Data Roaming and Quality Assessment for Composed Image Retrieval

Matan Levy[1], **Rami Ben-Ari**[2], **Nir Darshan**[2], **Dani Lischinski**[1]

[1]The Hebrew University of Jerusalem, Israel
[2]OriginAI, Israel
levy@cs.huji.ac.il

Abstract

The task of Composed Image Retrieval (CoIR) involves queries that combine image and text modalities, allowing users to express their intent more effectively. However, current CoIR datasets are orders of magnitude smaller compared to other vision and language (V&L) datasets. Additionally, some of these datasets have noticeable issues, such as queries containing redundant modalities. To address these shortcomings, we introduce the Large Scale Composed Image Retrieval (LaSCo) dataset, a new CoIR dataset which is ten times larger than existing ones. Pre-training on our LaSCo, shows a noteworthy improvement in performance, even in zero-shot. Furthermore, we propose a new approach for analyzing CoIR datasets and methods, which detects modality redundancy or necessity, in queries. We also introduce a new CoIR baseline, the Cross-Attention driven Shift Encoder (CASE). This baseline allows for early fusion of modalities using a cross-attention module and employs an additional auxiliary task during training. Our experiments demonstrate that this new baseline outperforms the current state-of-the-art methods on established benchmarks like FashionIQ and CIRR.

1 Introduction

Recent progress in the field of multi-modal learning (Radford et al. 2021; Lu et al. 2019) has been reflected in various downstream tasks, *e.g.*, VQA (Antol et al. 2015; Levy, Ben-Ari, and Lischinski 2022), Visual Dialog (Das et al. 2017), Image captioning (Li et al. 2020), Image Retrieval (in its variations) (Baldrati et al. 2022; Li et al. 2021) and Composed Image Retrieval (CoIR) (Vo et al. 2019; Chen, Gong, and Bazzani 2020). Image Retrieval (IR) is a longstanding task that aims to find a desired image in a large corpus, given a user query. While content-based image retrieval uses a single visual modality to convey the user intent (Barz and Denzler 2021; Dubey 2021; Zhong, Chen, and Qian 2020), providing a bi-modal query can mitigate miss-interpretations. In CoIR the gist and attributes are more succinctly described visually, and further intent is specified via a lingual modality (Han et al. 2017; Isola, Lim, and Adelson 2015; Vo et al. 2019; Guo et al. 2018; Wu et al. 2021; Liu et al. 2021; Couairon et al. 2022). Some examples of CoIR queries and their results are shown in Figures 1 and 5.

Despite the progress in foundation models and new CoIR architectures, curating a dataset for CoIR remains a challenging chore, where the samples are triplets of query-image, accompanying transition-text, and the target image, serving as the ground-truth answer. There are several existing datasets for CoIR that differ significantly from each other. Vo et al. (2019) propose a dataset of rendered images of simple 3D scenes. Other existing datasets suffer from a small amount of data, and some are domain-specific (*e.g.*, shoes (Guo et al. 2018)), while in others, the lingual modality is limited by transition-text used as a class label (Isola, Lim, and Adelson 2015), or generated based on pairs of image captions that differ by a single word (Han et al. 2017). Another dataset was labeled based on previous vision and language (V&L) models (Couairon et al. 2022). Recently, Wu et al. (2021) introduced *FashionIQ*, another domain-specific dataset for CoIR, which gained popularity (*e.g.* (Goenka et al. 2022; Lee, Kim, and Han 2021; Liu et al. 2021)) and contains human-annotated labels.

In addition to their small scale, shortcomings of these datasets include: 1) Not all acceptable target images for a given query are labeled as such, leading to incorrect count of false-negatives (*e.g.*, Fig. 5); 2) Lack of visual complexity (due to restriction to a specific domain); and 3) Modality redundancy, *i.e.* target images may often be retrieved using solely the query text, when descriptive enough to ignore the query image. We further refer to this issue as (lack of) *compositonality*: where the target should be determined by its query constituents combining the lingual and visual cues.

To break out from the previous domain-specific datasets to general and natural image domain, Liu et al. (2021) introduced the new *CIRR* (Composed Image Retrieval on Real-life images) dataset that contains open domain *natural images*, taken from NLVR2 (Suhr et al. 2019). To the best of our knowledge, *CIRR* is the only existing dataset for CoIR based on natural images with human annotated open-language texts. Despite attempts to reduce the false-negatives and relying on direct human labeling, *CIRR* still has two major shortcomings: 1) The image corpus is small, alleviating retrieval; 2) Modality redundancy still exists (see Sec. 5), as well as false-negatives (according to our observations), reflecting the challenge in creating a "clean" dataset.

