# FashionViL: Fashion-Focused Vision-and-Language Representation Learning

Xiao Han[1,2], Licheng Yu[3], Xiatian Zhu[1,4], Li Zhang[5]
Yi-Zhe Song[1,2], and Tao Xiang[1,2]

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
[3] Meta AI
[4] Surrey Institute for People-Centred Artificial Intelligence, University of Surrey
[5] School of Data Science, Fudan University
{xiao.han,xiatian.zhu,y.song,t.xiang}@surrey.ac.uk
lichengyu@fb.com lizhangfd@fudan.edu.cn

**Abstract.** Large-scale Vision-and-Language (V+L) pre-training for representation learning has proven to be effective in boosting various downstream V+L tasks. However, when it comes to the fashion domain, existing V+L methods are inadequate as they overlook the unique characteristics of both the fashion V+L data and downstream tasks. In this work, we propose a novel *fashion-focused* V+L representation learning framework, dubbed as FashionViL. It contains two novel fashion-specific pre-training tasks designed particularly to exploit two intrinsic attributes with fashion V+L data. First, in contrast to other domains where a V+L data point contains only a single image-text pair, there could be multiple images in the fashion domain. We thus propose a Multi-View Contrastive Learning task for pulling closer the visual representation of one image to the compositional multimodal representation of another image+text. Second, fashion text (*e.g.*, product description) often contains rich fine-grained concepts (attributes/noun phrases). To exploit this, a Pseudo-Attributes Classification task is introduced to encourage the learned unimodal (visual/textual) representations of the same concept to be adjacent. Further, fashion V+L tasks uniquely include ones that do not conform to the common one-stream or two-stream architectures (*e.g.*, text-guided image retrieval). We thus propose a flexible, versatile V+L model architecture consisting of a modality-agnostic Transformer so that it can be flexibly adapted to any downstream tasks. Extensive experiments show that our FashionViL achieves new state of the art across five downstream tasks. Code is available at https://github.com/BrandonHanx/mmf.

**Keywords:** Vision and Language, Representation learning, Fashion.

## 1 Introduction

Recently, Vision-and-Language (V+L) pre-training has received increasing attention [34,57,43,55,8,37,50,31,33,66]. The objective is to learn multimodal rep-

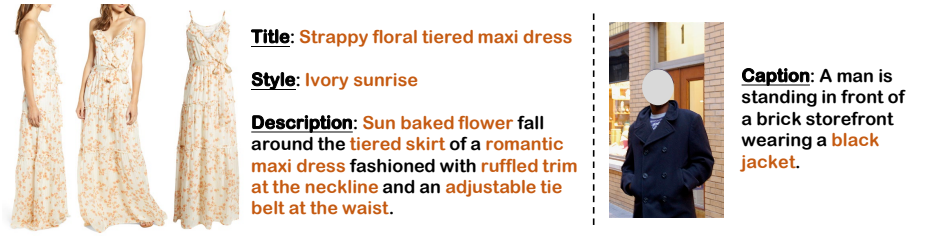


**Fig. 1.** Left and right are examples from fashion dataset FACAD [70] and Flickr30k [48], respectively. It can be seen that fashion data often have multiple images from different angles, associated with structured titles and descriptions with multiple fine-grained attributes (highlighted in color)

resentations from large-scale image-text pairs, in order to improve various downstream unimodal or multimodal tasks. These models have proven to be highly effective thanks to two main factors: (*i*) there are plenty of image-text pairs on the Web providing abundant training data for free (no additional annotation required), and (*ii*) Transformer-based model architectures have been widely used to learn the contextualized representation of multimodal inputs.

In this work, we focus on the fashion domain, for which V+L pre-training seems particularly suitable. First, fashion V+L data are not just copious in volume but also high in quality. Online fashion shopping is increasingly ubiquitous; on an e-commerce website, each product detail page (PDP) contains product images and text, both are of very high quality (*i.e.*, often generated by domain experts). Second, there are plenty of downstream tasks, more so than other domains, in real-world applications, ranging from multimodal product understanding [38,44], cross-modal retrieval [18], to text-guided image retrieval [67]. However, when applied to the fashion domain, we observe that existing SOTA V+L pre-training methods [18,79] are less effective compared to other domains (see Sec. 4). We believe that this is because they are not designed to exploit some unique characteristics of both fashion V+L data and downstream tasks.

