FashionSAP: Symbols and Attributes Prompt for Fine-grained Fashion Vision-Language Pre-training

Yunpeng Han\textsuperscript{1}, Lisai Zhang\textsuperscript{1}, Qingcai Chen\textsuperscript{1,2,} Zhijian Chen\textsuperscript{3}, Zhonghua Li\textsuperscript{3}, Jianxin Yang\textsuperscript{3}, Zhao Cao\textsuperscript{3}
\textsuperscript{1}Harbin Institute of Technology Shenzhen, China
\textsuperscript{2}PengCheng Laboratory, Shenzhen China, \textsuperscript{3}Huawei Technologies Co., Ltd

Abstract

Fashion vision-language pre-training models have shown efficacy for a wide range of downstream tasks. However, general vision-language pre-training models pay less attention to fine-grained domain features, while these features are important in distinguishing the specific domain tasks from general tasks. We propose a method for fine-grained fashion vision-language pre-training based on fashion Symbols and Attributes Prompt (FashionSAP) to model fine-grained multi-modalities fashion attributes and characteristics. Firstly, we propose the fashion symbols, a novel abstract fashion concept layer, to represent different fashion items and to generalize various kinds of fine-grained fashion features, making modelling fine-grained attributes more effective. Secondly, the attributes prompt method is proposed to make the model learn specific attributes of fashion items explicitly. We design proper prompt templates according to the format of fashion data. Comprehensive experiments are conducted on two public fashion benchmarks, i.e., FashionGen and FashionIQ, and FashionSAP gets SOTA performances for four popular fashion tasks. The ablation study also shows the proposed abstract fashion symbols, and the attribute prompt method enables the model to acquire fine-grained semantics in the fashion domain effectively. The obvious performance gains from FashionSAP provide a new baseline for future fashion task research.

1. Introduction

Vision-Language pre-training (VLP) attracts wide attention \cite{B8, B9, B10, B11} as a foundation for general multi-modal tasks. VLP methods aim at learning multimodal knowledge from large-scale text and image pairs data containing common objects in daily life. For example, MSCOCO \cite{B12}, a public vision-language benchmark, is introduced with common object labels. The fashion domain is an important application of VLP methods, where the online retail market needs retrieval and recommendation services. To satisfy such requirements, the VLP model needs to learn high-quality representations containing fine-grained attributes from the fashion data.

Many works have adapted general VLP models to fashion tasks directly. However, the general pre-training models are not effective for learning fashion knowledge to describe fashion items comprehensively, as the fashion descriptions are usually associated with fine-grained attribute-level features. As illustrated in Fig. 1, the description text of a fashion item (right) refers to fine-grained attributes like long sleeves, while such features are ignored by the de-
<table>
<thead>
<tr>
<th>Fashion Symbols</th>
<th>Categories</th>
<th>Definition Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOPS</strong></td>
<td>tops, shirt, polo, sweater, ...</td>
<td>upper body</td>
</tr>
<tr>
<td><strong>DRESSES</strong></td>
<td>dress, suit, shift, ...</td>
<td>up-to-lower body</td>
</tr>
<tr>
<td><strong>SKIRTS</strong></td>
<td>skirt, sarong, slit, kilt, ...</td>
<td>lower body</td>
</tr>
<tr>
<td><strong>COATS</strong></td>
<td>jacket, parka, blazer, duffle, ...</td>
<td>associated with others</td>
</tr>
<tr>
<td><strong>PANTS</strong></td>
<td>jeans, shorts, breeches, ...</td>
<td>lower body</td>
</tr>
<tr>
<td><strong>SHOES</strong></td>
<td>boots, sneakers, pump, loafers, ...</td>
<td>feet</td>
</tr>
<tr>
<td><strong>BAGS</strong></td>
<td>clutches, pouches, wristlet, ...</td>
<td>bag &amp; decorative</td>
</tr>
<tr>
<td><strong>ACCESSORIES</strong></td>
<td>ring, sunglasses, accessories,</td>
<td>decorative &amp; optional</td>
</tr>
<tr>
<td><strong>OTHERS</strong></td>
<td>hat, necklace, ...</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Fashion symbols and corresponding categories with definition rules.

scriptions from general vision-language data (left). Moreover, the public fashion producers from fashion platforms attach great importance to some definite attributes (e.g., season, gender) of fashion data and especially provide the fashion attributes annotations. However, these high-quality attributes are highly neglected by existing fashion VLP models. It is important for fashion VLP models to focus on these fine-grained attributes and learn fashion-specific knowledge.