In this work, we introduce a new large scale dataset for CoIR, dubbed LaSCo (Large Scale Composed Image Re-

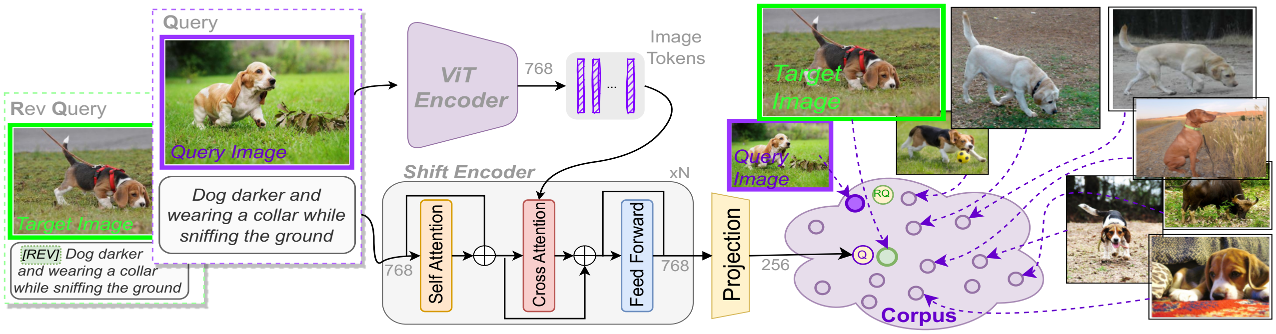

Figure 1: An overview of our CASE baseline. The query-image is fed to a ViT encoder. The query-text is handled by our shift encoder *i.e*. a BERT encoder with intermediate cross-attention layers, that receives the ViT output and fuses textual and visual information. The resulting $[CLS]$ token is then pooled and projected into a shared 256D space (circled Q). Finally, the K nearest cosine-similarity neighbors of Q are retrieved. For each training query, we also create a Reverse-Query by switching the roles of query and target images. A learnable special token $[REV]$ in our transformer, handles the *prediction of the query-image* (circled RQ) as the Reverse-Query task.

trieval dataset). To construct it with minimal human effort, we employ a simple and effective methodology to rephrase labels from an existing large scale VQA dataset (Goyal et al. 2017) into a form suited for CoIR. LaSCo contains an open and broad domain of natural images and rich text. Compared to *CIRR*, it has ×10 more queries, ×2 more unique tokens and ×17 more corpus images. We then propose a new approach for analyzing CoIR datasets and methods, which detects modality redundancy or necessity, in queries. Our analysis demonstrates that LaSCo shows a significantly smaller bias towards a single modality for retrieval.

SoTA approach (Baldrati et al. 2022) employ CLIP (Radford et al. 2021) to separately encode the textual and the visual query, followed by feature vector concatenation and a learnable projection head. We experiment with an additional end-to-end learnable baseline, which leverages the layers of BLIP's (Li et al. 2022) image-grounded text encoder and enables early interactions between textual tokens and individual areas (patches) in the image. This baseline, dubbed CASE (*Cross-Attention driven Shift Encoder*), builds upon bi-modal *Cross-Attention* to create an *Encoder* that *Shifts* the query-image towards the target in the embedding space (see Fig. 1). CASE is trained using a novel bi-directional objective, which we refer to as *Reverse-Query* (RQ), where the query-image is predicted from the target-image and the query-text. Being based on BLIP, CASE uses a lower dimension latent vector of 256D, reducing retrieval complexity by a factor of ×2.5 over previous SoTA. Furthermore, pre-training our baseline on LaSCo improves performance on *CIRR* dataset and even surpasses previous SoTA methods without training on *CIRR* (at zero-shot).

In summary, our key contributions in this paper are:

- *LaSCo:* A new large-scale, domain-free CoIR dataset, a few orders of magnitude larger than existing datasets.
- *Data Roaming:* A simple methodology for automatically generating CoIR triplets from an existing VQA dataset.
- *Modality Redundancy:* A method for analyzing redundancy between modalities in existing CoIR datasets.
- *CASE:* A new BLIP-based baseline, featuring early fusion and a novel bi-directional training objective, that achieves SoTA performance with a large gap on *FashionIQ*, *CIRR* and LaSCo benchmarks.

2 Related Work

Data Roaming: A major challenge in many multi-modal tasks, such as text-video and text-audio retrieval, is the lack of large-scale training data. Due to the complexity involved in creating multi-modal datasets such as text to image (Young et al. 2014; Lin et al. 2014) or video retrieval (Xu et al. 2016; Chen and Dolan 2011), several studies suggest using raw narrated video footage (Miech et al. 2019) for video retrieval or altering the narration to create a dataset for Visual Question Answering (VQA) (Yang et al. 2021). Other works try to enhance existing datasets, *e.g*., COCO captioning (Agrawal et al. 2019) to more diversity and object categories. In this line of work, Nagrani et al. (2022) propose a new video mining pipeline which involves automatically transferring captions from image captioning datasets to video clips, to create a new large-scale, weakly labelled audio-video captioning dataset. Nevertheless, for CoIR models to ever function in the wild, a much larger variety of visual concepts must be learned, ideally from less annotated datasets. In this paper (Section 3) we propose a methodology for leveraging VQA2.0 (Goyal et al. 2017), a large existing and labeled dataset for the VQA task.

CoIR datasets consist of triplets of query image, transition text and a target image. In order to differentiate these datasets from text-to-image and image-to-image retrieval, these triplets should ideally satisfy a condition where reaching the target image in the corpus will necessarily require both modalities. In this paper we further suggest an analysis tool for the "quality" of certain dataset, measured by "modality redundancy". We show that our newly generated large scale dataset exhibits higher quality, compared to CIRR on natural images, and is on-par with the domain-specific manually annotated FashionIQ dataset.