In particular, in most existing generic domain V+L datasets (*e.g.*, COCO [39] and Flickr30k [48]), each data point is a single image-text pair and the text is often brief (*e.g.*, an image caption as shown in Fig. 1). In contrast, fashion datasets are collected mostly from PDPs on e-commerce sites and thus have two specialties: (*i*) There are typically more than one image associated with a given text. One example is shown in Fig. 1. The garment 'maxi dress' is presented with three different views so that online shoppers can view the dress from different angles. (*ii*) There are many more fine-grained concepts in the text description as the text serves as the product description. As shown in Fig. 1, the fashion text is more focused on the garment itself with very detailed adjectives and nouns, describing its appearance in the title, style, and description. To show that this is statistically true, we calculate the ratio on four combined fashion datasets [52,23,70,60] and two combined generic datasets [48,39]. We found that

82% of the words in the fashion captions are adjectives or nouns, while this ratio becomes only 59% for the generic captions. None of the existing V+L models are capable of exploiting these specialties in fashion data.

Fashion downstream tasks are also more diverse than those in the generic domain, posing a challenge to the V+L pre-training model architecture design. More specifically, in the generic V+L domain, existing models are either single-stream or two-stream, depending on the intended downstream tasks. For example, the single-stream model [34,55,8,31,28] that operates on the concatenation of image and text tokens are suitable for multimodal fusion tasks such as VQA [2], VCR [73] and RefCOCO [72]. In contrast, the two-stream model [43,57,29,50,56] are typically designed for efficient cross-modal retrieval tasks[6]. However, in the fashion domain, apart from image-text fusion and cross-modal retrieval downstream tasks, there are also tasks for which neither single-stream nor two-stream architectures are suitable. For example, the text-guided image retrieval task [62,67,21] not only requires a high-quality fusion of the reference image and the modified text but also an efficient matching between the fused multimodal representation and the candidate image. Due to the diversity of fashion downstream tasks, the existing models, either one-stream or two-stream, do not have the required flexibility and versatility.

To overcome the limitations of existing models for fashion, we introduce a novel fashion-focused V+L representation learning framework termed FashionViL. Two fashion-focused pre-training tasks are proposed to fully exploit the specialties of fashion data. The first task is Multi-View Contrastive Learning (MVC). Given a fashion data item with multiple images/views and one text description, we assume that each modality (no matter it is unimodal or multimodal) should be semantically similar to each other since they are all referring to the same product. Thus, other than the common image-text matching, we propose to minimize the distance between ($a$) the multimodal representation of one of its views and text, and ($b$) the other views. The second task is Pseudo-Attributes Classification (PAC), designed to exploit the rich fine-grained fashion concepts in the description. Specifically, we extract those common attributes/noun phrases from the fashion datasets and construct a pseudo attribute set. The model then learns to predict those attributes during pre-training explicitly. PAC encourages the fashion items with the same attribute(s) to be clustered together so that the learned representations become more discriminative. We show that (see Sec 4.3) these new pre-training tasks are effective and complementary to conventional pre-training tasks such as Image-Text Contrastive Learning (ITC) and Masked Language modeling (MLM).

Furthermore, a flexible and versatile model architecture is designed to make the pre-trained model easily adaptable to a diverse set of downstream tasks. The new design keeps the superior fusion ability of single-stream model and the scalability of two-stream model. Crucially, it also caters for fashion-domain unique tasks such as text-guided image retrieval and outfit complementary item

[6] A single-stream model can also be applied but it needs to traverse every pair of query and gallery item, resulting in unacceptable retrieval speed in large-scale applications.

retrieval. Specifically, our model consists of an image encoder and a modality-agnostic Transformer module, which can be used as either a text encoder or a multimodal fusion encoder. It thus can be easily fine-tuned for three different downstream use cases: (*i*) early-fusion single-stream mode for joint representation learning, *e.g.*, multimodal classification; (*ii*) late-fusion two-stream mode for unimodal representation learning, *e.g.*, cross-modal retrieval; (*iii*) early-fusion two-stream architecture for compositional representation learning, *e.g.*, text-guided image retrieval.

In summary, our contributions are as follows: (1) A novel V+L pre-training framework is proposed specifically for the fashion domain, which can exploit the specialties of fashion data through two new V+L pre-training tasks. (2) A flexible architecture design is introduced with a shared text encoder and fusion encoder, which can be easily adapted to a set of diverse fashion downstream tasks. (3) To demonstrate the generalization of FashionViL, we evaluate our model on 5 fashion V+L tasks: image-to-text retrieval, text-to-image retrieval [52], text-guided image retrieval [67], (sub)category recognition [52] and outfit complementary item retrieval [60]. The experiments show that FashionViL achieves a new state of the art (SOTA) with a consistent and significant performance boost across every downstream task. To the best of our knowledge, this is the first work capable of addressing 5 diverse fashion tasks together.

## 2 Related work

With the advent of Transformer [61] and its success in NLP [10] and CV [13], there has been great success in applying large-scale V+L pre-training to generic domain [34,8,33,50]. Some recent studies started to focus on e-commerce domains including fashion [18,79,78,11,76]. Existing works differ in two main aspects: architecture design and pre-training tasks.

