Appendix for SAC: Semantic Attention Composition for Text-Conditioned Image Retrieval

Our supplementary material is organized as follows: In Appendix A, we provide implementation details for training the SAC models. We provide qualitative results on this task in Appendix C with a detailed analysis of the results obtained.

A. Implementation Details

We conduct our experiments on a machine with $4 \times T E S L A ~ V - 100$ GPU’s, $16 \times C P U$ cores with 2 threads per-core and 16 GB RAM installed with Ubuntu 16.04 LTS operating system. We used the Pytorch==1.14.0 framework for our experiments. The architecture of the Discriminator $D$ used in SAC is provided below

```
nn.Sequential(nn.Linear(512,512),
nn.LeakyReLU(0.2, inplace=True),
nn.Linear(512,256),
nn.LeakyReLU(0.2, inplace=True),
nn.Linear(256,1),
nn.Sigmoid())
```

B. Attention maps

We provide additional attention maps for different images in Figure 5.

C. Additional Qualitative Results

Figure 7 presents additional qualitative results for SAC on each of the three datasets.

In Figure 6, we present comparative visual results for SAC with VAL, the strongest baseline, on the FashionIQ and Shoes datasets. In particular, we notice that SAC is able to i) retrieve images while changing cross-granular attributes conditioned on text feedback (from Figure 6 (A), SAC retrieves shirt which is “blue colored” with “less buttons” while VAL incorrectly retrieves “orange” colored clothes) ii) concurrently focus on global appearance and multiple local fine-grained designs (from Figure 6 (B), SAC retrieves “shoes” that have “silver color” and “ice blue accents” while VAL incorrectly retrieves sandals).
Figure 7: Additional Qualitative results for our proposed method SAC on FashionIQ, Shoes, and Birds-to-Words datasets. The first column denotes the source image followed by the support text description. The retrieved results then follow with ranking in decreasing relevance from left-to-right.

Figure 8: On the left, we provide the query image and the corresponding textual query along with the ground truth from the FashionIQ dataset. On the right, we provide the retrieved results ranked by decreasing relevance results from our approach SAC. Though the retrieved results don’t contain the actual annotated sample, the results are extremely relevant to the query.