Composed Image Retrieval: CoIR methods commonly learn a shared embedding space between the text and visual modalities. These methods often differ by encoding models, *e.g.*, (Vo et al. 2019) that uses ResNet and LSTM and learns a shift encoder. Other methods suggest different attention mechanisms, *e.g.*, Chen, Gong, and Bazzani (2020); Hosseinzadeh and Wang (2020). Goenka et al. (2022) focuses on specific domain characteristics such as the fashion domain (with FashionIQ dataset). Different fusion strategies between visual and textual modalities has gained high attention suggesting early (Hosseinzadeh and Wang 2020) and late (Delmas et al. 2022) fusion methods.

Recent works leverage VLM foundation models, *e.g.* Baldrati et al. (2022); Couairon et al. (2022) use CLIP features reaching top performance. Encouraged by (Baldrati et al. 2022) we suggest a strong baseline built from pretrained BLIP components, finetuned on CoIR task.

Lastly, Kim et al. (2021) suggested enforcing cyclic consistency from query/target images back to the transition-text. To this end, they jointly optimized two separate networks, one devoted to the CoIR task; another predicting the query text, given the query and target images. The latter is used for re-ranking the target candidates. In this work, we suggest a different auxiliary task, dubbed *reverse objective*, inspired by the CoIR task. Our reverse objective maps the target image, conditioned on the query text, back to the query image.

3 LaSCo Dataset

In this section we introduce LaSCo (*Large Scale Composed Image Retrieval*), a new CoIR dataset consisting of open-domain natural images, that elevates the scale of existing datasets. To construct LaSCo, we leverage the carefully labeled datasets that exist for the well-studied VQA task (Yang et al. 2021). Specifically, we utilize the *VQA2.0* (Goyal et al. 2017) dataset to create LaSCo with minimal human effort. *VQA2.0* introduces two important features: 1) A balanced answer set for VQA; 2) Inclusion of "complementary" samples, with counter examples. A complementary image I_c is one that is similar to an original image I in VQA, but yields a different answer for the same question.

3.1 Data Roaming

We generate CoIR triplets from *VQA2.0* samples and their "complementary" counterparts, as demonstrated in Figure 2. For brevity, let us denote *VQA2.0* by set $\mathcal{D}$. Consider two complementary triplets $(I, Q, A) \in \mathcal{D}$ and $(I_c, Q, A_c) \in \mathcal{D}$. By construction, each I_c image was manually selected from the 24 (visually) nearest neighbors of I such that: 1) any premise assumed in Q must hold in I_c; 2) it makes sense to ask Q about I_c; and 3) the answer A_c to Q about I_c differs from A. The actual answer A_c was provided by a different annotator, in a second round. These properties of $\mathcal{D}$ enable building our new dataset by using the existing (Q, A_c) pair to construct the transition text S_c from I to I_c.

Transition text: The text conversion task is defined by $G : (Q, A_c) \mapsto S_c$. We employ a strong language model, *e.g.*, G :=*GPT-3* due to its few-shot capabilities (Brown et al. 2020), to perform the conversion. To create an example set for the task, we recruit 20 annotators to manually

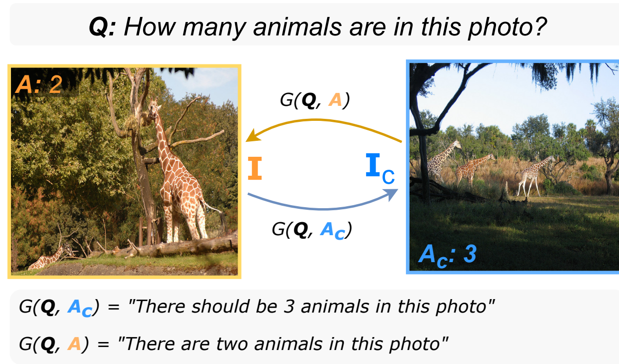

Figure 2: Generating transition texts from *VQA2.0* samples. Given two paired triplets (I, Q, A), (I_c, Q, A_c), where I and I_c are visually similar, but yield different answers A and A_c for the same question Q, a transition text from I to I_c is generated by G :=*GPT-3* (Brown et al. 2020), based on (Q, A_c).

rephrase ~ 300 randomly sampled (Q, A_c) pairs to valid transition texts S_c. We then provide G a short description of the task, with three annotated examples of (Q, A_c, S_c), and ask the model to perform the task on a new pair (Q, A_c). We further exploit the symmetry in the transition, $I \rightarrow I_c$ and $I \leftarrow I_c$ to generate more triplets. Finally, we organize the triplets as (I, S_c, I_c), (I_c, S, I), with S_c and S indicating the corresponding transition texts. For an extensive list of examples we refer the reader to our suppl. material.

3.2 Quality Control

We further conduct a data curation process and preliminary evaluation for the quality of our generated CoIR triplets. We first remove triplets identical query/target image. For text, we apply automatic rule-based scripts to filter out potentially wrong conversions (*e.g.*, too short, unexpected characters such as '\n', etc.). In our manual examination of 1000 random text conversions we judged 91.7% of them to produce well-phrased and reasonable transitions. Next, we conduct a short user study to compare the quality of our transition texts to fully human-annotated ones. We sample ~ 300 random triplets (Q_i, Q_t, T_i), of query-image, query-text, and target image (respectively). Triplets were shown to a total of 30 users, who were asked whether Q_t "adequately describes the transition/modification", from Q_i to T_i, or not. This experiment was conducted on three different CoIR datasets. The results show 82.02% positive rate for LaSCo samples, compared to 81.15% for *FashionIQ*, and 82.65% for *CIRR*.

