and a single target image. In the second one, we extend each triplet by annotating additional ground truths. The whole annotation process has been carried out by the authors of this paper. We also report a detailed analysis of CIRCO, along with a comparison with CIRR [6]. CIRCO is available at https://github.com/miccunifi/CIRCO.



(a) Interface for selecting the reference-target image pair. On the left, there is a randomly sampled reference image and a button to skip it. On the right, a gallery of images displays the candidate target images.



(b) Interface for writing the shared concept and the relative caption of a given reference-target image pair.



(c) Interface for selecting the multiple ground truths. On the left, the tool displays the current triplet and checkboxes to assign the semantic aspects covered by the relative caption. On the right, there is a gallery of images from which we select the ground truths.

Figure S1: Screenshots of the annotation tool user interface. (a) and (b) correspond to the first annotation phase, while (c) is related to the second one.

S2.1. Triplets Annotation

CIRCO consists of images belonging to COCO 2017 [5] unlabeled set, which comprises 123,403 images. We chose this dataset as it contains open-domain real-life images depicting a large variety of subjects. We rely on the unlabeled set of COCO instead of the training one as the latter is often employed as a pre-training dataset, and we do not want any model to have a prior on the images. Every object in each image of COCO labeled sets is associated with a supercategory. The supercategories are 12 and are as follows: *person*, *animal*, *sports*, *vehicle*, *food*, *accessory*, *electronic*, *kitchen*, *furniture*, *indoor*, *outdoor*, and *appliance*.

We start by associating every image of the unlabeled set to a supercategory relying on CLIP ViT-L/14 zero-shot classification capabilities. We assume that each image is classified according to its main subject. Our objective is to obtain a rough estimation of the content of each image to later build a balanced dataset. Indeed, we create the queries so that we evenly distribute the reference images of CIRCO among the supercategories. This balancing process is crucial, as we have noticed a significant domain bias within the COCO images. For instance, some objects like stop signs or fire hydrants are over-represented.

Figure S1a shows the annotation tool we employed for creating the triplets. The tool randomly samples a reference image and displays it next to a gallery of 50 candidate target images. Since CIR requires the differences between the reference and target images to be describable with a relative caption, they should be similar but with appreciable disparities. Therefore, the candidate target images are the most visually similar to the reference one according to the CLIP features. To avoid near-identical images, we filter out those with a similarity higher than 0.92. The tool allows the annotators to skip the current reference image if there is no suitable target in the gallery. Otherwise, when the annotator selects a target image, the tool displays the user interface shown in Fig. S1b. In this stage, the user must write the *shared concept*, *i.e.* the shared characteristics between the reference and target images. This concept is collected to clarify possible ambiguities. For instance, the shared concept for the reference-target pair shown in Fig. S1b is "a close-up dog wearing a hat indoors". Finally, the annotator writes the relative caption from the prefix "Unlike the provided image, I want a photo of {shared concept} that". To build a more challenging dataset containing truly relative captions, we ensure they do not refer to the subjects mentioned in the shared concept. Indeed, we want the subject of the relative caption to be inferred from the reference image.

At the end of this phase, we have 1020 triplets composed of a reference image, a relative caption, and a single target image.

S2.2. Multiple Ground Truths Annotation

For each triplet, we want to label as ground truths all the images beside the target one that are valid matches for the corresponding query. Figure S1c shows the annotation tool we relied on in this phase. We provide the annotator with

the starting triplet, and they have to select the ground truths from a gallery of images. In addition, the user also has to choose the semantic aspects of the query (see Sec. S2.3 for more details).

We propose leveraging our approach to ease the annotation process by employing it to retrieve the images from which we select the multiple ground truths. Specifically, we use SEARLE-XL to obtain the pseudo-word S_* associated with the reference image. Then, we perform text-to-image retrieval with the query "a photo of {shared concept} S_* that {relative caption}". During the annotation process, we also include the shared concept in the query, as it improves performance. In fact, considering the single ground truth triplets obtained in Sec. S2.1, with the shared concept we achieve a Recall@100 of 82.15, compared to 66.25 without it. In the gallery of images from which we select the multiple ground truths, the annotation tool provides the top 100 retrieved images with our approach and the top 50 ones most similar to the target one.