**Model architecture.** All V+L pre-training methods use image and text embedding sequences as input for modeling inter-modal and optionally intra-modal interactions through a CNN or Transformer architecture, and output a contextualized feature sequence [6]. There are many options on architecture designs on different aspects, including singe-stream early fusion [34,55,8,37] *vs.* two stream late fusion [57,43,29,50,17], or different visual features (*e.g.,* detector-based regions [75] *vs.* ConvNet patches [28] *vs.* linear projections [31,69]). In many case, the design is driven by the intended downstream tasks (*e.g.*, VQA requires earlier fusion to enhance joint representation whereas cross-modal retrieval requires later fusion to speed up inference). There are also efforts for alleviating the gap between different architectures through retrieve-and-rerank strategy [56,19] or knowledge distillation [65,41]. Unlike them, inspired by the recent advances in modality-agnostic models [1,71,64,63,35], we introduce a unified architecture that can be easily switched between the single-stream or two-stream mode, so there is no need to modify the architecture for different downstream tasks.

**Pre-training tasks.** Various tasks have been proposed for V+L pre-training. Masked Language Modeling (MLM) and Image-Text Matching (ITM) are the

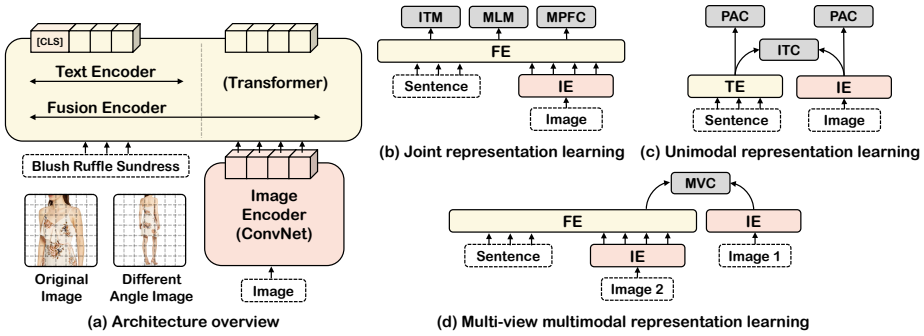


**Fig. 2.** Overview of the proposed FashionViL model architecture, consisting of an image encoder, a text encoder and a fusion encoder. Text encoder and fusion encoder share the same parameters. We adopt six pre-training tasks to learn different representations

direct counterparts of the BERT objectives [10,34]. Masked Image Modeling (MIM) is the extension of MLM on the visual modality, including several variants like masked region classification [43,55] and masked region feature regression [8]. Some other tasks are also proved to be effective, such as predicting object tags [37,27], sequential caption generation [77,66] and image-text contrastive learning [33,50,36]. However, none of these tasks are able to take advantage of the two specialities of fashion data as discussed earlier. We therefore propose two fashion-focused pre-training tasks in this work.

## 3 Methodology

### 3.1 Model overview

The model architecture of FashionViL is illustrated in Fig. 2(a), which is composed of an image encoder (IE) and a Transformer module that can be used for both text encoder (TE) and fusion encoder (FE). Specifically, our image encoder uses ConvNet as its backbone to convert the raw pixels into a sequence of visual embeddings by rasterizing the grid features of the final feature map. For the text encoder, we follow BERT [10] to tokenize the input sentence into WordPieces [68]. Each sub-word token's embedding is obtained by summing up its word embedding and learnable position embedding, followed by LN [3].

One novelty of the model design lies in the shared Transformer for TE and FE, which allows us to flexibly build various multimodal model architectures, each of which is suited for different types of downstream tasks. For example, Fig. 2(b) shows an early-fusion model architecture, where the raw sentence and the computed image embeddings are jointly fed into the multimodal fusion encoder. Note that when we use the Transformer as the fusion encoder, we will further add the modality embeddings to the visual embeddings and word embeddings, helping the model distinguish the modality type. This architecture

is exactly the same as the well-known single-stream models in many previous pre-training works [34,8,18]. Then in Fig. 2(c) we show a late-fusion two-stream model architecture, where we apply the shareable Transformer as the text encoder. The outputs from image encoder and text encoder are interacted with a simple dot product to compute the similarity between two modalities. This architecture has been widely adopted for efficient large-scale cross-modal retrieval [56,19]. Furthermore, we can fine-tune this shared Transformer to a more complicated two-stream architecture variant, shown in Fig. 2(d). Here, one stream operates in an early-fusion manner while the other stream is an image encoder. This architecture is needed for some fashion-focused retrieval tasks with multimodal query, *e.g.*, text-guided image retrieval [62,67]. Note that all FE and TE in the above three architectures are actually the same Transformer, and the mere difference lies in its input.