Finally, we performed a larger scale user study using the Amazon Mechanical Turk (AMT) platform. We randomly presented 1000 samples (triplets) from each dataset, and asked 3 different AMT workers to rate each sample using a 1–5 rating scale (worst–best, respectively). A mean opinion score (MOS) was computed for each sample as the average of the three ratings. Binarization of the ratings (considering 1,2 as 'Bad', otherwise as 'Good') yields a positive (Good) rate of 90.9%, 93.8% and 97.1% for LaSCo, FashionIQ and CIRR, respectively. The overall (relative) gap between LaSCo and the other datasets is under 7%, indicating that the generated texts are on-par with human annotations. For further information, please see our suppl. material.

Dataset	↑#Query Triplets	↑#Image Corpus	#Unique Tokens	Image Domain
CIRR	36,554	19,039	6,880	Natural
FashionIQ	30,132	30,271	4,425	Fashion
LaSCo (ours)	389,305	121,479	13,488	Natural

Table 1: Comparison of LaSCo to existing Composed Image Retrieval (CoIR) datasets, CIRR and FashionIQ.

Table 1 compares statistics of LaSCo to previous CoIR datasets. LaSCo contains over 389K queries, $\times 10$ larger than previous datasets, with an image set containing 121.5K different images, compared to previous 19K–41K. The size of the test image corpus, determining the target search space, is almost 40K, compared to 2.3K in *CIRR* and 15.4K in *FashionIQ*. In terms of natural language, LaSCo is richer with 13.5K different language tokens, compared to 4.4K in *FashionIQ* and 6.8K in *CIRR*. Moreover, LaSCo and *VQA2.0* are both derived from COCO's image set; thus, captions are available for each of LaSCo's images. Utilizing captions as an additional cue (see Section 6) allows creating a rich dataset for training CoIR methods to achieve high performance in both Text-to-Image and CoIR tasks.

4 Cross-Attention Driven Shift Encoder

Here we introduce a new strong baseline for CoIR that leverages pre-trained BLIP components with early fusion, named *Cross-Attention driven Shift Encoder (CASE)*.

4.1 CASE Architecture

The CASE architecture, depicted in Figure 1, consists of two transformer components (Vaswani et al. 2017). The first is a shift-encoder, based on an image-grounded text encoder, previously introduced in (Li et al. 2022). It is a BERT (Devlin et al. 2019) encoder with additional intermediate cross-attention layers, to model vision-language interactions. The second component is a ViT (Dosovitskiy et al. 2021) encoder. ViT divides an input image into patches and encodes them as a sequence of *image tokens*. The image tokens are then fed into cross-attention layers, allowing interaction between the lingual and visual branches. The output, a bi-modality conditioned sequence (text on image and image on text), is then pooled to a single vector and projected to a 256D latent space. CASE allows *early fusion* between modalities, in contrast to previous *late fusion* methods (Baldrati et al. 2022; Vo et al. 2019; Kim et al. 2021; Delmas et al. 2022) or methods that take a middle way (Liu et al. 2021; Goenka et al. 2022; Chen, Gong, and Bazzani 2020; Hosseinzadeh and Wang 2020), as discussed in Section 2.

Utilizing Vision-Language Pre-training: We initialize our model's weights using BLIP (Li et al. 2022) pre-trained weights, as follows: Our shift-encoder's Self-Attention, Cross-Attention and Feed-Forward layers are initialized with the corresponding layers of BLIP's image-grounded encoder. Our final projection layer is initialized with the final projection of BLIP's text-encoder. Finally, we initialize our ViT component with BLIP's image encoder. Our model is end-to-end trainable (see Figure 1).

4.2 Adding a Reverse Objective

A common training approach for image retrieval tasks uses an objective of contrastive loss with the target image as positive *e.g.*, (Goenka et al. 2022; Baldrati et al. 2022; Liu et al. 2021; Li et al. 2020). Here, we propose an additional *reverse* objective, where the goal is to retrieve the query image given the transition-text and the target image. One can view the reverse objective as flipping the shift vector of the original query (in the latent space) to point in the opposite direction, from the target image to the embedding of the query image. Our reverse objective further suggests an additional task and can be viewed as a valid augmentation, effectively enlarging the dataset. We therefore train our model jointly with the reverse objective (see Figure 1). Namely, given a triplet (Q_i, Q_t, T_i) of query-image, query-text and target-image (respectively), our model objective M requires: $M(Q_i, Q_t) = T_i$ (standard CoIR task), while simultaneously enforcing $M(T_i, [REV]; Q_t) = Q_i$, where $[REV]$ is a special token provided to the model. Although the reverse task is not one-to-one (multiple queries may be suitable), this objective has proven to be beneficial in practice.

