# Supplementary: MEDAL: multi-modal MEta-space Distillation and ALignment for Visual Compatibility Learning

[1]Dween Rabius Sanny, [2]Vinay Kumar Verma, [1]Prateek Sircar, [1]Deepak Gupta
International Machine Learning, Amazon India[1], Private Brands & Discovery[2]
{drsanny, sircarp, dgupt, vkvermaa} @amazon.com

## 1. Qualitative retrieval results

In this section, we present additional qualitative retrieval results. Fig- **??** displays complementary item sets retrieved for given query items. Comparing our model's results with those of baseline models demonstrates that our approach generates more compatible item sets with higher confidence scores. The qualitative results highlight our model's superior performance in identifying complementary items. In Fig-**??** and Fig-**??** we show more qualitative retrieval results.


Figure 1. Visual compatibility set retrieval wrt query product. Score shows the similarity with ground truth product

## 2. Additional Implementation details

Our proposed model is fairly simple, utilizing only two pairwise losses: a self-supervised loss and a triplet loss. We freeze the parameters of the CLIP and freeze the gradient of the teacher models during training. Only the student's model parameters are trained, while the teacher's parameters are updated using an exponential moving average. It is also reason that adding triplets only to the student is sufficient to capture the desired features. The self-supervised loss is computed using the cross-entropy loss between the teacher's and student's predictions.

Before calculating the loss, we project the outputs of both the teacher and student models to a high-dimensional space with a projection dimension of 64K. In our method, the final dimension used for retrieval is 128, which includes both ViT and color embeddings. We do not use the student's head for any inference retrieval task; instead, only the teacher latent space embeddings are utilized as image embeddings.During the calculation of the self-supervised loss, we exclude color information, which is used solely in the triplet loss. We use $224 \times 224$ and $96 \times 96$ for the global and local crop image sizes respectively.

## 3. Some Key points

In our method, the inclusion of text embeddings is solely intended to harness textual information to aid in the retrieval task. We believe that our method does not align text and visual information for retrieving items based on a query prompt such as "I want party wear shoes that match this pant." Our method does not incorporate any outfit-level loss; instead, it focuses on pairwise losses, making it more suitable for e-commerce applications. In e-commerce, customers are typically more inclined towards purchasing a single item and expect to see complementary pairs (e.g., shirts with pants, pants with shoes, dresses with handbags) rather than buying an entire outfit.

Figure 2. Retrieval result of our method on Polyvore disjoint set. **Left:** is the query outfit with item in black box in the right being ground truth. **Right:** top-10 retrieval results for the ground truth category are shown.


Figure 3. Retrieval result of our method on Polyvore nondisjoint set. **Left:** is the query outfit with item in black box in the right being ground truth. **Right:** top-10 retrieval results for the ground truth category are shown.