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

# Supplementary Material

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

## A. Balancing the Forward and Reversed Loss Terms

We investigate the effect of varying the hyperparameter $\alpha$ in the bi-directional loss (Equation 3). As a general rule of thumb, we discover that large $\alpha$ values close to, or beyond 1.0 adversely harm the performance, which corroborates with our hypothesis on the effect of false negatives in the reversed direction in Section 3.3. We, therefore, seek to balance the forward and reversed loss terms by reducing $\alpha$. We also note that the second-stage combiner training is more sensitive to tunings in $\alpha$ compared to the first stage. We suspect the reason to be related to the model capacity, as the first-stage finetuning is relatively light in architecture, while the second-stage combiner module is of much higher complexity (Figure 2 right). To this end, the combiner could more easily, and quickly, overfit to the noise brought by the false negatives.

Our choices of $\alpha$ for each training stage on both datasets for results reported in Tables 1 and 2 are detailed as follows. For Fashion-IQ [2], in both stages, we discover that an $\alpha$ of around 0.5 is optimal. We note that for the first stage, further decreasing it to 0.4 yields a slightly better result. On CIRR [1], we find that the training consistently benefits from a relatively small $\alpha$, we set it to 0.1 in both stages.

In Figure S1 we illustrate the effect of varying $\alpha$ on performance in the second-stage combiner training. We notice that as long as $\alpha$ sits within a certain range that is smaller than 1.0, the results are fairly robust.

## B. Inference on Reversed Queries

Section 3.3 details the impact of false-negatives. In Table S1 we demonstrate that validating on the reversed queries yields subpar results, which collaborates with our observation of a higher loss in the reversed path. This leads to our inference strategy that only takes into account the forward queries.

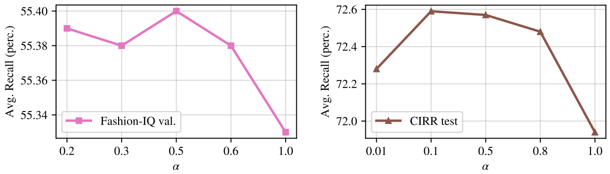


Figure S1. Performance *vs.* $\alpha$ in the second stage bi-directional training. **(Left)** Fashion-IQ validation set. **(Right)** CIRR test set. We select a few $\alpha$ values to examine the trend surrounding optimality. Note the relatively small scale in performance (y-axis), suggesting the performance is fairly robust against varying $\alpha$ within a certain range. Compare the results with Tables 1 and 2.

## C. Analysis on the Learned Reversed Semantics

We perform both quantitative and qualitative analyses to examine if our bi-directional training is encouraging the learning of the reversed semantics. Specifically, Table S2 (rows 2 *vs.* 1) compares the retrieval performance on the Fashion-IQ reversed queries with or without bi-directional training. We examine the model after the first-stage text encoder finetuning, as in Figure 2 (left). The result suggests that a model specifically trained with bi-directional queries is better equipped at reasoning over reversed semantics, which substantiates our claim. However, note that the performance on said queries is generally much lower than on the (standard) forward ones due to the larger number of potential false negatives, which has been discussed in Section 3.3.

We additionally present four qualitative examples of CIRR retrieved on the reversed queries. In Figure S2 (a)

| | | Fashion-IQ | | | CIRR | | | |
|---|---|---|---|---|---|---|---|---|
| | **BLIP4CIR+Bi** | R@10 | R@50 | Average | R@1 | R@5 | $\text{R}_{\text{Subset}}$@1 | Average |
| **1** | on forward queries | 43.49 | 67.31 | 55.40 | 42.36 | 75.46 | 72.90 | 74.18 |
| **2** | on reversed queries | 23.08 | 45.05 | 34.07 | 18.08 | 49.25 | 44.51 | 46.88 |

Table S1. Comparison of performance when validating on the forward and reversed queries. Results obtained on validation sets after the second-stage combiner training, directly comparable to results in Table 3.

| | | **Dress** | | **Shirt** | | **Toptee** | | **Average** | | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Methods** | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | **Metric** |
| **1** | BLIP4CIR (first-stage) | 4.81 | 15.42 | 8.10 | 16.63 | 7.75 | 17.64 | 6.89 | 16.57 | 11.72 |
| **2** | BLIP4CIR+Bi (first-stage) | 22.91 | 45.96 | 23.80 | 41.22 | 27.03 | 45.44 | 24.58 | 44.20 | 34.39 |

Table S2. Performance comparison on the *reversed query* retrieval with or without bi-directional training, Fashion-IQ validation set. We report the average Recall@10 and 50 of all three categories. Note that the comparison is on the first-stage text encoder finetuning (Figure 2 left).