4.3 Retrieval Approach

We follow the most common approach for image retrieval: searching for matches in an embedding space shared by queries and targets (see Figure 1). First, corpus images (potential targets) are encoded into a single feature vector per image by a ViT encoder. Then, a given query (composed of image and text) is projected by CASE to the shared embedding space. Finally, the target candidates are ranked by cosine-similarity distance w.r.t the shifted query embedding. By using a relatively small embedding space dimension of 256D, compared to 640D in the previous SoTA (Baldrati et al. 2022), the retrieval is sped up by $\times 2.5$.

5 Modality Redundancy

In this section, we first propose a simple analysis of existing CoIR datasets to examine the degree to which their queries require *both* modalities for successful retrieval. Next, a similar analysis is proposed for assessing the bias of CoIR methods towards modality redundancies.

An ideal composed query should require both modalities for retrieving the desired target. For example, a transition-text such "Change the color to be more cream colored" in the top row of Figure 5, will only succeed in finding the proper target in conjuction with the query image, since the type of object cannot be inferred from the text alone. However, in practice, one of the modalities can become redundant, with the degree of redundancy depending on the information conveyed by the other modality. On one extreme, the query-image might be completely redundant, reducing the task to Text-to-Image retrieval; on the other extreme, the query-text might be redundant, with the task becoming Image-to-Image retrieval. To quantitatively assess the degree of redundancy

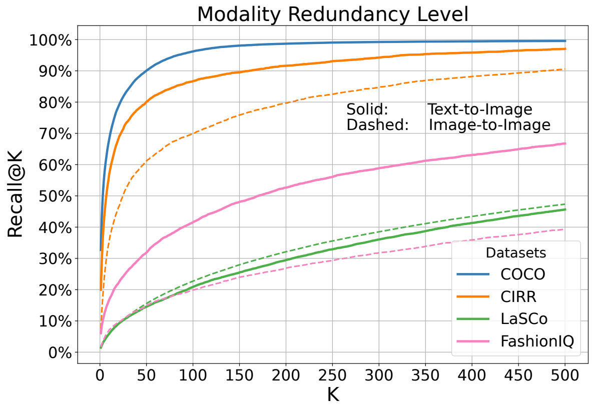

Figure 3: Modality redundancy level in several datasets (lower is better). Recall@K values for *Text-to-Image* and *Image-to-Image* retrieval using off-the-shelf CLIP (Radford et al. 2021). Lower values indicate higher image and text compositionality. The results for COCO (Lin et al. 2014) are shown as a reference for a purely text-to-image (single modality) retrieval task, presenting an upper bound.

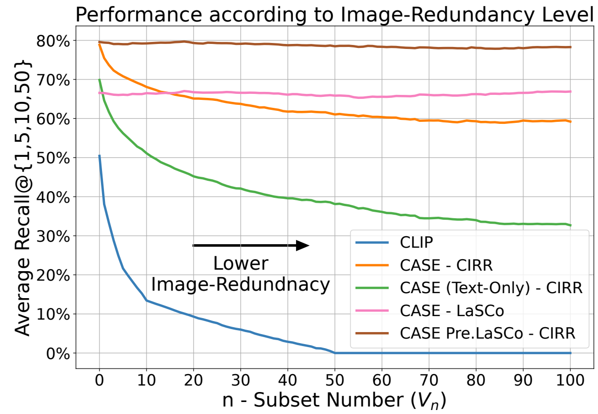

Figure 4: Average retrieval performance on subsets of *CIRR* determined by *image-redundancy* levels. Higher values of Average Recall, imply a desirable trend of higher level of required *compositionality* between text and image (lower modality redundancy). See Section 5.

in CoIR datasets, we measure the Recall@K performance of naive Text-to-Image and Image-to-Image retrieval using the embeddings produced by an independent off-the-shelf CLIP model (Radford et al. 2021). To create a continuous measure we compute the Recall@K for varying K values. These measurements computed on several datasets are plotted in Figure 3. A lower curve indicates that the corresponding dataset is more challenging for a uni-modal query. Note that the LaSCo and *FashionIQ* curves are much lower than *CIRR*, implying that more of the queries in *CIRR* are modality-redundant. For reference, we also plot the performance of the CLIP-based Text-to-Image retriever using COCO captions as query text (a commonly used benchmark for Text-to-Image retrieval (Li et al. 2022, 2021, 2020; Radford et al. 2021)). While COCO may be viewed as an "upper bound" for this task, note that the *CIRR* curve is quite close to it.

Next, we employ a similar analysis for studying the degree to which CoIR methods (trained on a certain dataset) are affected by the presence of modality redundancies in the dataset. Starting from the full *CIRR* validation set, denoted as V, we generate a sequence of progressively "purified" subsets $V_n \subset V$, with each subset containing fewer modality redundancies. Specifically, subset V_n is generated by removing from V all of the queries for which the naive CLIP-based Text-to-Image retriever, retrieves the correct target image among it's top-n results. In Figure 4 we plot the average of Recall@$\{1, 5, 10, 50\}$ as a function of n, measured by applying our baseline, CASE on each dataset.

Note that the performance of the CLIP-based retriever (blue line) vanishes at V_{50}, since by construction, V_{50} contains only queries for which CLIP failed to retrieve the target within its top 50 results. A similar trend is observed with CASE-(Text-Only), a variant of our model trained on transition-texts only, ignoring the image. CASE-(Text-Only) exhibits performance degradation with increased n, as it relies solely on query-text. While CASE that was trained on *CIRR* (orange line) shows better performance, it still suffers some degradation as n grows, implying that some bias towards modality redundancies might still exist. However, when CASE is trained or pre-trained on LaSCo dataset (pink and brown lines, respectively), it achieves the best performance, which is roughly constant regardless of n. Thus, LaSCo appears to be effective at removing bias towards redundancies, and CASE pre-trained on it is better suited for datasets with high compositionality.

