Leveraging Off-the-shelf Diffusion Model for Multi-attribute Fashion Image Manipulation –Supplementary Materials–

1. Implementation Details

We finetune the attention pooler added on top of a pretrained ViT model with a single A100 GPU, with a batch size of 64. We implement the attention pooler as a multi-head cross-attention layer with 8 heads. AdamW optimizer is used with learning rate of 1e-4 and weight decay of 0.1. No sophisticated learning rate scheduling or warm up are used. We train for 60 epochs, where we generally reach the optimum at around 50-th epoch.

For the image editing pipeline, the total guidance consists of the attribute guidance term and the background preservation term. The former is multiplied by 100 (from our hyperparameter search). For the attention map used to compute the background preservation loss, we interpolate the patch-level attention map to the image scale using bicubic interpolation.

2. Additional Synthesis Results

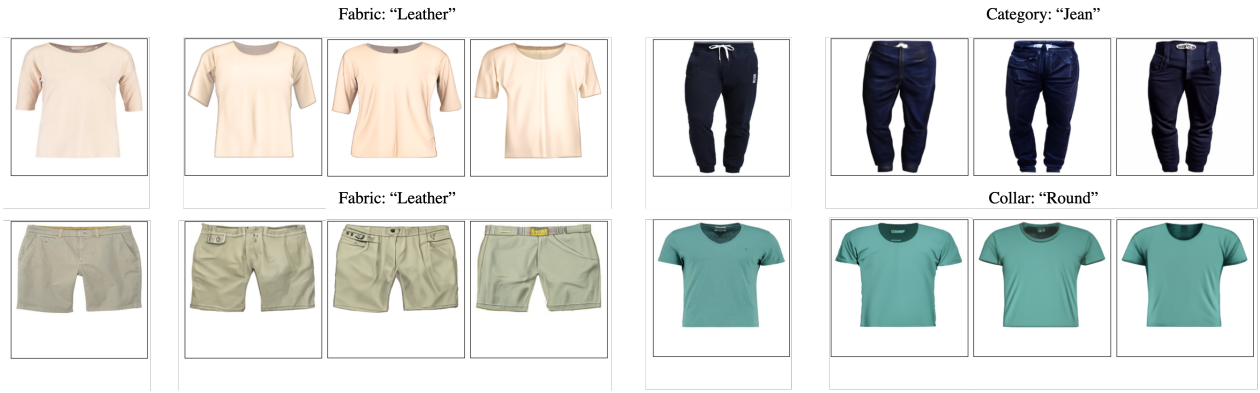

Figure 1. Additional qualitative results for fashion attribute editing.

We present additional qualitative results for our framework. We observe from the left column that the fabric manipulation works well for both the top and the bottom. In the right top, we convert sweat pants into jeans. In right bottom, we edit the collar of the given t-shirts, from v-neck to round neck.

3. Discussions

We find the class imbalance problem to hinder manipulations for some under-represented attributes. For example, in the Shopping100k dataset we have used to finetune the classifier, 51.9% of the images were labeled *regular* fit, rendering fit manipulation very challenging. We observe that overall, manipulation towards a well represented attribute works well, but the classifier suffers to provide useful guidances otherwise. Training a classifier against severe class imbalance is an ongoing research topic, and we defer this challenge to the future works.