Given an image-text pair, we denote its raw visual inputs as $\mathbf{v}_i = \left\{\mathbf{v}_i^1, \ldots, \mathbf{v}_i^K\right\}$, and its input words as $\mathbf{w}_i = \left\{\mathbf{w}_i^{\text{cls}}, \mathbf{w}_i^1, \ldots, \mathbf{w}_i^T\right\}$, where the subscript $i$ indicates the i-th pair in the dataset. An additional special `[CLS]` token is inserted at the beginning of the text sequence, as well as the multimodal sequence when modalities are concatenated. We follow the common pre-training + fine-tuning pipeline when applying the model to downstream tasks.

### 3.2 Pre-training tasks

We first introduce two new pre-training tasks. This is followed by the other conventional pre-training tasks adopted in our framework.

**Multi-view contrastive learning (MVC).** As can be seen in Fig. 1, each fashion item is often associated with multiple views to provide a comprehensive overview of the product. To take advantage of the reciprocal information between different views, we propose to build a correlation between (a) the visual representation of the original view $\mathbf{v}$, and (b) the compositional representation of another view $\mathbf{d}$ and the text $\mathbf{w}$. In cases where there is only one view of the product, we augment another view by randomly cropping or horizontally flipping the given view. As shown in Fig. 2(d), the visual representation of the original view is extracted by the image encoder while the compositional representation is calculated in an early fusion way. Therefore, the similarity between the multimodal input $[\mathbf{w}; \mathbf{d}]$[7] and $\mathbf{v}$ can be computed as:

$$s\left([\mathbf{w}_i; \mathbf{d}_i], \mathbf{v}_j\right) = g_\theta\left(\mathbf{d}_i^{\text{avg}} | \mathbf{w}_i\right)^T g_\theta\left(\mathbf{v}_j^{\text{avg}}\right), \tag{1}$$

where $g$ represents a linear transformation that projects the average pooled features into the normalized low-dimensional latent space. Next, we apply two symmetrical InfoNCE losses [46] to pull closer the matched compositional representations and visual representations in the shared latent space:

$$\mathcal{L}_{\text{InfoNCE}}(x, y) = -\mathbb{E}_{(x,y)\sim B} \log \frac{\exp(s(x, y)/\tau)}{\sum_{\hat{y}\in\hat{B}} \exp(s(x, \hat{y})/\tau)}, \tag{2}$$

[7] We randomly dropout some words in $\mathbf{w}$ and patches in $\mathbf{d}$ with the probability of 15% to make the learning process more robust.

$$\mathcal{L}_{\mathrm{MVC}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{InfoNCE}}([\mathbf{w};\mathbf{d}],\mathbf{v}) + \mathcal{L}_{\mathrm{InfoNCE}}(\mathbf{v},[\mathbf{w};\mathbf{d}])\right], \tag{3}$$

where $\tau$ is a learnable temperature and $\hat{B}$ contains the positive sample $y$ and $|\hat{B}|-1$ negative samples drawn from a mini-batch $B$.

**Pseudo-attribute classification (PAC).** As mentioned in Sec. 1, we found that there are a large number of fine-grained attributes in the fashion description. We propose to mine the pseudo-attribute concepts from all the available textual information, including title, description and meta-info. Specifically, we extract all nouns and adjectives via NLTK tagger [5] and only keep those that appear more than 100 times, resulting in a list of 2,232 attributes. We show the histogram of the top-50 pseudo attributes in Fig. 3. It is observed that all of them are truly highly-related to the fashion domain.

Then we explore how to utilize such mined concepts. We aim to let our model learn to explicitly recognize those pseudo attributes during the pre-training stage. We model this task as a multi-label classification problem, called Pseudo-Attribute Classification (PAC). As shown in Fig. 2(c), we apply the PAC to both visual and textual modalities so that both encoders can learn to capture the fine-grained concepts. As this is a weakly-supervised learning setting, we leverage label smoothing to generate the labels [25] considering that the mined labels can be noisy. We use $A$ to denote the whole 2,232 pseudo-attribute set and $a$ as the smoothed soft-target for each class. For example, if one sample has two ground truth labels at position 0 and 1, then $a_0 = a_1 = 0.5$ while $a_i = 0$ $(i \neq 0, 1)$. Our objective is as follows:

$$\mathcal{L}_{\mathrm{PAC}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}\mathbb{E}_{a\sim A}\left[a\log P_\theta\left(a|\mathbf{w}\right) + a\log P_\theta\left(a|\mathbf{v}\right)\right], \tag{4}$$

where $\theta$ is the learnable parameters and each pair $(\mathbf{w},\mathbf{v})$ is sampled from the whole training set $D$.

