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

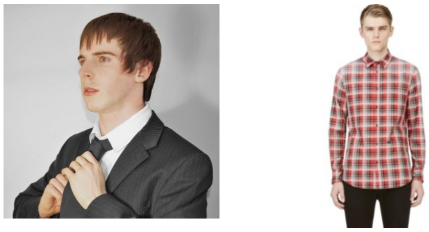
Yunpeng Han[1],Lisai Zhang[1],Qingcai Chen[1,2]*, Zhijian Chen[3],Zhonghua Li[3],Jianxin Yang[3],Zhao Cao[3]
[1]Harbin Institute of Technology Shenzhen, China
[2]PengCheng Laboratory, Shenzhen China, [3]Huawei Technologies Co., Ltd
hanyunpeng.hyp@gmail.com, lisaizhang@foxmail.com, qingcai.chen@hit.edu.cn
{chenzhijian13, lizhonghua3, yangjianxin4, caozhao1}@huawei.com

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 **S**ymbols and **A**ttributes **P**rompt (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]

1. Introduction

Vision-Language pre-training (VLP) attracts wide attention [10, 16–18, 43] 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 [19], 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.



(a) General item
Caption: *a young man in a suit securing his tie.*

(b) Fashion item
Caption: *long sleeve shirt in red, white, and black plaid, single-button barrel cuffs, ...*

Attribute(b): ***Season***: *spring-summer;* ***Gender***: *men, ...*

Figure 1. Two text-image instances from general (a) and fashion domain (b). The captions of the general domain only describe object-level (underlined words) image content, while fashion domain captions emphasise attribute-level semantics.

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-

*Corresponding author
[1] The source code is available at https://github.com/hssip/FashionSAP

Fashion Symbols	Categories	Definition Rules
TOPS	tops, shirt, polo, sweater, ...	upper body
DRESSES	dress, suit, shift, ...	up-to-lower body
SKIRTS	skirt, sarong, slit, kilt, ...	lower body
COATS	jacket, parka, blazer, duffle, ...	associated with others
PANTS	jeans, shorts, breeches, ...	lower body
SHOES	boots, sneakers, pump, loafers, ...	feet
BAGS	clutches, pouches, wristlet, ...	bag & decorative
ACCESSORIES	ring, sunglasses, accessories, hat, necklace, ...	decorative & optional
OTHERS	swim-wear, lingerie, lounge-wear, ...,	-

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-steam 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:

$$\mathcal{T}_e = \texttt{the image attribute [An] is [Au]}$$

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:

$$\mathcal{T}_b = \texttt{is image attribute [Ab]? [As]}$$

where `[Ab]` is filled with binary attribute label, `[As]` is filled with positive answer word `yes` or negative `no` as attribute value. We concatenate $\mathcal{T}_e$ or $\mathcal{T}_b$ to the tail of the caption tokens during pre-training stage.