6 Evaluation

We evaluate our baseline on *FashionIQ*, *CIRR*, and *LaSCo* benchmarks, one domain-specific and the others more general and broad, based on natural images. First, we show the results from our newly suggested baseline, CASE. Next we examine the effect of using our new LaSCo dataset for training and pre-training (train/val. split of 92% & 8%). We also present results with pre-training with a mixture of COCO captions, that are very descriptive to better handle samples where the transition text is highly detailed, making the the query image often redundant (*i.e.* text-to-Image retrieval). To this end, we conduct an experiment where we train CASE on LaSCo, replacing 50% of transition-texts Q_t, with captions, corresponding to the target image. Namely, we change the train distribution to combine both CoIR and text-to-image samples, as discussed in Sec. 3.1. We then explain the results thru the properties of different datasets in terms of modality redundancy.

6.1 Datasets

FashionIQ (Wu et al. 2021) contains crowdsourced descriptions of differences between images of fashion products. Images are collected from the web and divided to three categories of *Shirts*, *Dresses* and *Tops&Tees*. The query and target images were automatically paired based on title similarities (crawled from the web). This dataset consists of 30K

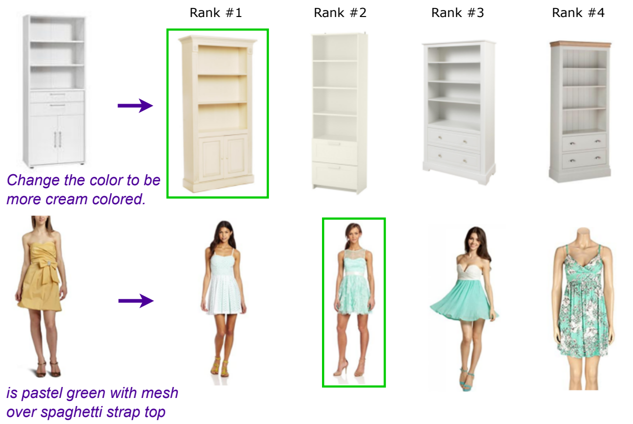

Figure 5: CASE top-4 retrievals (from left to right) for a query in *CIRR* (top) and *FashionIQ* (bottom). The query (image and text) is shown in the left column. The single ground truth target is framed in green. Arguably, additional images could be marked acceptable (referred as false-negatives).

queries (see Table 1), annotated on 40.5K different fashion images. There are 4.4K different tokens in the transition-texts (according to *BERT* tokenizer). The validation corpus contains 15.4K different images, from which target should be retrieved). Figure 5 (bottom) shows a *FashionIQ* retrieval example.

CIRR contains open domain natural images, taken from NLVR2 (Suhr et al. 2019). It contains a total of 36.5K queries annotated on 19K different images, with 6.8K unique tokens in the transition texts. Examples may be seen in the top two rows of Fig. 5. Its validation corpus is relatively small, with a size of 2.3K. The authors further suggest two benchmarks, one *general*, with the target search space as the entire validation corpus, and a *subset*, where the search space is a subgroup of six images similar to the query image (based on pre-trained ResNet15 feature distance), demonstrating a fine-grained retrieval task.

6.2 Implementation Details

We set an *AdamW* optimizer, initializing learning rate by 5×10^{-5} with a exponential decay rate of 0.93 to 1×10^{-6}. We train CASE on *CIRR* with a batch size of 2048 for 6 epochs. For *FashionIQ*, we train with a batch size of 1024 for 20 epochs (further ablation on the batch size is available in suppl. material). On LaSCo we train with a batch size of 3840 for 10 epochs. We use the Recall@K surrogate loss (Patel, Tolias, and Matas 2022) as the differentiable version of the Recall@K metric. Training on four A100 nodes takes 0.5-6 minutes per epoch, depending on dataset size.

6.3 Results

We start by showing the performance of our CASE baseline on *FashionIQ*, in Table 2. The results are broken down to different clothing categories and Recall@K values. For demonstrating modality redundancies, we train two baselines only on query-image (Image-only) or query-text (Text-only). Interestingly, CASE achieves SoTA results, surpassing the previous top performing method (LF-CLIP (Baldrati et al. 2022)) by a large margin (13.4% and 11.6% absolute points at Recall@10,50 respectively). The poor results for CASE (Image-only) baseline, show that visual information is not sufficient for *FashionIQ*, as often the transition text asks for a certain change in the image (see Section 5). However, the CASE (Text-only) baseline results are close to the previous method of LF-CLIP, indicating the high level of redundancy in the image, as the single text modality is sufficient to reach the previous SoTA performance.