Fashion attributes describe not only item details but also the overall item features. The category for fashion items is an essential attribute highlighted by many benchmark datasets [31, 42, 50]. We notice that categories have a deep correlation to fine-grained attributes, although they describe the general information of a fashion item. For example, the length is an important attribute for both pants and jeans, while it is rarely mentioned in the description of a pair of shoes. However, most existing fashion VLP methods neglect the importance of the relationship between similar categories. In this paper, we explore the usage of category attributes as a global concept layer during pre-training. According to the human description of a fashion product, we believe categories declare the basic understanding of a fashion product. Therefore, we attach the fashion category to the beginning of captions to guide the representation learning. Since the fashion products are designed for the decoration of people, we summarize nine fashion symbols corresponding to human body parts, as shown in Tab. 1, to unify all the categories of fashion items.

We propose a method for the fashion domain to learn fine-grained semantics. This method is able to capture the similarity of fine-grained features based on fashion symbols and learn explicit fine-grained fashion attributes by the prompt. Our method gets the SOTA performance for four popular fashion tasks on the two public datasets, and the obvious performance gains provide new baselines for further research. Our main contributions are summarized below:

- An effective fine-grained vision-language pre-training model is proposed to learn the attribute-level fashion knowledge.
- An abstract fashion concept layer is proposed, and 9 fashion symbols are summarized to represent various fashion concepts according to their similarities on body parts and product functions.
- The attributes prompt method enables the cross-modalities pre-training model to explicitly learn fine-grained fashion characteristics.

2. Related Work

**Vision-Language Pre-training** The pre-training of the vision-language model has been used in many works [10, 16–18, 48]. The structure of the VLP model mainly includes two types, single-stream and two-stream. The single-stream models [18, 48] generate the image and text into preliminary representations and concatenate them so that they can interact with each other in a unified model (e.g. transformer [38]). Two-stream models [11, 30, 41] try to encode text and image respectively and the features interact with each other through semantic alignment tasks. Some works [16, 17, 43, 47] combine single-stream and two-stream by designing multi-step semantic alignment tasks. The backbones for text and image encoder refer to the stricture of unimodal [2, 3, 9]. The one-stream models usually perform better than two-stream models, while the latter is better than the former in time complexity. We design a model combined with one-stream and two-stream to adapt to the fashion tasks.

**Vision-Language Model for Fashion** Tasks in the fashion domain include the retrieval, match and generation of cross-modal [6] similar to the general vision-language. There are also many datasets collected and released for fashion tasks [6, 21, 31, 37, 40, 42]. KaleidoBERT [4] designs multiple stages to refine the salient features of fashion
items by utilizing multiple single-task frameworks. FashionViL [7] uses an end-to-end framework to pre-train the model in multiple single-tasks by referring to the general vision-language model. These works try to use attributes of fashion items by attaching all the category attributes to the same classification task. There are also some works aiming at specific fashion tasks [1, 5, 13, 26, 39, 45] by setting a variety of gating and route structures for the latent features of fashion items. The exact representation of each attribute respectively is essential for fashion models. We propose a model that can obtain latent features and knowledge in the fashion domain at the pre-training stage.

**Prompt Learning** Prompt learning is an effective method to transfer the pre-training model to accomplish downstream tasks in Natural Language Processing [15, 20, 32, 33, 35]. It can utilize the knowledge from the large-scale pre-training model in low-resource scenarios by appropriate prompts. The expression of the prompting is the crucial aspect in task transfer [14, 23] as proper trigger words can activate the knowledge from the pre-training model better. Prompted-base methods are also used in task-aimed model training [15, 35] for utilizing the resource in the task. These methods diversify a single instance from multiple perspectives into multiple instances. There are also some works applying prompt learning to multi-modal. We design two prompt templates from the text side for the adaptation of different kinds of attributes. [36, 44].

3. Methodology

In this section, we first introduce the preliminary of fashion symbols and attributes prompt in Sec. 3.1. Then we describe the architecture of FashionSAP in Sec. 3.2 network. Afterward, we elaborate on five pre-training tasks in Sec. 3.3.