**Masked patch feature classification (MPFC).** While the naive masked feature regression has been shown not helpful in V+L pre-training [31,14], we found empirically our version of masked patch modeling being effective in the fashion domain. Specifically, we disregard the feature reconstruction of each masked patch, but instead predict the patch label given by an offline image tokenizer. To this end, we first train a discrete VAE [59,51,15] as the image tokenizer on our collected fashion images with the perceputal loss [12]. We also adopt exponential moving average (EMA) to update the codebook, which is proved to be useful for increasing the utilization of codewords [59,12]. We randomly replace 25% patch features with zeros through block-wise masking strategy [4][8]. Since now we have discrete labels for each patch, the model can be trained to predict the label of each masked patches $\mathbf{v_m}$ given the remaining patches $\mathbf{v}_{\backslash\mathbf{m}}$ by optimizing:

$$\mathcal{L}_{\mathrm{MPFC}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}\log P_\theta\left(\mathbf{v_m^t}|\mathbf{v}_{\backslash\mathbf{m}},\mathbf{w}\right), \tag{5}$$

where $\mathbf{v_m^t}$ is the estimated target label for the masked patch.

[8] Following UNITER, we use conditional masking for MLM/MPFC, *i.e.*, only masking one modality while keeping the other one intact at each time.

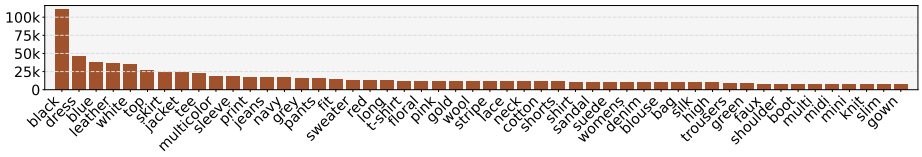


**Fig. 3.** Histogram of the top-50 pseudo attributes

**Image-text contrastive learning (ITC).** We also use ITC to encourage the two unimodal representations to be close in the latent space. As shown in Fig. 2(c), the similarity of $\mathbf{w}$ and $\mathbf{v}$ is measured by the dot product of their average pooled features after being projected to the latent space with two linear transformations $f$ and $g$: $s\left(\mathbf{w}_i, \mathbf{v}_j\right) = f_\theta\left(\mathbf{w}_i^{\text{avg}}\right)^T g_\theta\left(\mathbf{v}_j^{\text{avg}}\right)$. The ITC loss is:

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{w}, \mathbf{v}) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{v}, \mathbf{w})\right]. \tag{6}$$

**Masked language modeling (MLM).** In MLM, we randomly mask out the input words with a probability of 15%, and replace all subwords belonging to the masked words $\mathbf{w_m}$ with special token `[MASK]`[9]. The goal of MLM is to predict these masked sub-words based on the observation of their surrounding words $\mathbf{w}_{\backslash \mathbf{m}}$ and all image patches $\mathbf{v}$, by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta\left(\mathbf{w_m} | \mathbf{w}_{\backslash \mathbf{m}}, \mathbf{v}\right). \tag{7}$$

**Image-text matching (ITM).** In ITM, the input is an image-text pair and the target is a binary label $z \in \{0, 1\}$, indicating if each input pair is a match. Following [33], we sample the hard negative pairs from the similarity matrix $s\left(\mathbf{w}_i, \mathbf{v}_j\right)$ computed by ITC and then make a mini-batch $H$ containing 50% negative pairs. We extract the hidden output of `[CLS]` at the last layer to represent the joint representation of both modalities, then feed it into a FC layer to do a two-class classification. We apply cross-entropy loss for ITM:

$$\mathcal{L}_{\text{ITM}} = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim H} \log P_\theta\left(z | \mathbf{w}, \mathbf{v}\right). \tag{8}$$

## 4 Experiments

In this section, we introduce our pre-training dataset and 5 practical downstream tasks. We use MMF [54] and PyTorch [47] for the implementation. For the image encoder, we use an off-the-shelf ResNet50 [24] to fairly compare with previous methods, most of which also used ResNet50. For the text encoder and multimodal fusion encoder (using the shared Transformer), we use the BERT-base-uncased [55] as the initialization. We use 4 RTX 3090 GPUs for the pre-training. The details of the hyper-parameters are listed in the supplementary file.

[9] Following BERT and UNITER, we decompose this 15% into 10% random words, 10% unchanged, and 80% `[MASK]`.