At the end of this phase, we have 4624 ground truths, of which 4097 were retrieved employing our method and 527 using the similarity with the target image. Given that with SEARLE we achieve a Recall@100 of 82.15, by approximation, we can assume that in the top 100 retrieved images, there will be an average of 82.15% of the total ground truths. This implies that the estimated number of ground truths in the entire dataset is $4097/0.8215 \approx 4,987$. Since we labeled 4624 images as ground truth, we can infer that the annotated ones are $4,624/4,987 \approx 92.7\%$ of the total. Therefore, we estimate that this annotation strategy allows us to reduce the percentage of missing ground truths in the dataset to less than 10%.

Thanks to this second annotation phase, we labeled additional $4624 - 1020 = 3604$ ground truths that otherwise would have been false negatives. Moreover, it allows us to estimate the fraction of missing ground truths in the dataset. It is not possible to make this estimation for CIR datasets with a single ground truth, such as FashionIQ [10] and CIRR [6], as they have not any information about the total number of ground truths. Indeed, their annotation process stops after building the triplets.

S2.3. Dataset Analysis

CIRCO comprises 1020 queries with a total of 4624 ground truths, 4.53 per query on average. Figure S2 shows a histogram representing the number of queries per number of ground truths. The maximum number of ground truths annotated for a query is 21, while the modal value is 2.

The relative captions are composed of an average of 10.4 words. Figure S3 illustrates a word cloud of the most frequent words in the annotations. Following CIRR [6], we analyze the semantic aspects covered by the relative captions. Table S2 reports the results. We observe that the

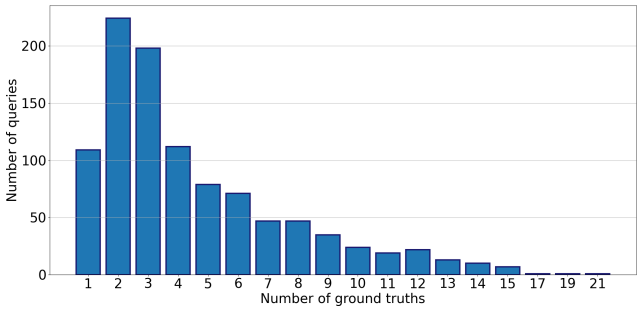


Figure S2: Histogram of the number of queries per number of ground truths for the CIRCO dataset.

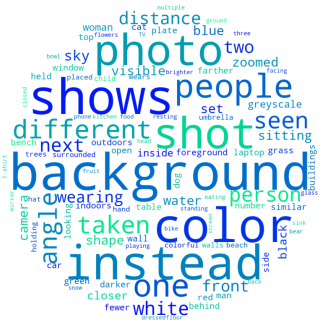


Figure S3: Vocabulary of the most frequent words in the relative captions scaled by frequency.

Semantic Aspect	Coverage (%)		
	CIRCO	CIRR	FashionIQ
Cardinality	22.1	29.3†	–
Addition	26.7	15.2†	15.7†
Negation	12.2	11.9†	4.0†
Direct Addressing	50.7	57.4†	49.0†
Compare & Change	32.0	31.7†	3.0†
Comparative Statement	35.3	51.7†	32.0†
Statement with Conjunction	75.1	43.7†	19.0†
Spatial Relations & Background	45.7	61.4†	–
Viewpoint	22.6	12.7†	–
Avg. Caption Length (words)	10.4	11.3†	5.3†

Table S2: Analysis of the semantic aspects covered by the relative captions. † indicates results taken from [6]. – denotes no reported results.

average length of the captions and the distribution of the semantic concepts is comparable with CIRR. However, in CIRCO about 75% of the annotations are composed of multiple statements, more than the ∼43% of CIRR, thus revealing a higher complexity. We provide a query example for each semantic aspect in Fig. S4.