3.1. Preliminary

3.1.1 Fashion Symbols Definition

The category is an essential attribute of a fashion item. However, the categories terms in different datasets are various. For example, the widely used FashionIQ [40] provides 3 kinds of categories while 48 in FashionGen [31]. To address the problem, we propose a concept semantic layer to embed similar category terms into the same fashion symbol. The symbols are defined by the following rules:

1. **Body Part**: fashion items that are associated with a specific part of the human body.

2. **Function**: fashion items that are optionally used for decoration and can be dressed on multiple body parts.

For the datasets in this paper, we propose nine symbols to summarize different categories of fashion items. As shown in Tab. 1, the fashion symbols PANTS, SKIRTS, SHOES, BAGS have their unique features. TOPS is a kind of upper clothing that can be worn independently. DRESSES can cover the whole body and exist independently. COATS represents the outwear usually worn with other clothing. ACCESSORIES represents the accessories that aim to enhance the whole outfit but are not necessary for a basic outfit. OTHERS includes fashion items that do not appear in everyday dressing and public occasions. We use an embedding layer to learn the representation of these fashion symbols as shown in Fig. 2. We enumerate all categories and corresponding fashion symbols in practice.

3.1.2 Fine-grained Attribute Prompt

Existing works suggest that the fashion items are usually annotated from multiple perspectives [50]. Most benchmark datasets [21, 31, 40, 42] focus on fine-grained attributes when annotating fashion items. However, most general vision-language models focus on object-level semantics and seldom pay attention to attribute-level semantics, which contain many fine-grained characteristics for fashion items. Therefore, we propose a method to utilize these fine-grained attributes by prompt.

The attribute format of *key-value* is concordant with the prompt format of *description-value* [33, 35]. According to this schema, we encode fine-grained fashion attributes in sequence format so that our model can capture the inner interaction between name and value. Attributes prompt tells the model precisely the ownership between attribute value and name to utilize the latent semantics from the language model. We design two prompt templates to tackle the diversity of patterns of fashion attributes. The first template covers the enumerable attributes. This kind of attribute has a textual name and the enumerable value from a defined finite set, where each attribute has a unique value. The first template is:

\[ T_e = \text{the image attribute [An] is [Au]} \]

where [An] is the slot to be filled with the attribute name, and [Au] is filled with the attribute value. Another template covers the binary attributes. This attribute annotates a fashion item with a one-hot vector with binary to illustrate whether the fashion item has a certain feature, *e.g.* \{red, pure cotton\}. For binary attributes, the template is:

\[ T_b = \text{is image attribute [Ab]? [As]} \]

where [Ab] is filled with binary attribute label, [As] is filled with positive answer word yes or negative no as attribute value. We concatenate \( T_e \) or \( T_b \) to the tail of the caption tokens during pre-training stage.
3.2. Model Architecture

As illustrated in Fig. 2, FashionSAP consists of an image encoder, a text encoder and a feature fusion module. An image is encoded to $I$.

$$I = \{v_{img}, v_{i0}, v_{i1}, ..., v_{iN}\} \in \mathbb{R}^{(iN+1) \times d}$$

where $v_i$ is a feature vector of a patch of the image generated by IE, $d$ is the dimension of latent semantic space and $i_N$ is the number of patches of the input image. We concatenate the fashion symbol between BERT token [CLS] and fashion text to form a new text sequence shown in the upper-left of Fig. 2. The text sequence is embedded to $E_t$.

$$E_t = \{e_{cls}, e_{symbol}, e_{i0}, ..., e_{iN}\} \in \mathbb{R}^{(tN+2) \times d_e}$$

where $t_N$ is the length of fashion text tokens sequence and $d_e$ is the dimension of text embedding space. The embedding $E_t$ is encoded into $T$.

$$T = \{v_{cls}, v_{symbol}, v_{i0}, ..., v_{iN}\} \in \mathbb{R}^{(tN+2) \times d}$$

For the case in Fig. 2, the $e_{symbol}$ is specific to $e_{tops}$ and $v_{symbol}$ is specific to $v_{tops}$.