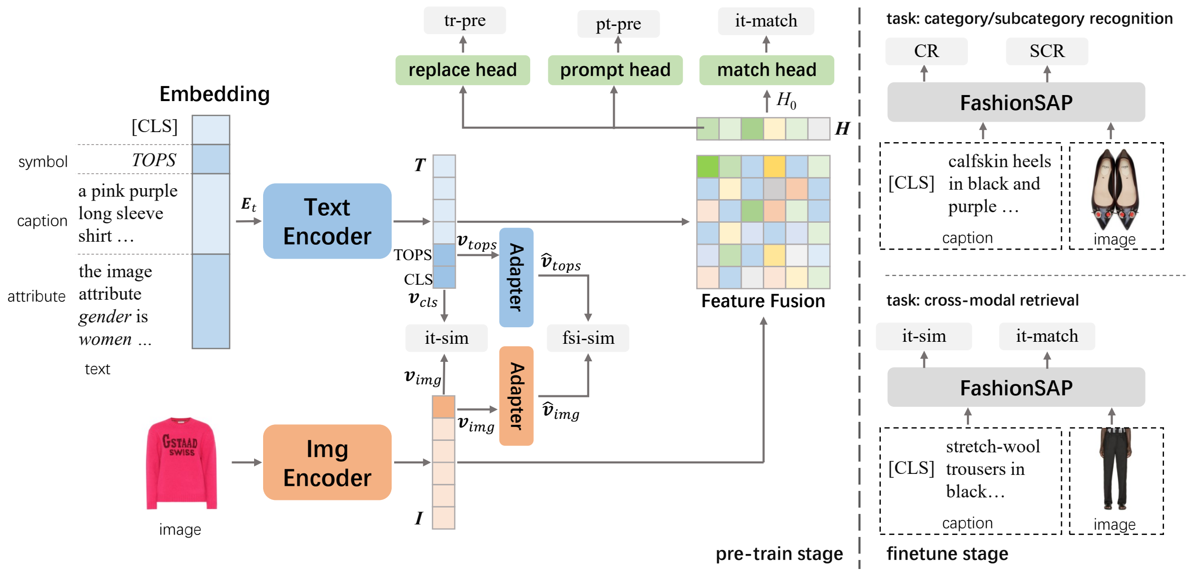


Figure 2. An overview of the FashionSAP framework. The fashion symbol($fsis$ task), attribute prompt (ptp task) and token replace(trp task) are all removed in finetune stage.

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 $\boldsymbol{I}$,

$$\boldsymbol{I}=\{\boldsymbol{v}_{img},\boldsymbol{v}_{i_0},\boldsymbol{v}_{i_1},\boldsymbol{v}_{i_2},...,\boldsymbol{v}_{i_N}\}\in\mathbb{R}^{(i_N+1)\times d}$$

where $\boldsymbol{v}_i$ 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 $\boldsymbol{E}_t$,

$$\boldsymbol{E}_t=\{\boldsymbol{e}_{cls},\boldsymbol{e}_{symbol},\boldsymbol{e}_{t_0},...,\boldsymbol{e}_{t_N}\}\in\mathbb{R}^{(t_N+2)\times d_e}$$

where t_N is the length of fashion text tokens sequence and d_e is the dimension of text embedding space. The embedding $\boldsymbol{E}_t$ is encoded into $\boldsymbol{T}$,

$$\boldsymbol{T}=\{\boldsymbol{v}_{cls},\boldsymbol{v}_{symbol},\boldsymbol{v}_{t_0},...,\boldsymbol{v}_{t_N}\}\in\mathbb{R}^{(t_N+2)\times d}$$

For the case in Fig. 2, the $\boldsymbol{e}_{symbol}$ is specific to $\boldsymbol{e}_{tops}$ and $\boldsymbol{v}_{symbol}$ is specific to $\boldsymbol{v}_{tops}$.

Then FashionSAP uses a feature fusion module to fuse the features from the text and image into hybrid feature $\boldsymbol{H}$. 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)

$$CA^k(\boldsymbol{T},\boldsymbol{I})=softmax(\frac{(W_T^k\boldsymbol{T})(W_{I_1}^k\boldsymbol{I})^\top}{\sqrt{d}})(W_{I_2}^k\boldsymbol{I})^\top \tag{1}$$

where W_T^k, $W_{I_1}^k$ and $W_{I_2}^k\in\mathbb{R}^{d\times d}$ 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 $\boldsymbol{v}_{symbol}$ denote the feature vector of the fashion symbol in the text side and $\boldsymbol{v}_{img}$ 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 $\hat{\boldsymbol{v}}_{symbol}=\mathrm{norm}(Adp(\boldsymbol{v}_{symbol}))\in\mathbb{R}^{d_1}$ denote the adapted fashion symbol feature and $\hat{\boldsymbol{v}}_{img}=\mathrm{norm}(Adp(\boldsymbol{v}_{img})\in\mathbb{R}^{d_1}$ denote the adapted image feature and d_1 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)

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

where norm is the normalize function, B is the size of mini-batch, $\hat{\boldsymbol{v}}_{img}^b$ is the *b-th* image adapted feature vector and $\hat{\boldsymbol{v}}_{symbol}^b$ 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 $\mathcal{T}$ 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)

$$\mathcal{L}_{ptp} = \mathbb{E}_{(\boldsymbol{T}_{ptp},\boldsymbol{I})\sim D} G(\boldsymbol{y}_{ptp}, \boldsymbol{g}_{ptp}(\boldsymbol{H}_{ptp})) \tag{3}$$

where $\boldsymbol{H}_{ptp} = [\boldsymbol{H}_{mlm} \oplus \boldsymbol{H}_{pmt}]$ and $\oplus$ means the concatenation between two sequences, $\boldsymbol{y}_{ptp} = [\boldsymbol{y}_{mlm} \oplus \boldsymbol{y}_{pmt}]$ and $\boldsymbol{H}_{mlm}, \boldsymbol{H}_{pmt}$ are hybrid features generated by feature fusion module with masked tokens and prompt tokens input respectively. $\boldsymbol{y}_{ptp}$ is the ground-truth and $\boldsymbol{g}_{ptp}(\boldsymbol{H}_{ptp})$ 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)

$$\mathcal{L}_{trp} = \mathbb{E}_{(\boldsymbol{T}_{trp},I)\sim D} G(\boldsymbol{y}_{trp}, \boldsymbol{g}_{trp}(\boldsymbol{H}_{trp})) \tag{4}$$