CIRR [6] validation and test sets contain 4K triplets each. We recall that during the data collection process,

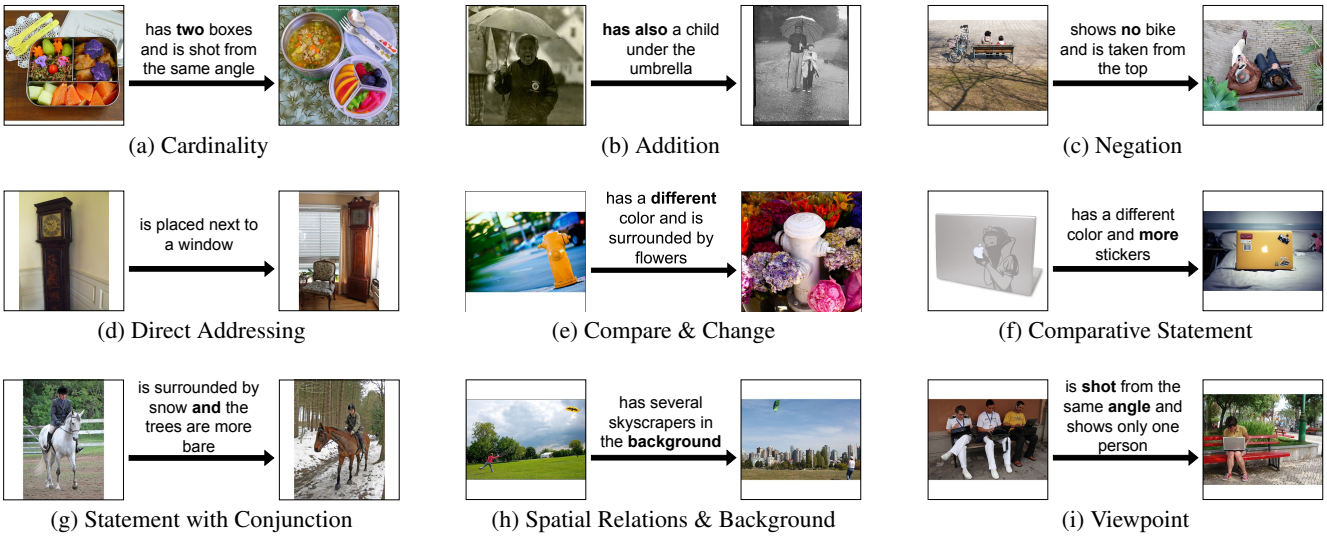


Figure S4: Examples of queries of the proposed CIRCO dataset for different semantic aspects. For simplicity, we report only one ground truth. We highlight the keywords of each semantic aspect in bold.

CIRR constructs subsets of 6 visually similar images in an automated way according to the features of a ResNet152 [3]. Then, the queries are built such that the reference and the target images belong to the same subset. However, despite the feature similarity, the images of the subset often depict very different subjects. This makes writing a relative caption unfeasible for a human annotator and leads to an absolute description of the target image. We report some examples of this issue in Fig. S5. We observe that, for instance, the annotator needs to rely on an absolute caption to describe the differences between an image depicting some pillows and one with a group of penguins.

CIRCO annotation strategy aims to address the issue mentioned above. Indeed, we let the annotators choose the reference-target pair without any constraint. This way, we ensure that the annotators only write captions that are actually relative, thereby increasing the quality of the dataset.

CIRCO comprises 1020 queries, randomly divided into 220 and 800 for the validation and test set, respectively. Compared to CIRR, we have fewer queries, but our two-phase annotation strategy ensures higher quality, reduced false negatives, and the availability of multiple ground truths. Moreover, we provide significantly more distractors than the 2K images of the CIRR test set by employing all the 120K images of COCO as the index set. Figure S4 shows some query examples.

S2.4. Dataset Evaluation

Thanks to the reduced false negatives and multiple ground truths, for performance evaluation on CIRCO we adopt the fine-grained metric mean Average Precision (mAP). In particular, we compute mAP@K, with K ranging from 5 to 50, as follows:

$$\text{mAP@}K = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\min(K, G_n)}\sum_{k=1}^{K} P@k * \text{rel@}k \quad \text{(S1)}$$

where N is the number of queries, G_n is the number of ground truths of the n-th query, $P@k$ is the precision at rank k, rel@k is a relevance function. The relevance function is an indicator function that equals 1 if the image at rank k is labeled as positive and equals 0 otherwise.

S3. Additional Experimental Results

S3.1. Visual Information in v_*

To evaluate the effectiveness of the pseudo-word tokens in capturing visual information, we conduct an image retrieval experiment. Specifically, we investigate whether the pseudo-word tokens are able to retrieve the corresponding images.