Then FashionSAP uses a feature fusion module to fuse the features from the text and image into hybrid feature $H$. The feature fusion module is implemented as multiple cross-attention layers from transformer [38]. The feature of $k$-th cross-attention layer is calculated as Eq. (1)

$$CA^k(T, I) = \text{softmax}(\frac{(W^k_T T)(W^k_i I)^\top}{\sqrt{d}})(W^k_{i,I})^\top$$

where $W^k_T$, $W^k_i$ and $W^k_{i,I}$ are attention parameters in $k$-th cross-attention layer.

3.3. FashionSAP Pre-training Tasks

3.3.1 Fashion Symbol Image Similarity (FSIS)

This task makes the model capture the features from both text and image by maximizing the similarity between the image and the fashion symbol. In this task, the fashion symbol is concatenated between [CLS] and the description tokens as shown in the upper-left of Fig. 2. Let $v_{symbol}$ denote the feature vector of the fashion symbol in the text side and $v_{img}$ denote the feature vector of the image side. We use an adaptive layer $Adp(\cdot)$ to project the feature vector into adapted latent space. Let $\hat{v}_{symbol} = \text{norm}(Adp(v_{symbol})) \in \mathbb{R}^{d_2}$ denote the adapted fashion symbol feature and $\hat{v}_{img} = \text{norm}(Adp(v_{img})) \in \mathbb{R}^{d_2}$ denote the adapted image feature and $d_2$ is the dimension of adapted latent space.

The similarity between the fashion symbols and images is measured by modified vector cosine distance as Eq. (2)

$$\mathcal{L}_{fsis} = \frac{1}{B} \left[ 1 - \frac{1}{B} \sum_{b=1}^{B} \frac{1}{2} [\hat{v}_b^{img}(\hat{v}_b^{symbol})^\top + 1] \right]$$

where $\text{norm}$ is the normalize function, $B$ is the size of mini-batch, $\hat{v}_b^{img}$ is the $b$-th image adapted feature vector and $\hat{v}_b^{symbol}$ is the $b$-th fashion symbol feature vector.
3.3.2 Prompt Token Prediction (PTP)

The goal of the PTP task is to improve the capacity of the model for learning from fine-grained attributes through predicting the correct token under a prompt. In this task, we choose a proper template $T$ and use blank tokens to randomly hold the places of the name or value tokens with a probability of 0.5 to generate attribute input. This task minimizes the cross-entropy loss ($G$). In addition, we use masked language modeling (MLM) task in model pre-training with loss calculated by $G$ as well. So we merge these two losses as Eq. (3)

$$
L_{ptp} = \mathbb{E}_{(T_{ptp}, I) \sim D} G(y_{ptp} \mid g_{ptp}(H_{ptp})) \tag{3}
$$

where $H_{ptp} = [H_{mlm} \oplus H_{ptm}]$ and $\oplus$ means the concatenation between two sequences, $y_{ptp} = [y_{mlm} \oplus y_{ptm}]$ and $H_{mlm}, H_{ptm}$ are hybrid features generated by feature fusion module with masked tokens and prompt tokens input respectively. $y_{ptp}$ is the ground-truth and $g_{ptp}(H_{ptp})$ is the predicted probability distribution of the prompt token prediction task.

3.3.3 Token Replace Prediction (TRP)

In this task, first, we choose some tokens (ratio of 0.15) from the caption and one of the attribute values. Then, half of the chosen tokens are replaced by the antonyms searched by WordNet [27] like [46] and the other half are replaced by random tokens from the vocabulary. This task aims at predicting whether the input tokens are substituted (labels 0 or 1). The loss is shown in Eq. (4)

$$
L_{trp} = \mathbb{E}_{(T_{trp}, I) \sim D} G(y_{trp} \mid g_{trp}(H_{trp})) \tag{4}
$$

$y_{trp}$ is the ground-truth binary label and $g_{trp}(H_{trp})$ is the predicted probability distribution of the replacement task.