**Table 1.** Statistics on the datasets used for pre-training

| Datasets | FashionGen [52] | | FACAD [70] | | Fashion200k [23] | | PolyvoreOutfits [60] | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *#products* | *#pairs* | *#products* | *#pairs* | *#products* | *#pairs* | *#products* | *#pairs* | *#products* | *#pairs* |
| **Train** | 60k | 260k | 164.5k | 847k | 77k | 172k | 72k | 72k | 373.5k | 1.35M |
| **Val** | 7.5k | 32.5k | 18k | 94k | 13k | 30k | 14.5k | 14.5k | 53k | 171k |

### 4.1 Pre-training dataset and downstream tasks

**Pre-training dataset.** Our pre-training dataset consists of 4 public fashion-related datasets, namely, FashionGen [52], FACAD [70], Fashion200K [23] and PolyvoreOutfits [60]. In total, these datasets provide us with 373.5K fashion products for pre-training. Because each product may contain multiple images from different angles, we have about 1.35 million image-text pairs on hand. The detailed statistics are provided in Table 1.

**Cross-modal retrieval.** Image-to-Text Retrieval (ITR) is a cross-modal retrieval task. Given an image query, our model finds the most aligned text from a large candidate pool. Previous fashion-domain pre-training works [18,79] use the joint representation over the `[CLS]` token to predict the matching score, which results in an impractical time complexity due to the exhaustive matching between each query item and all gallery items in the early-fusion model [56,65,41,74,19]. While one of our model architectures can do the same (as Fig. 2(b)), we opt to use the two-stream late-fusion model in Fig. 2(c) to compute the cosine similarity for a far more efficient retrieval as [29,50]. Text-to-Image Retrieval (TIR) is an inverse problem of ITR, where the query modality and gallery modality are swapped. The architecture for TIR is the same as ITR.

**Text-guided image retrieval (TGIR).** TGIR is a special type of image retrieval problem, whose query is a multimodal composition [20,62,67,21]. Specifically, given a query image and a modified sentence, the model is required to retrieve another image which has the similar outlook as the query image but with some appearance changes according to the query text. It has many practical applications in fashion, such as retrieving another garment according to a user's reference garment and his/her feedback. To handle the uniqueness of the multimodal query, several interesting fusion approaches have been proposed in the past, such as the gating mechanism [62,53], hierarchical attention [7], and style-content modification [32]. In this work, we follow [42] to simply apply an early fusion model to encode the compositional representation of the query image and modified text, which is shown in Fig. 2(d).

**Category/Subcategory recognition (CR/SCR).** The (sub)category is a vital attribute for describing a product. (S)CR requires the model to produce a reliable joint representation. Following previous works [18,79], we directly append a linear layer on top of `[CLS]` to predict the label for these tasks.

**Outfit complementary item retrieval (OCIR).** OCIR aims at finding visually compatible item(s) of several given items to complete an outfit. This is a very practical task as people often buy garments that match previously selected or purchased ones. OCIR can be a helpful recommendation feature for online retailers [40,26]. To address this task, we replace the backbone of CSA-

**Table 2.** Results of cross-modal retrieval on FashionGen [52] with the protocol used in KaleidoBERT [79]. *-e2e*: without end-to-end training, *i.e.*, the image encoder is fixed. *-pt*: directly fine-tuning without multimodal pre-training

| Methods | | VSE++ [16] | ViLBERT [43] | VLBERT [55] | Image-BERT [49] | Fashion-BERT [18] | OSCAR [37] | Kaleido-BERT [79] | Ours *-e2e* | Ours *-pt* | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ITR | R@1 | 4.59 | 20.97 | 19.26 | 22.76 | 23.96 | 23.39 | 27.99 | 21.13 | 58.84 | **65.54** |
| | R@5 | 14.99 | 40.49 | 39.90 | 41.89 | 46.31 | 44.67 | 60.09 | 46.82 | 89.46 | **91.34** |
| | R@10 | 24.10 | 48.21 | 46.05 | 50.77 | 52.12 | 52.55 | 68.37 | 58.71 | 95.84 | **96.30** |
| TIR | R@1 | 4.60 | 21.12 | 22.63 | 24.78 | 26.75 | 25.10 | 33.88 | 25.83 | 57.16 | **61.88** |
| | R@5 | 16.89 | 37.23 | 36.48 | 45.20 | 46.48 | 49.14 | 60.60 | 51.54 | 84.34 | **87.32** |
| | R@10 | 28.99 | 50.11 | 48.52 | 55.90 | 55.74 | 56.68 | 68.59 | 63.53 | 91.90 | **93.22** |
| Mean | | 15.69 | 36.36 | 35.47 | 40.22 | 41.89 | 41.92 | 53.25 | 44.59 | 79.59 | **82.60** |

Net [40] with the pre-trained image encoder of FashionViL. Note that unlike all multimodal/cross-modal tasks above, only the pre-trained image encoder is used in this downstream task. We leverage this task to evaluate the performance of our image encoder under the proposed multimodal pre-training.