Given an input image I, we perform textual inversion to obtain the corresponding pseudo-word token v_* and its associated pseudo-word S_*. We build a generic prompt using the pseudo-word S_* such as "a photo of S_*". We extract the text features using CLIP text encoder ψ_T and use them to query an image database. If the pseudo-word token manages to capture the visual content of the input image, we expect the image I to be the top-ranked result.

Table S3 shows the results for Image Retrieval (IR) next to the corresponding ones for Composed Image Retrieval (CIR). We carry out all the experiments on the CIRR validation set. We report the results obtained by all the ablation studies on the regularization loss for both OTI and ϕ.

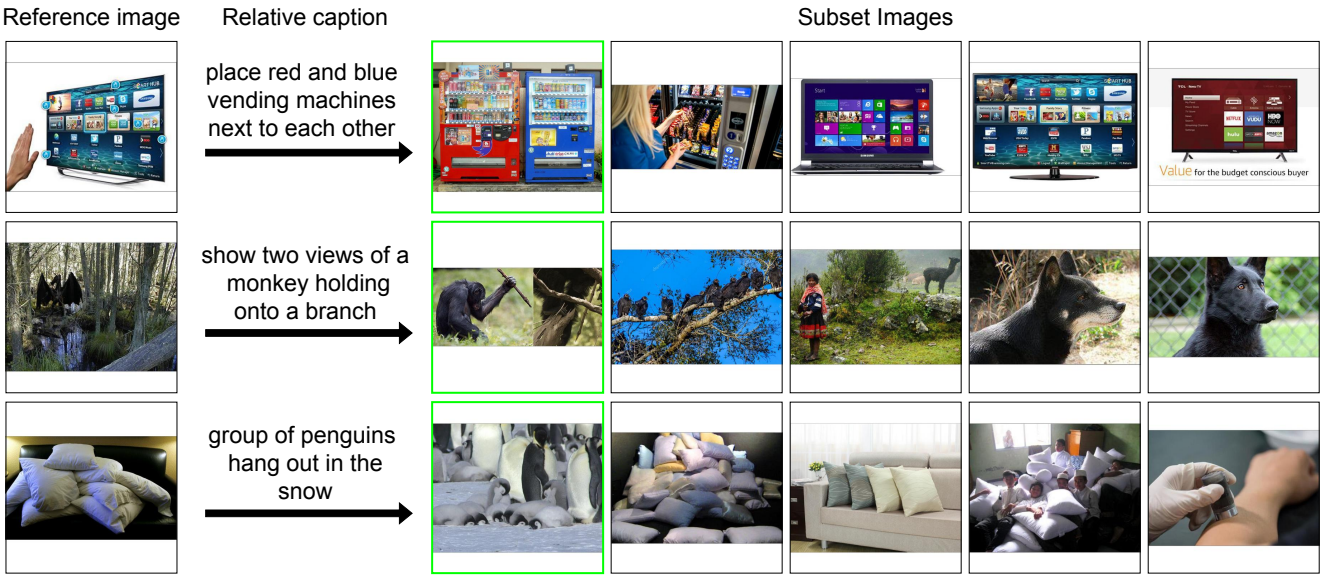


Figure S5: Examples of queries belonging to the CIRR dataset [6]. The subsets of images depict very different subjects and the relative captions do not consider the reference images. We highlight the target image with a green border.

Ablation	Method	IR			CIR			
		R@1	R@3	R@5	R@1	R@5	R@10	R@50
OTI	w/o GPT reg	99.58	99.90	99.96	21.63	50.51	64.07	88.04
	random reg	99.03	99.21	99.55	21.09	50.42	63.84	87.30
	w/o reg	99.72	99.91	**100**	19.30	46.81	59.96	84.74
	SEARLE-OTI	**99.81**	**100**	**100**	**23.54**	**53.93**	**67.69**	**90.31**
ϕ	w/o reg	**99.35**	**100**	**100**	22.41	53.00	66.90	89.95
	SEARLE	98.66	99.86	**100**	**25.09**	**55.18**	**68.79**	**90.82**

Table S3: Evaluation of the visual information embedded in v_* for different regularization techniques on CIRR validation set. IR and CIR stand for Image Retrieval and Composed Image Retrieval, respectively. Best and second-best scores are highlighted in bold and underlined, respectively.