$\boldsymbol{y}_{trp}$ is the ground-truth binary label and $\boldsymbol{g}_{trp}(\boldsymbol{H}_{trp})$ 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 $\boldsymbol{v}_{cls}$ denotes the whole semantics from the text and $\boldsymbol{v}'_{cls}$ is the corresponding vector generated by momentum text encoder. Let vector $\boldsymbol{v}_{img}$ denote the whole feature from the image and $\boldsymbol{v}'_i$ 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)

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

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

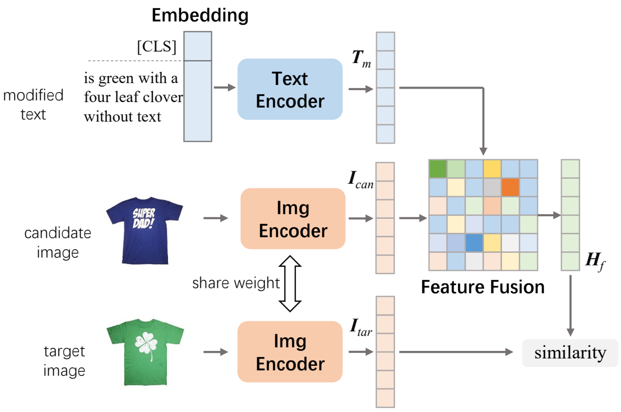


Figure 3. Model structure for TMIR task.

where the W_T and $W_I \in \mathbb{R}^{(d\times d)}$ are transfer weights to unify feature representations. The similarity between images and texts is measured by $\boldsymbol{g}_{i2t}$ and $\boldsymbol{g}_{t2i}$ and for *k-th* image and text as Eq. (7) ans Eq. (8)

$$g^k_{i2t}(\boldsymbol{I}) = \frac{\exp(\text{sim}(\boldsymbol{I}, \boldsymbol{T}^k)/\tau)}{\sum_{m=1}^{M} \exp(\text{sim}(\boldsymbol{I}, \boldsymbol{T}^m))} \tag{7}$$

$$g^k_{t2i}(\boldsymbol{T}) = \frac{\exp(\text{sim}(\boldsymbol{T}, \boldsymbol{I}^k)/\tau)}{\sum_{m=1}^{M} \exp(\text{sim}(\boldsymbol{T}, \boldsymbol{I}^m))} \tag{8}$$

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

$$\begin{aligned}\mathcal{L}_{its} = \frac{1}{2}\mathbb{E}_{(\boldsymbol{T},\boldsymbol{I})\sim D}[G(\boldsymbol{y}_{i2t}(\boldsymbol{I}), \boldsymbol{g}_{i2t}(\boldsymbol{I})) +\\ G(\boldsymbol{y}_{t2i}(\boldsymbol{T}), \boldsymbol{g}_{t2i}(\boldsymbol{T}))]\end{aligned} \tag{9}$$

where $\boldsymbol{y}_{i2t}(\boldsymbol{I})$ and $\boldsymbol{y}_{t2i}(\boldsymbol{T})$ 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)

$$\mathcal{L}_{itm} = \mathbb{E}_{(T,I)\sim D} G(\boldsymbol{y}_{itm}, \boldsymbol{g}_{itm}(H_0)) \tag{10}$$

where $\boldsymbol{g}_{itm}$ denote the predicted probability distribution by match head and $\boldsymbol{y}_{itm}$ 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),

$$\mathcal{L} = \mathcal{L}_{fsis} + \mathcal{L}_{ptp} + \mathcal{L}_{trp} + \mathcal{L}_{its} + \mathcal{L}_{itm} \tag{11}$$