3.3.4 Image Text Similarity (ITS)

This task aims at measuring the similarity between the text and the image. We use momentum contrastive learning [8, 17, 28] in this task to take full advantage of text-image pairs. As momentum contrastive learning requires mirror encoders for momentum updating, the vector $v_{cls}$ denotes the whole semantics from the text and $v'_{cls}$ is the corresponding vector generated by momentum text encoder. Let vector $v_{img}$ denote the whole feature from the image and $v'_{img}$ is generated by the momentum image encoder. The momentum distillation [16,17] is also used for label smoothing. For each pair of text and image, the similarities between them are Eq. (5) and Eq. (6)

$$
sim(T, I) = \|W_T v_{cls} \| \|W_I v'_{img} \|^\top \tag{5}
$$

$$
sim(I, T) = \|W_I v_{img} \| \|W_T v'_{cls} \|^\top \tag{6}
$$

where the $W_T$ and $W_I$ are $\mathbb{R}^{(d \times d)}$ are transfer weights to unify feature representations. The similarity between images and texts is measured by $g_{trp}$ and $g_{trp}$ and for $k$-th image and text as Eq. (7) ans Eq. (8)

$$
g^{k}_{trp}(I) = \frac{\exp(\text{sim}(I, T^k) / \tau)}{\sum_{m=1}^{M} \exp(\text{sim}(I, T^m) / \tau)} \tag{7}
$$

$$
g^{k}_{trp}(T) = \frac{\exp(\text{sim}(T, I^k) / \tau)}{\sum_{m=1}^{M} \exp(\text{sim}(T, I^m) / \tau)} \tag{8}
$$

where $\tau$ is a temperature parameter. The loss of the similarity of image and text is as Eq. (9)

$$
L_{its} = \frac{1}{2} \mathbb{E}_{(T, I) \sim D}[G(y_{2t1}(I), g_{trp}(I)) + G(y_{2t1}(T), g_{trp}(T))] \tag{9}
$$

where $y_{2t1}(I)$ and $y_{2t1}(T)$ denote the label of the similarity between images and texts.

3.3.5 Image Text Match (ITM)

In the task of image text match, the first vector of hybrid feature $H_0$ is sent to match head to predict the probability of text-image pair. The loss of this task is Eq. (10)

$$
L_{itm} = \mathbb{E}_{(T, I) \sim D} G(y_{itm}, g_{itm}(H_0)) \tag{10}
$$

where $g_{itm}$ denote the predicted probability distribution by match head and $y_{itm}$ denote the label(1 or 0) of image and text matching. The label is positive if the text-image is matched and negative if mismatched.

The complete pre-training objective of FashionSAP is the combination of the motioned terms above as Eq. (11),

$$
L = L_{its} + L_{ptp} + L_{trp} + L_{itm} \tag{11}
$$

The model is optimized end-to-end on the pre-training datasets by minimizing $L$. 

---

**Figure 3.** Model structure for TMIR task.
### 4. Experiments

#### 4.1. Datasets

We use the FashionGen [31] and FashionIQ [40] datasets for pre-training and downstream tasks. FashionGen [31] includes 320k pairs of text-image and 40k unique fashion items, which are shown as multiple images from multiple views. The detailed description and enumeration attributes are attached to all fashion items. FashionIQ [40] dataset includes 77k unique fashion items and 18k modified text for text modified image retrieval task. We use the train set of FashionGen [31] as pre-training data containing about 260k pairs of text-image. We evaluate downstream tasks text-to-image retrieval, image-to-text retrieval, category recognition and subcategory recognition in FashionGen [31] and text modified image retrieval task in FashionIQ [40].

#### 4.2. Downstream Tasks and Results

**Cross-modal Retrieval** We retrain only two losses, $L_{its}$ and $L_{itm}$ shown in Fig. 2 (lower-right) in this task. Cross-modal retrieval includes two tasks. One task is Image-to-Text (I2T), aiming to retrieve a matched text given a query image. Another task is Text-to-Image (T2I), which aims to retrieve a target image given a query text. We evaluate the performance of the model only by calculating the similarity between text and image following previous works.

FashionSAP gets the SOTA performance as the comparable results shown in Tab. 2. We report the average result of 5 randomly chosen retrieval test sets and each of them contains 1k queries by following previous works. For each query in test sets, only one candidate is matched (positive), while the other 100 candidates are mismatched (negative) and chosen from the same subcategory. For the T2I task, there are 101 candidate images for each query text, and only one image in candidates is matched.

In order to test the performance of our model thoroughly, we also evaluate our model in the full test set of FashionGen [31] in Tab. 3 following [7, 26]. Our model also gets the SOTA performance. Moreover, the differences between the results of our model and others are significant.
We take a fine-tuning stage to the general VLP model (ALBEF) and report the results in Tab. 2 and Tab. 3. We also provide the results of training FashionSAP from scratch with different image encoders following previous works.