### 4.2 Comparative results

**Cross-modal retrieval.** We evaluate the cross-modal retrieval on the FashionGen [52] test split (not included in pre-training), including both ITR and TIR. Table 2 compares the performance of the previous V+L pre-training methods with our FashaionViL. Because previous works [18,79] are designed with a single-stream architecture, they can only be evaluated on a small retrieval set. For example, for TIR, the models are required to pick the best-matched image from only 101 images given a text query[10]. Recall (over 1K retrievals) is reported as the metric. The same setting is used for ITR. For a fair comparison, we strictly follow the same evaluation protocol, reporting the recall for 1K retrievals[11].

In Table 2, we compare our FashionViL and its two variants with existing methods. In particular, *-e2e* and *-pt* denotes our model without end-to-end training (image encoder is fixed) and multimodal pre-training respectively. We have the following observations: (1) Even with the fixed image encoder and without pre-training, FashionViL already achieves comparable results with the existing methods. This suggests that the performance of late fusion can be as effective as early-fusion for such fine-grained cross-modal retrieval. (2) When we unfreeze the image encoder for end-to-end training, we observe that $R$@1 jumps from 21.13 to 58.84, suggesting that end-to-end training is very efficient and redundant pre-processing may be unnecessary. (3) When we further utilize our proposed multimodal pre-training, our model achieves SOTA performance as in the last column of Table. 2, whose $R$@1 is more than twice of the previous SOTA.

Note that our model architecture for this task is two-stream. This means that it can be applied to large-scale retrieval, unlike the compared baselines.

[10] In the 101 images, 1 is positively paired with the text and the other 100 are randomly paired but sharing the same sub-category as the positive, increasing the difficulty.

[11] Because the authors did not release their 1K retrieval set, we report the average recall of 5 experiments with 5 randomly selected 1K retrieval sets.

**Table 3.** Results of cross-modal retrieval on FashionGen [52] with full evaluation

| ITR | | | TIR | | | Mean |
|---|---|---|---|---|---|---|
| R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| 42.88 | 71.57 | 80.55 | 51.34 | 75.42 | 84.75 | 67.75 |

**Table 4.** Results of text-guided image retrieval on FashionIQ [67]

| Image Encoder | | Fixed ResNet 152 | | | | ResNet 50 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fusion Module / Text Encoder | | CIRR-*pt* | CIRR [42] | Ours-*pt* | Ours | TIRG [62] GRU [9] | VAL [7] GRU [9] | CoSMo [32] GRU [9] | TIRG [62] BERT [55] | Ours-*pt* | Ours |
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Dress | R@10 | 14.38 | 17.45 | 20.97 | **22.66** | 23.65 | 26.28 | 24.49 | 27.17 | 28.46 | **33.47** |
| | R@50 | 34.66 | 40.41 | 42.64 | **46.60** | 49.93 | 50.25 | 51.01 | 53.25 | 54.24 | **59.94** |
| Shirt | R@10 | 13.64 | 17.53 | 17.62 | **18.74** | 21.98 | 21.69 | 18.99 | 22.28 | 22.33 | **25.17** |
| | R@50 | 33.56 | 38.31 | 41.32 | **41.56** | 46.61 | 45.53 | 43.57 | 45.58 | 46.07 | **50.39** |
| Toptee | R@10 | 16.44 | 21.64 | 21.67 | **25.29** | 27.84 | 27.43 | 25.19 | 27.84 | 29.02 | **34.98** |
| | R@50 | 38.34 | 45.38 | 46.46 | **50.28** | 55.07 | 56.25 | 54.00 | 57.11 | 57.93 | **60.79** |
| Mean | | 25.17 | 30.20 | 31.78 | **34.19** | 37.51 | 37.91 | 36.21 | 38.87 | 39.67 | **44.12** |

Therefore, we additionally report the evaluation results on the full test set (of 32K image-text pairs), *i.e.*, each query item is compared with every gallery item in the full test set. The results can be found in Table 3. We encourage the future works to also follow such a full evaluation protocol to measure the performance.

**Text-guided image retrieval.** For TGIR, we compare our FashionViL with the previous V+L pre-training methods and the task-specific methods on FashionIQ [67][12]. The results are shown in Table 4. For more comprehensive comparisons, we use two different implementations adopted by previous methods, *i.e.*, training with fixed image encoder [42] or end-to-end training [62,7,32].

We first report the results with the fixed ResNet 152 from Column 1 to Column 4 (C1-C4). CIRR adopts OSCAR [37] as the fusion module and uses the global image features as the input. We find FashionViL consistently outperforms CIRR with a relative 10%∼20% gain with or without the multimodal pre-training (C1 *vs.* C3, C2 *vs.* C4). This improvement demonstrates that the patch-level features are superior to the global features for the compositional multimodal fusion. With our proposed pre-training, the performance further improves from 31.78 to 34.19 (C3 *vs.* C4), showing our pre-training also works well on the off-the-shelf fixed image encoder.