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

Methods	I2T			T2I			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	R@1
VL-BERT [34]	19.26	39.90	46.05	22.63	36.48	48.52	20.95
ViLBERT [25]	20.97	40.49	48.21	21.12	37.23	50.11	21.05
Image-BERT [29]	22.76	41.89	50.77	24.78	45.20	55.90	23.77
OSCAR [18]	23.39	44.67	52.55	25.10	49.14	56.68	24.25
FashionBERT [4]	23.96	46.31	52.12	26.75	46.48	55.74	25.36
KaleidoBERT [49]	27.99	60.09	68.37	33.88	60.60	68.59	30.94
EI-CLIP [26]	38.70	72.20	84.25	40.06	71.99	82.90	39.38
CommerceMM [45]	41.60	64.00	72.80	39.60	61.50	72.70	62.75
ALBEF [17]	63.97	88.92	94.41	60.52	84.99	91.45	62.20
FashionViL [7]	65.54	91.34	96.30	61.88	87.32	93.22	63.71
FashionSAP(Resnet50)	67.23	91.30	96.41	64.11	88.24	94.31	65.67
FashionSAP(ViT-B16)	71.14	92.21	96.52	69.07	89.81	94.75	70.11
FashionSAP	**73.14**	**92.80**	**96.87**	**70.12**	**91.76**	**96.38**	**71.63**

Table 2. Cross-modal retrieval result on FashionGen [31] in the sub set of evaluation following previous work.

Methods	I2T			T2I			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	R@1
EI-CLIP [26]	25.70	54.50	66.80	28.40	57.10	69.40	27.05
ALBEF [17]	41.68	67.39	75.50	50.95	75.36	84.15	46.32
FashionViL [7]	42.88	71.57	80.55	51.34	75.42	84.57	47.11
FashionSAP(Resnet50)	44.92	71.49	81.64	52.45	76.63	84.71	48.69
FashionSAP(ViT-B16)	50.34	74.34	81.67	58.43	80.06	87.02	54.39
FashionSAP	**54.43**	**77.30**	**83.15**	**62.82**	**83.96**	**90.16**	**58.63**

Table 3. Cross-modal retrieval result on FashionGen [31] with full evaluation

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, $\mathcal{L}_{its}$ and $\mathcal{L}_{itm}$ 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.

Methods	Dress		Toptee		Shirt		Mean	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CIRR [22]	17.45	40.41	21.64	45.38	17.53	38.81	18.87	41.53
VAL [1]	22.53	44.00	27.53	51.68	22.38	44.15	24.15	46.61
CosMo [13]	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31
DCNet [12]	28.95	56.7	30.44	58.29	23.95	47.3	27.78	54.10
FashionVLP [5]	32.42	60.29	38.51	68.79	31.89	58.44	34.27	62.51
FashionViL [7]	33.47	59.94	34.98	60.79	25.17	50.39	31.21	57.04
FashionSAP	**33.71**	**60.43**	**41.91**	**70.93**	**33.17**	**61.33**	**36.26**	**64.23**

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

Methods	CR		SCR	
	Acc	Macro-F	Acc	Macro-F
F-BERT [4]	91.25	70.50	85.27	62.00
K-BERT [49]	95.07	71.40	88.07	63.60
F-ViL [7]	97.48	88.60	92.23	83.02
FashionSAP	**98.34**	**89.84**	**94.33**	**87.67**

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

ptp	*trp*	*fsis*	I2T	T2I	CR	SCR	TMIR
			R@1	R@1	Macro-F	Macro-F	R@10
			43.84	53.24	84.50	84.42	30.02
✓			51.99	53.78	86.32	86.03	34.40
✓	✓		52.09	55.54	86.51	86.65	35.01
✓	✓	✓	**54.43**	**62.82**	**89.84**	**87.67**	**36.26**

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

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 $\boldsymbol{T}_m$ meanwhile candidate image and target image are encoded into $\boldsymbol{I}_{can}$ and $\boldsymbol{I}_{tar}$. Then $\boldsymbol{T}_m$ and $\boldsymbol{I}_{can}$ are blended into hybrid feature $\boldsymbol{H}_f$. The cosine similarity between $\boldsymbol{H}_f$ and $\boldsymbol{I}_{tar}$ 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.