**Category/Subcategory Recognition (CR&SCR)** In this task, we only use cross-entropy loss for classification Fig. 2 (upper-right). This downstream tries to recognize the category and the subcategory, given the text and image of the fashion item. We extract the first vector of the fusion feature $H_0$ and input it to a linear layer to predict the category and the subcategory as shown in upper-right in Fig. 2. FashionSAP gets the SOTA performance in both accuracy (Acc) and Macro-F as shown in Tab. 5.

**Text Modified Image Retrieval (TMIR)** This task aims at retrieving a target image of the fashion item by referring to the semantics of the query containing the features from a pair of candidate text-image while the text modifies some elements in the candidate image. As the original pre-training model can not be applied to this task directly, we design a new model structure for this task, shown in Fig. 3. The modified text is encoded into $T_m$ meanwhile candidate image and target image are encoded into $I_{can}$ and $I_{tar}$. Then $T_m$ and $I_{can}$ are blended into hybrid feature $H_f$. The cosine similarity between $H_f$ and $I_{tar}$ is the score between query and target and our model optimizes the similarity between them. Our model gets the SOTA performance compared with previous models, shown in Tab. 4.

### Table 4. Text modified image retrieval performance in FashionIQ [40]

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dress R@10</th>
<th>Dress R@50</th>
<th>Toptee R@10</th>
<th>Toptee R@50</th>
<th>Shirt R@10</th>
<th>Shirt R@50</th>
<th>Mean R@10</th>
<th>Mean R@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIRR [22]</td>
<td>17.45</td>
<td>40.41</td>
<td>21.64</td>
<td>45.38</td>
<td>17.53</td>
<td>38.81</td>
<td>18.87</td>
<td>41.53</td>
</tr>
<tr>
<td>VAL [1]</td>
<td>22.53</td>
<td>44.00</td>
<td>27.53</td>
<td>51.68</td>
<td>22.38</td>
<td>44.15</td>
<td>24.15</td>
<td>46.61</td>
</tr>
<tr>
<td>CosMo [13]</td>
<td>25.64</td>
<td>50.30</td>
<td>29.21</td>
<td>57.46</td>
<td>23.95</td>
<td>47.3</td>
<td>27.78</td>
<td>54.10</td>
</tr>
<tr>
<td>DCNet [12]</td>
<td>28.95</td>
<td>56.7</td>
<td>30.44</td>
<td>58.29</td>
<td>23.95</td>
<td>47.3</td>
<td>27.78</td>
<td>54.10</td>
</tr>
<tr>
<td>FashionVLP [5]</td>
<td>32.42</td>
<td>60.29</td>
<td>38.51</td>
<td>68.79</td>
<td>31.89</td>
<td>58.44</td>
<td>34.27</td>
<td>62.51</td>
</tr>
<tr>
<td>FashionViL [7]</td>
<td>33.47</td>
<td>59.94</td>
<td>34.98</td>
<td>60.79</td>
<td>25.17</td>
<td>50.39</td>
<td>31.21</td>
<td>57.04</td>
</tr>
<tr>
<td>FashionSAP</td>
<td>33.71</td>
<td>60.43</td>
<td>41.91</td>
<td>70.93</td>
<td>33.17</td>
<td>61.33</td>
<td>36.26</td>
<td>64.23</td>
</tr>
</tbody>
</table>

### Table 5. CR and SCR results on FashionGen [31]

<table>
<thead>
<tr>
<th>Methods</th>
<th>CR Acc</th>
<th>CR Macro-F</th>
<th>SCR Acc</th>
<th>SCR Macro-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-BERT [4]</td>
<td>91.25</td>
<td>70.50</td>
<td>85.27</td>
<td>62.00</td>
</tr>
<tr>
<td>K-BERT [49]</td>
<td>95.07</td>
<td>71.40</td>
<td>88.07</td>
<td>63.60</td>
</tr>
<tr>
<td>F-ViL [7]</td>
<td>97.48</td>
<td>88.60</td>
<td>92.23</td>
<td>83.02</td>
</tr>
<tr>
<td>FashionSAP</td>
<td>98.34</td>
<td>88.84</td>
<td>94.33</td>
<td>87.67</td>
</tr>
</tbody>
</table>