Table 3 shows results on *CIRR*. At the top we present the results from previous methods. The right columns $R_{subset}@K$, correspond to the fine-grained retrieval task (introduced originally in (Liu et al. 2021)). Here, we show five different variants of our model. As in *FashionIQ*, the poor results for CASE (Image-Only) imply that the query-image alone is not sufficient for retrieval also on *CIRR*. Interestingly, CASE (Text-only) reference surpass previous methods in most metrics, further demonstrating the high level of modality redundancy in *CIRR* as shown in Sec. 5. This baseline is also top performing in the *subset* benchmark. We believe this is caused by the existing image similarity in the subset, making the query image redundant for the task (see visual examples in suppl. material).

Next, we observe the performance of CASE, which consistently outperforms previous methods. We further show visual and textual explainability maps in the suppl. material.

Now we examine the impact of our LaSCo dataset in two main aspects 1) Pr-training and 2) Zero-Shot inference. The results are shown under CASE Pre-LaSCo, gaining ~0.5-1% improvement. We found that this relatively small improvement is due to the modality redundancy in the *CIRR* dataset. We justify this assumption by the analysis shown in Fig. 4, and by shifting CASE towards the distribution of *CIRR*. Specifically, we train CASE on a mix of LaSCo transition-texts with full target captions (taken from COCO captions), which we denote LaSCo.Ca, thus biasing the model towards highly descriptive texts. As shown in Tab. 3, it further boosts performance on *CIRR*, a fact that we attribute to improvement of the Text-to-Image (T2I) search capability (see Section 5). The use of LaSCo.Ca significantly boosts performance also in *zero-shot* on *CIRR* test set (*i.e.* without even training on *CIRR*), surpassing previous methods in most metrics, indicating again the impact of modality redundancy on the results.

Finally, in Table 4, we benchmark on LaSCo. To this end, we apply the two CASE variants (Text-Only, Image-Only) that result in poor performance, implying the necessity of both modalities in this dataset. We further test LF-CLIP (Baldrati et al. 2022) trained on LaSCo, and observe its significant drop in performance (compared to *CIRR*), implying that LaSCo dataset introduces a higher true CoIR challenge. Finally, CASE performs best also here, raising *e.g.*, Recall@1 from 4.01% (by prior LF-CLIP) to 7.08%, and Recall@50 from 32.08% to 50.25%.

Ablation Study: First, we construct a reference by replacing pre-trained encoders of CLIP, with BLIP's, presented in Tables 2 to 4, named LF-BLIP. Next, we ablate

Method	Shirt R@10	Shirt R@50	Dress R@10	Dress R@50	Toptee R@10	Toptee R@50	Average R@10	Average R@50
MAAF (Dodds et al. 2020)	21.3	44.2	23.8	48.6	27.9	53.6	24.3	48.8
FashionVLP(Goenka et al. 2022)	31.89	<u>58.44</u>	32.42	<u>60.29</u>	<u>38.51</u>	<u>68.79</u>	34.27	<u>62.51</u>
LF-BLIP	25.39	43.57	25.31	44.05	26.54	44.48	25.75	43.98
LF-CLIP (Baldrati et al. 2022)	<u>36.36</u>	58.00	31.63	56.67	38.19	62.42	<u>35.39</u>	59.03
CASE	**48.48**	**70.23**	**47.44**	**69.36**	**50.18**	**72.24**	**48.79**	**70.68**
CASE (Image-only)	6.72	17.52	8.02	18.6	7.93	18.05	7.43	17.95
CASE (Text-only)	34.84	56.93	<u>33.47</u>	55.81	35.99	58.52	34.89	57.18

Table 2: Recall@K comparison, on *FashionIQ* validation set. Best performance is in bold and second best is underlined. CASE outperforms previous results in all metrics by a large margin. CASE(Text-only) results indicate high modality redundancy.

Mode	Method	Recall@K K=1	K=5	K=10	K=50	R_{subset}@K K=1	K=2	K=3
Train	TIRG (Vo et al. 2019)	14.61	48.37	64.08	90.03	22.67	44.97	65.14
	MAAF (Dodds et al. 2020)	10.31	33.03	48.30	80.06	21.05	41.81	61.60
	ARTEMIS (Delmas et al. 2022)	19.96	46.10	61.31	87.73	39.99	62.20	75.67
	LF-BLIP	20.89	48.07	61.16	83.71	50.22	73.16	86.82
	LF-CLIP (Baldrati et al. 2022)	33.59	65.35	77.35	95.21	62.39	81.81	92.02
	CASE	**48.00**	**79.11**	**87.25**	**97.57**	**75.88**	**90.58**	**96.00**
Zero Shot	CASE Init.	16.63	33.54	42.65	65.30	55.74	77.10	88.48
	CASE - LaSCo	30.89	60.75	73.88	92.84	60.17	80.17	90.41
	CASE - LaSCo.Ca.	**35.40**	**65.78**	**78.53**	**94.63**	**64.29**	**82.66**	**91.61**
Pre-Train	CASE Pre-LaSCo	48.68	79.98	88.51	**97.49**	76.39	90.12	**95.86**
	CASE Pre-LaSCo.Ca.	**49.35**	**80.02**	**88.75**	97.47	**76.48**	**90.37**	95.71
Modality Redundancy	CASE (Image-only)	0.00	0.19	0.41	2.12	19.78	39.49	59.87
	CASE (Text-only)	39.01	69.53	79.24	91.32	78.68	91.70	96.08

Table 3: Recall@K comparison on *CIRR* test set. CASE shows state-of-the-art results in all cases. The methods in the "Train" mode were all trained on the *CIRR* train set, in contrast to "Zero-shot" mode. CASE (Text/Image-only) was trained solely on a single modality (text/image, respectively), while ignoring the other.