We observe that, regardless of the regularization technique, v_* captures the visual information of the image effectively. However, we notice a significant improvement in the performance of CIR when using the proposed GPT-powered loss. This proves how our regularization technique enhances the ability of the pseudo-word tokens to interact with the actual words that compose the relative caption.

S3.2. Training ϕ on Different Datasets

We perform several experiments to investigate the impact of the pre-training dataset we employed for training the textual inversion network ϕ. In particular, besides the version of ϕ trained on the test split of ImageNet1K, we also train two variants using the training sets of CIRR and FashionIQ, named SEARLE-CIRR and SEARLE-FIQ, respectively. Notably, we rely only on the raw images of these datasets without considering the associated labels. This way, our approach is still unsupervised. The FashionIQ and CIRR training sets comprise 45,429 and 16,939 images, respectively. For both datasets, the number of images is lower than the 100K contained in the ImageNet1K test split. To assess the generalization capabilities of our approach, we test all three variants of the ϕ network on both the FashionIQ and CIRR validation sets.

Table S4 shows the results on the FashionIQ validation set. We notice how SEARLE-FIQ and SEARLE-XL-FIQ improve the performance over the ImageNet-based variants. We suppose this gain is related to the narrow domain of FashionIQ, which has a much more limited scope than the natural images of ImageNet. Furthermore, both SEARLE-FIQ and SEARLE-XL-FIQ outperform the OTI-based methods, showing the effectiveness of our distillation-based approach. Regarding the CIRR variant of ϕ, we observe that with the B/32 backbone, it

Backbone	Method	Shirt		Dress		Toptee		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
B/32	SEARLE-FIQ	**26.15**	**43.57**	**20.62**	**42.69**	**27.89**	**49.36**	**24.89**	**45.21**
	SEARLE-CIRR	24.44	41.24	18.29	38.92	25.40	45.69	22.71	41.95
	SEARLE	24.44	<u>41.61</u>	<u>18.54</u>	39.51	<u>25.70</u>	<u>46.46</u>	<u>22.89</u>	<u>42.53</u>
	SEARLE-OTI	<u>25.37</u>	41.32	17.85	<u>39.91</u>	24.12	45.79	22.44	42.34
L/14	SEARLE-XL-FIQ	<u>29.54</u>	**48.04**	**23.15**	**46.36**	**31.16**	**53.34**	**27.95**	**49.24**
	SEARLE-XL-CIRR	25.22	42.44	19.29	41.00	27.38	48.29	23.96	43.91
	SEARLE-XL	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	SEARLE-XL-OTI	**30.37**	<u>47.49</u>	<u>21.57</u>	<u>44.47</u>	<u>30.90</u>	<u>51.76</u>	<u>27.61</u>	<u>47.90</u>

Table S4: Quantitative results of our approach on FashionIQ validation set varying the pre-training dataset of ϕ. Best and second-best scores are highlighted in bold and underlined, respectively.

Backbone	Method	Recall@K				Recall$_{\text{Subset}}$@K		
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
B/32	SEARLE-FIQ	23.99	53.53	67.33	89.17	**57.07**	**78.21**	**89.28**
	SEARLE-CIRR	**25.28**	**55.32**	<u>68.74</u>	**90.89**	<u>55.18</u>	76.27	<u>88.07</u>
	SEARLE	<u>25.09</u>	<u>55.18</u>	**68.79**	<u>90.82</u>	54.84	<u>76.63</u>	87.95
	SEARLE-OTI	23.54	53.93	67.69	90.31	51.26	73.02	86.51
L/14	SEARLE-XL-FIQ	24.04	53.67	66.92	88.07	**55.80**	**77.13**	**88.26**
	SEARLE-XL-CIRR	**24.61**	53.79	<u>67.06</u>	88.85	<u>54.39</u>	<u>75.68</u>	87.37
	SEARLE-XL	24.11	<u>54.25</u>	66.95	**89.48**	53.77	75.29	<u>87.56</u>
	SEARLE-XL-OTI	<u>24.40</u>	**54.68**	**68.02**	<u>89.09</u>	52.27	74.53	86.80

Table S5: Quantitative results of our approach on CIRR validation set varying the pre-training dataset of ϕ. Best and second-best scores are highlighted in bold and underlined, respectively.

achieves comparable performance to SEARLE, while with the L/14 one, the results are worse than those of SEARLE-XL but still noteworthy. Considering that the CIRR training set consists of only 16K images, we can infer that our approach is effective even in a low-data regime.