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 $fsis$ 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

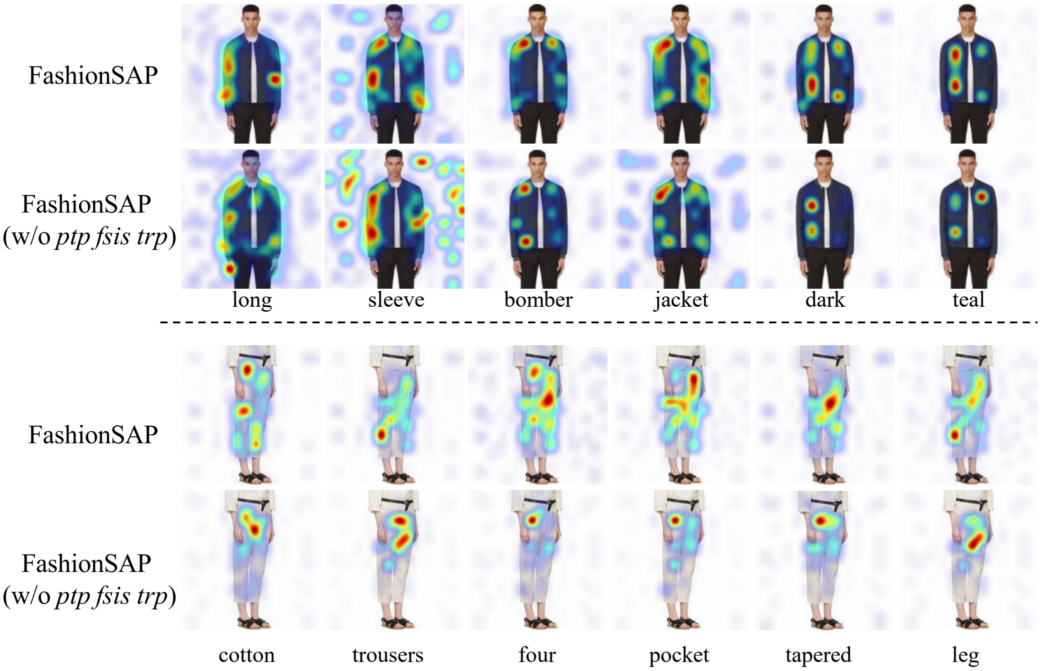


Figure 4. Instances of the comparison of Grad-CAM cross-attention maps for the 1st layer of the feature fusion module from FashionSAP (upper) and FashionSAP without three tasks (lower), Prompt Token Prediction task(ptp), Fashion Symbol Image Similarity task($fsis$) and Token Replace Prediction task(trp).

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 × 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. Acknowledegements

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), Natural 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

[1] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 3, 7

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 2, 8

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 8

[4] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2020. 2, 6, 7

[5] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115, 2022. 3, 7

[6] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017. 2

[7] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*. Springer, 2022. 2, 6, 7

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[10] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 1, 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[12] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, pages 1771–1779, 2021. 7

[13] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 3, 7

[14] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3

[15] Dongfang Li, Baotian Hu, and Qingcai Chen. Prompt-based text entailment for low-resource named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1896–1903, 2022. 3

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2, 5

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 5, 6, 8

[18] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2, 3

[22] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pretrained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 7

[23] Robert Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down

on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, 2022. 3

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 6

[26] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061, 2022. 3, 6

[27] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5, 8

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[29] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 6

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[31] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 2, 3, 6, 7

[32] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 3

[33] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, 2021. 3

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 6

[35] Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, 2021. 3

[36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[37] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*, pages 390–405, 2018. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[39] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 3

[40] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 2, 3, 6, 7

[41] Hongfa Wu, Lisai Zhang, Qingcai Chen, Yimeng Deng, Joanna Siebert, Yunpeng Han, Zhonghua Li, Dejiang Kong, and Zhao Cao. Contrastive label correlation enhanced unified hashing encoder for cross-modal retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2158–2168, 2022. 2

[42] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2, 3

[43] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1, 2

[44] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3

[45] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4433–4442, 2022. 3, 6

[46] Lisai Zhang, Qingcai Chen, Zhijian Chen, Yunpeng Han, Zhonghua Li, and Zhao Cao. Replacement as a self-supervision for fine-grained vision-language pre-training. 2023. 5

[47] Lisai Zhang, Hongfa Wu, Qingcai Chen, Yimeng Deng, Joanna Siebert, Zhonghua Li, Yunpeng Han, Dejiang Kong, and Zhao Cao. Vldeformer: Vision–language decomposed transformer for fast cross-modal retrieval. *Knowledge-Based Systems*, 252:109316, 2022. 2

[48] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao.

Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[49] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657, 2021. 6, 7

[50] Xingxing Zou, Kaicheng Pang, Wen Zhang, and Waikeung Wong. How good is aesthetic ability of a fashion model? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21200–21209, 2022. 2, 3