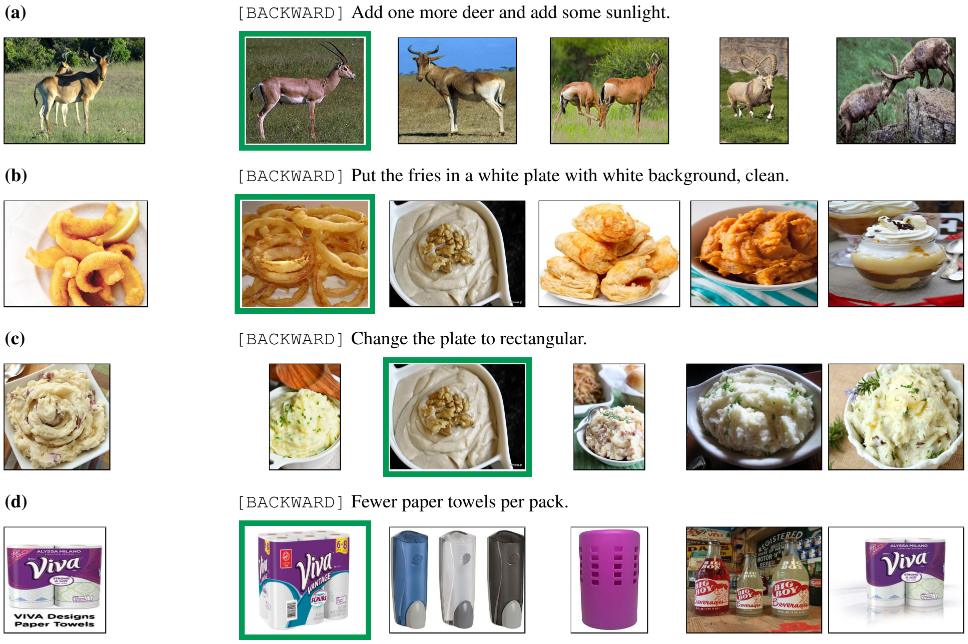
and (d) where the reversed text is unambiguous (i.e., “add” is negated to “remove”, “fewer” is negated to “more”), we show the model is capable of reasoning over such reversed semantics. We demonstrate a more complicated case in (b), where one might not definitively predict the ground truth content by examining the query. Still, among the top-5 ranked candidates, we argue that the model produces a plausible result, with the ground truth ranked the highest.

We especially illustrate the existence of false negatives among candidates in Figure S2 (c) — though the issue is present in multiple examples. Here, in particular, “change to rectangular” shall be reversed to “change *from* rectangular”, which points to a range of possible shapes. Indeed, the top-5 ranked candidates all contain non-rectangular plates — though only one of them is labelled positive. Here, we note that not all such reversed examples with false negatives can be successfully retrieved. Evidence can be seen when comparing the performance on the reversed queries (Table S2 row 2) to the performance on the forward ones (Table 1 row 19), where the former is much lower than the latter. This further validates our decisions to not perform inference on the reversed queries (Section 3.3) and to down-scale the reversed contrastive loss (Section A).

## References

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

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



(a) [BACKWARD] Add one more deer and add some sunlight.

(b) [BACKWARD] Put the fries in a white plate with white background, clean.

(c) [BACKWARD] Change the plate to rectangular.

(d) [BACKWARD] Fewer paper towels per pack.

Figure S2. Qualitative examples of *reversed query* retrieval on the first-stage text encoder finetuning (CIRR). In each example, leftmost is target image, green box denotes the ground truth (reference image), the reversed modification text is provided above the images. We show the top-5 candidates in ranking. Note that the reference image and target image exchange roles here and that the modification text shall be interpreted in its reversed semantic — for this, we specifically show the prepended text token.