In Tab. S5, we report the results on the CIRR validation set. Interestingly, we notice that SEARLE-FIQ and SEARLE-XL-FIQ manage to generalize to a broader domain achieving promising performance. In addition, due to the domain similarity between the CIRR and ImageNet datasets, we observe that the CIRR-based versions of ϕ obtain comparable results to the ImageNet-based ones.

S3.3. CIRCO

Table S6 shows the results on the CIRCO validation set. We observe that the same considerations made for the test set (Tab. 3 in Sec. 5.1) hold true. We still report these results for completeness.

Additionally, we evaluate the performance of SEARLE on the CIRCO test set considering only the first annotated ground truth (end of Sec. S2.1) in Tab. S7. Since we leverage SEARLE-XL only during the second annotation phase, we can fairly compare with the baselines. We employ Recall@K as the evaluation metric. We observe that even with a single annotated ground truth, the Image + Text baseline outperforms Image-only and Text-only. This confirms that we need both the reference image and the relative caption to retrieve the target image. SEARLE and SEARLE-XL achieve the best performance improving over all the other methods and baselines. In particular, we notice that SEARLE-XL significantly outperforms Pic2Word while leveraging the same CLIP backbone.

S3.4. Comparison with Supervised Baselines

We compare our zero-shot approach to a supervised method. Specifically, we consider Combiner [1], which integrates image and text CLIP features using a combiner network. Since we also rely on an out-of-the-box CLIP model, we believe Combiner constitutes the most similar method to ours among the supervised ones. We train Combiner both on FashionIQ and CIRR training sets with the official repository using the B/32 backbone. To evaluate the generalization capabilities of supervised models, we test both Combiner versions on FashionIQ and CIRR validation sets and compare them with our zero-shot method. We report the results in Tab. S8. As expected, when the training and test-

Backbone	Method	mAP@K			
		$K=5$	$K=10$	$K=25$	$K=50$
B/32	Image-only	1.61	2.16	2.73	3.10
	Text-only	2.96	3.29	3.74	3.89
	Image + Text	2.63	3.58	4.52	4.94
	Captioning	5.12	5.31	6.38	6.77
	PALAVRA [2]	5.15	6.13	7.20	7.78
	SEARLE-OTI	6.61	7.24	8.30	8.97
	SEARLE	**6.82**	**7.83**	**9.15**	**9.77**
L/14	Pic2Word [9]	7.92	9.02	10.18	10.83
	SEARLE-XL-OTI	**10.85**	**12.15**	**13.63**	**14.46**
	SEARLE-XL	10.09	11.15	12.83	13.60

Table S6: Quantitative results on CIRCO validation set. Best and second-best scores are highlighted in bold and underlined, respectively.

Backbone	Method	Recall@K			
		$K=5$	$K=10$	$K=25$	$K=50$
B/32	Image-only	3.88	6.63	14.13	22.00
	Text-only	4.75	6.63	9.50	13.50
	Image + Text	8.25	14.13	25.50	34.75
	Captioning	10.25	14.33	21.38	29.00
	PALAVRA [2]	12.63	20.63	32.00	41.75
	SEARLE-OTI	16.88	25.00	37.00	46.38
	SEARLE	**19.75**	**28.00**	**39.50**	**50.63**
L/14	Pic2Word [9]	16.13	24.38	37.25	46.50
	SEARLE-XL-OTI	22.75	32.00	45.13	**58.00**
	SEARLE-XL	**23.50**	**32.63**	**45.25**	55.63

Table S7: Quantitative results on CIRCO test set considering only the first annotated ground truth. Best and second-best scores are highlighted in bold and underlined, respectively.

ing datasets correspond, Combiner achieves the best results. However, we observe that both the supervised models struggle to generalize to different domains, as also noticed by [9]. On the contrary, SEARLE obtains noteworthy performance on both datasets in a zero-shot manner. Therefore, as we do not require an expensive manually-annotated training set, our approach proves to be more scalable and more suitable for the broad applicability of CIR.