### Table 6. Ablation study results for proposed tasks($ptp$, $sis$, $trp$) on five downstream tasks.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$ptp$ R@1</th>
<th>$trp$ R@1</th>
<th>$sis$ R@1</th>
<th>I2T R@1</th>
<th>T2I R@1</th>
<th>CR Macro-F</th>
<th>SCR Macro-F</th>
<th>TMIR R@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-BERT</td>
<td>43.84</td>
<td>51.99</td>
<td>52.09</td>
<td>84.50</td>
<td>84.02</td>
<td>30.22</td>
<td>34.40</td>
<td>30.97</td>
</tr>
<tr>
<td>FashionSAP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>54.43</td>
<td>62.82</td>
<td>89.84</td>
<td>87.67</td>
<td>36.26</td>
</tr>
</tbody>
</table>

### 4.3. Ablation Study

We evaluate the effectiveness of the proposed pre-training tasks in the section. For comparability, the settings in the same series of ablation are consistent. Considering the ITM task and ITS task are similar to general vision-language pre-training, we set the two tasks as basic ones and evaluate the three tasks proposed by this paper in downstream tasks Tab. 6. For conciseness, we list only the index R@1 for both image-to-text and text-to-image tasks, index Macro-F for category (subcategory) recognition and index mean R@10 of three sets in FashionIQ [40] for text modified image retrieval (TMIR).

As we can see from the results of the ablation study in Tab. 6, the loss $sis$ brings a distinct improvement for T2I task as the fashion symbol is an essential structure capturing implicit semantics from the text side to the image side. The loss $ptp$ brings a distinct improvement for I2T task because the prompted fine-grained attributes are encoded as text tokens and share the same embedding layer with text. The loss $trp$ also brings an improvement in downstream tasks as the model learns synonym characteristics through this task.

### 4.4. Fine-grained Alignment Analysis

We choose two instances from FashionGen [31] and show the cross-attention map in the T2I task using the Grad-CAM method Fig. 4 to visualize the improvement of attention score. For each instance, we list the Grad-CAM visualizations from FashionSAP and FashionSAP without
Figure 4. Instances of the comparison of Grad-CAM cross-attention maps for the 1st layer of the feature fusion module from FashionSAP (upper) and FashionSAP without three tasks (lower), Prompt Token Prediction task ($ptp$), Fashion Symbol Image Similarity task ($fsis$) and Token Replace Prediction task ($trp$).

losses $ptp$, $fsis$, $trp$. Compared with the instances without proposed methods, FashionSAP concentrates on the corresponding region precisely. The two instances show that FashionSAP pays proper attention to the whole region of the object (e.g., trousers, leg) rather than the sub-region. FashionSAP can also find all positions of pockets in the attention maps rather than only one.

4.5. Implementation Details

The text encoder is the front 6-layer transformer of BERT-base [2], the image encoder is ViT-B16 [3]. The feature fusion module is a 6-layer transformer. The feed-forward neural network implements the adapters, both on the text and image side. The FashionSAP is initialized by the checkpoint from ALBEF [17] except for the results trained from scratch. The prompt predictor is a multi-layer feed-forward neural network. An AdamW [24] optimizer is adopted with a learning rate $6e^{-5}$. The batch size is 16 with momentum queue size 65535. The size of input images is $256 \times 256$. For training costs, we perform the pre-training stage in 8 Tesla V100*32G GPUs for 20 hours and fine-tuning stage for 10 hours. We randomly choose the attribute name or attribute value and replace them with their synonyms searched by WordNet [27] for raw data preprocessing.

5. Conclusion

This paper introduced a fine-grained fashion VLP model for based on fashion symbols and attributes prompt. We used nine fashion symbols and attributes prompt to enhance the model to capture multi-modal fine-grained semantics. The comparative results and ablation study demonstrated that the FashionSAP was effective in learning fashion representation and outperforms SOTA models significantly.

Several future directions could be considered. Our main goal was to show the potential of the attribute prompt framework to learn fine-grained fashion representation. The fashion symbol only considered category attributes and diversified symbols could be proposed.

6. Acknowledgements

We thank the reviewers for their thoughtful and constructive comments. This work was supported in part by the National Key R&D Program of China (2022ZD0116002), National Science Foundation of China (62276075, 61872113), Science and Technology Planning Project of Shenzhen (JCYJ20190806112210067), Huawei Technologies Co., Ltd. and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China (2022B1212010005).
References


