1. Qualitative Results

In this supplementary materials, we present more examples of both text-to-image and image-to-text on the Fashion-Gen dataset, as shown Figure 1 and Figure 2. As we encode tags separately, the text encoder can also attend on more detailed features of the products. For text-to-image task, in Figure 1 (a), the top-1 retrieved images from the baseline ignores the information “striping in white and blue”, and in Figure 1 (b), the baseline ignores the detailed feature “signet ring”. Similarly, for image-to-text task, in Figure 2 (a), the retrieved texts from the baseline ignores the “white printed logo” information from the image.
Figure 1. Examples of text-to-image results. For each example, the first row is the query text, the second row is the top-5 retrieved results of fine-tuning CLIP (Experiment (c) in Sec. 5.2, and the third row is the top-5 results from EI-CLIP. The correct answers are boxed in red.


---


---


---


---


---


---

Figure 2. Examples of image-to-text results. For each example, the left image is the query image. The 5 text descriptions above are results from the fine-tuning CLIP. The 5 text descriptions below are results from our EI-CLIP. The ground truth text descriptions are in blue color.