S3.5. Qualitative Results

Figure S6 shows the qualitative results for FashionIQ and CIRR. We observe that SEARLE manages to integrate the visual features of the reference image and the text features of the relative caption to retrieve the correct image. On the contrary, the baselines either focus too much on the reference image or the relative caption. The second and fourth rows of the figure highlight the problem of false negatives in existing CIR datasets. Indeed, SEARLE retrieves images that are valid matches for the query but are not labeled as such.

Method	CIRR				FashionIQ	
	R@1	R@5	R@10	R@50	R@10	R@50
Combiner-FIQ [1]	19.88	48.05	61.11	85.51	**32.96**	**54.55**
Combiner-CIRR [1]	**32.24**	**65.46**	**78.21**	**95.19**	20.91	40.40
SEARLE-OTI	23.54	53.93	67.69	90.31	22.44	42.34
SEARLE	25.09	55.18	68.79	90.82	22.89	42.53

Table S8: Comparison with supervised baselines on CIRR and FashionIQ validation sets. Combiner-FIQ and Combiner-CIRR denote the models from [1] trained on FashionIQ and CIRR, respectively. For FashionIQ, we consider the average recall. Best and second-best scores are highlighted in bold and underlined, respectively.

In Fig. S7 we compare the top-5 images retrieved by SEARLE and PALAVRA for two queries belonging to CIRCO. SEARLE manages to retrieve more relevant images compared to PALAVRA. We stress that without the second phase of the annotation process (see Sec. S2.2) the additional ground truths would have been false negatives.

References

[1] 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. 1, 6, 7

[2] 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. 7

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[4] 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. 1