Method	R@1	R@5	R@10	R@50	R@500
CASE (Img.-Only)	2.21	7.39	11.82	30.62	72.64
CASE (Txt.-Only)	2.39	6.89	10.39	24.92	61.23
LF-CLIP	4.01	10.23	14.68	32.08	72.69
LF-BLIP	4.26	12.01	17.11	36.54	74.62
CASE	**7.08**	**18.50**	**26.16**	**50.25**	**85.46**

Table 4: Results on LaSCo validation set.

various key-components of CASE to examine their impact on performance. We report ablation results on the frequently used *FashionIQ* dataset (also on CIRR) in the suppl. material). We train CASE without the reverse queries (RQ) objective described in Sec. 4.2. We observe that RQ improve performance by 0.5–1.2% ($\sim 5\%$ relative performance boost at R@1). Using surrogate Recall@K loss instead of common contrastive loss, further improves results by roughly 0.5%. Finally, we examine the influence of fine-tuning the ViT where results vary depending on the dataset. On LaSCo, all R@K metrics were improved by an absolute $0.4 - 2\%$. On *CIRR* and *FashionIQ*, which contain fewer images, some metrics improved but with trade-offs in others.

7 Discussion

We shed more light on the task of CoIR. Data labeling for CoIR appears to be difficult and costly, exposed to serious biases (*e.g.*, redundancy of the image-query), bounded by inevitable flaws (*e.g.*, false-negatives), and eventually ending up with a low quality and size of data. We suggest a remedy for most of these shortcomings via an inexpensive solution: leveraging labels from a popular related task, to create a new labeled dataset, LaSCo. We extensively analyze current CoIR datasets, in order to show their effectiveness, generalization and the capability of a certain model (trained on specific dataset) in handling the desired compositionality in the CoIR task. We also suggest the CASE baseline, that relies on early fusion of the query modalities through a cross-attention module. We demonstrate the effectiveness of CASE by achieving top results on labelled CoIR benchmarks from two different domains. To the best of our knowledge, our new baseline, also leverages the smallest dimension of the search space (shared embedding space) among the methods being compared, resulting in a further reduction in computational expenses. We believe this work, including our newly introduced dataset, might serve as a useful and practical tool not solely limited to the intricate CoIR task but also extending to the broader realm of multi-modal learning.

Acknowledgments

This work was supported in part by the Israel Science Foundation (grants 2492/20 and 3611/21). We thank Or Kedar for his help with parts of this research.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, 8948–8957.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Bimbo, A. D. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In *CVPR*, 21434–21442. IEEE.

Barz, B.; and Denzler, J. 2021. Content-based image retrieval and the semantic gap in the deep learning era. In *ICPR*, 245–260.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*, volume 33, 1877–1901.

Chen, D. L.; and Dolan, W. B. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *ACL*, 190–200.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *CVPR*, 2998–3008.

Couairon, G.; Douze, M.; Cord, M.; and Schwenk, H. 2022. Embedding Arithmetic of Multimodal Queries for Image Retrieval. In *CVPRW*, 4946–4954. IEEE.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *CVPR*.

Delmas, G.; de Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Dodds, E.; Culpepper, J.; Herdade, S.; Zhang, Y.; and Boakye, K. 2020. Modality-Agnostic Attention Fusion for visual search with text feedback. *CoRR*, abs/2007.00145.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Dubey, S. R. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2687–2704.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback. In *CVPR*, 14085–14095. IEEE.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. S. 2018. Dialog-based Interactive Image Retrieval. In *NeurIPS*, 676–686.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *ICCV*, 1472–1480.

Hosseinzadeh, M.; and Wang, Y. 2020. Composed Query Image Retrieval Using Locally Bounded Features. In *CVPR*, 3593–3602. Computer Vision Foundation / IEEE.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *CVPR*, 1383–1391.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual Compositional Learning in Interactive Image Retrieval. *AAAI*, 35(2): 1771–1779.

Lee, S.; Kim, D.; and Han, B. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *CVPR*, 802–812.

Levy, M.; Ben-Ari, R.; and Lischinski, D. 2022. Classification-Regression for Chart comprehension. In *ECCV*, 469–484.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 12888–12900.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*, 9694–9705.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, 740–755.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *ICCV*, 2105–2114.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 13–23.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *CVPR*, 2630–2640.

Nagrani, A.; Seo, P. H.; Seybold, B.; Hauth, A.; Manen, S.; Sun, C.; and Schmid, C. 2022. Learning Audio-Video Modalities from Image Captions. *ECCV*.

Patel, Y.; Tolias, G.; and Matas, J. 2022. Recall@k surrogate loss with large batches and similarity mixup. In *CVPR*, 7502–7511.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *ICML*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *ACL*, 6418–6428.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*, 6432–6441.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 11307–11317.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 5288–5296. IEEE Computer Society.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 1686–1697.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2: 67–78.

Zhong, Q.; Chen, L.; and Qian, Y. 2020. Few-shot learning for remote sensing image retrieval with maml. In *ICIP*, 2446–2450. IEEE.