[5] 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. 2

[6] 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. 1, 3, 5

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

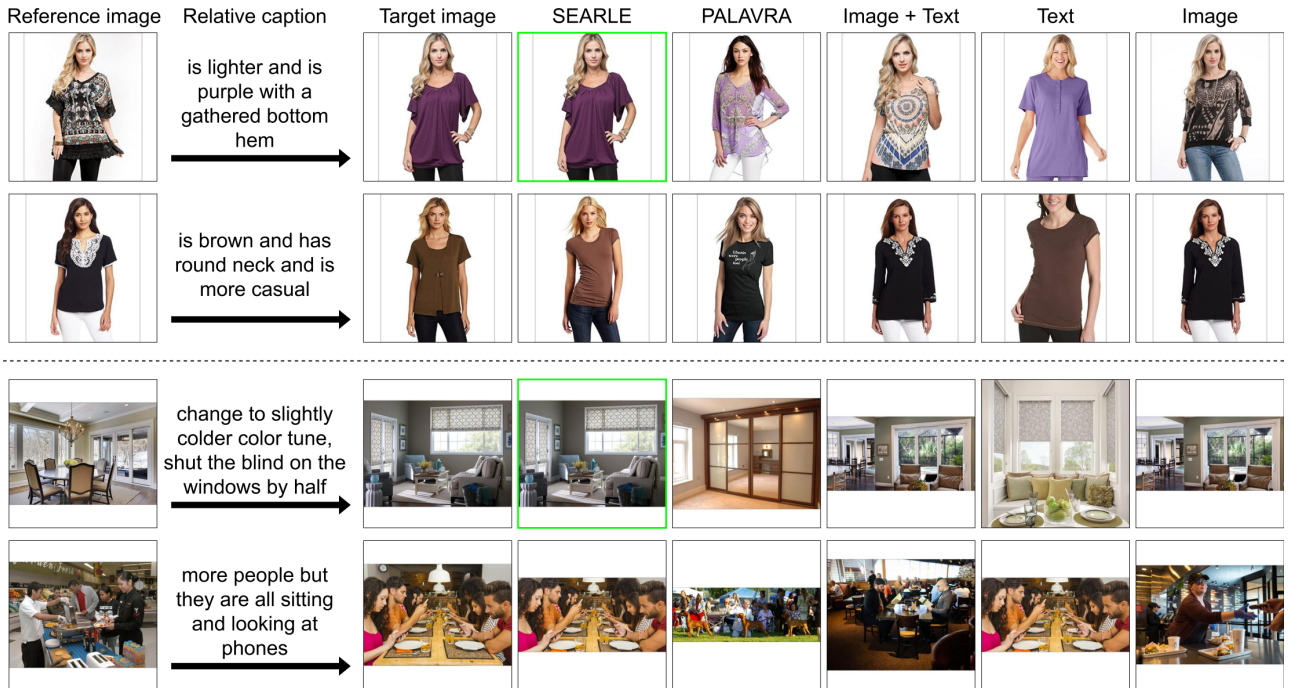


Figure S6: Qualitative results for the FashionIQ (top) and CIRR (bottom) datasets. For a clearer visualization, we do not consider the reference image in the retrieval results. We highlight with a green border when the retrieved image is the labeled ground truth. The second and fourth rows show examples in which SEARLE retrieves a false negative.

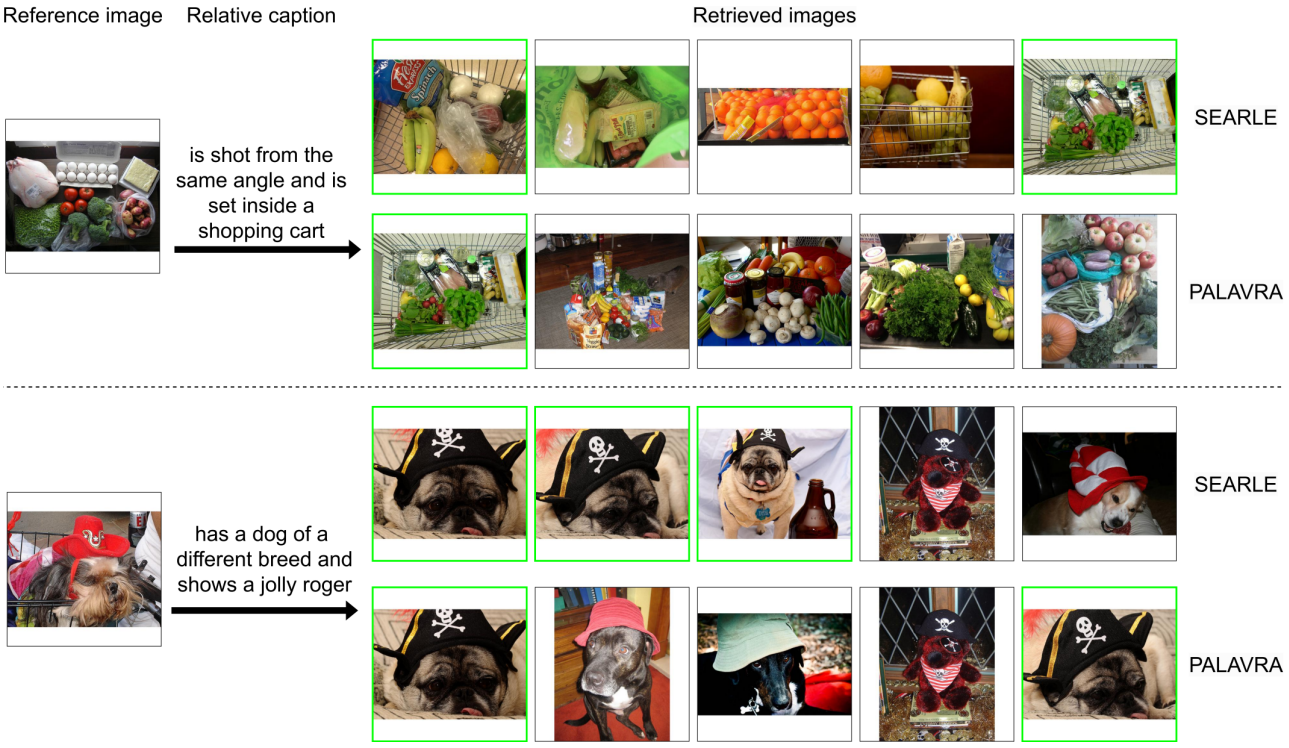


Figure S7: Qualitative results for the CIRCO dataset. We compare the top-5 retrieved images of SEARLE and the best-performing baseline. We highlight ground truths with a green border.

[8] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *International Conference on Learning Representations*, 2018. 1

[9] 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. 7

[10] 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. 3