We then report the results under the end-to-end training paradigm (C5-C10). We find that simply replacing GRU with BERT (C5 *vs.* C8) already leads to a 4% relative gain (from 23.65 to 27.17), indicating the importance of having a higher-quality text encoder. Additionally, all previous works apply a late interaction between the image embeddings and modified text embeddings with an elaborately designed fusion module, *e.g.*, TIRG [62]. We argue that an earlier fusion of the two modalities should result in an even better compositional embedding for the query purpose. Comparing C9 and C8, our FashionViL without pre-training already outperforms TIRG+BERT, indicating better query multimodal embeddings are learned in our model. Note that our text encoder and fusion encoder are shared, so FashionViL also saves more training parameters

[12] Details for the reproduction of previous methods are in the supplementary file.

**Table 5.** Results of category / subcategory recognition on FashionGen [52]

| Methods | | FashionBERT [18] | ImageBERT [49] | OSCAR [37] | KaleidoBERT [79] | Ours *-pt* | |
|---|---|---|---|---|---|---|---|
| CR | Acc | 91.25 | 90.77 | 91.79 | 95.07 | 97.07 | **97.48** |
| | Macro$\mathcal{F}$ | 70.50 | 69.90 | 72.70 | 71.40 | 84.72 | **88.60** |
| SCR | Acc | 85.27 | 80.11 | 84.23 | 88.07 | 91.45 | **92.23** |
| | Macro$\mathcal{F}$ | 62.00 | 57.50 | 59.10 | 63.60 | 78.13 | **83.02** |
| Mean | | 77.76 | 74.57 | 76.96 | 79.54 | 87.84 | **90.33** |

**Table 6.** Results of outfit complementary item retrieval on PolyvoreOutfits [60]

| Methods | | Type-aware [60] | SCE-Net [58] | CSA-Net [40] | ADDE-O [26] | CSA-Net *reproduced* | Ours *-pt* | |
|---|---|---|---|---|---|---|---|---|
| OCIR | R@10 | 3.66 | 4.41 | 5.93 | 6.18 | 2.69 | 4.38 | **5.83** |
| | R@30 | 8.26 | 9.85 | 12.31 | 13.79 | 6.29 | 10.54 | **12.61** |
| | R@50 | 11.98 | 13.87 | 17.85 | 18.60 | 9.14 | 14.77 | **17.49** |
| Mean | | 7.97 | 9.38 | 12.03 | 12.86 | 6.04 | 9.90 | **11.98** |

than TIRG+BERT. With the help of pre-training, our FashionViL achieves the new SOTA result with another significant 11.2% relative gain (C9 *vs.* C10).

**Category / Subcategory recognition.** Following KaleidoBERT [79], we evaluate CR and SCR on the FashionGen dataset [52]. The joint representation of the model architecture in Fig. 2(b) is used to predict the classification score. The results are shown in Table 5. Once again, the end-to-end learning and the well-designed fashion-specific pre-training tasks help our FashionViL outperform the two previous works by significant margins (10.4% and 3.2%, respectively). Furthermore, we also simulate a new task – multi-image subcategory recognition (M-SCR) to evaluate the performance of FashionViL with multiple input images. See more results in the supplementary file.

**Outfit complementary item retrieval.** In addition to the aforementioned multimodal and instance-level downstream tasks, we also examine FashionViL on the unimodal outfit-level task, *i.e.*, OCIR. We compare our model with the previous task-specific methods [40,26] on the Disjoint split of Polyvore Outfits [60][13]. As shown in Table 6, our multimodal pre-training benefits the performance with a 21.0% improvement, even when only the image encoder is tuned.

### 4.3 Ablation study

We analyze the effectiveness of different pre-training tasks and the sharing TE/FE strategy through ablation studies over the aforementioned five downstream tasks. The complete results are listed in Table 7. In addition to the standard metrics for each benchmark, we use the Meta-sum (sum of all scores across all the benchmarks) as a global metric.

First, we establish a baseline without any multimodal pre-training in Line 0 (L0), *i.e.*, the image/text encoder is initialized with the off-the-shelf ResNet50 or BERT, which is pre-trained in vision-only or language-only domain.

[13] We have no access to the data splits of CSA-Net, so constructed the Polyvore Outfits [60] and reproduced CSA-Net by ourselves according to the original paper [40,26].

**Table 7.** Evaluation on pre-training tasks using ITR, TIR, TGIR, SCR and OCIR as downstream tasks. Each number is the mean value of all metrics for one specific downstream task. Meta-sum stands for the summation of all numbers in each row. The three shades of grey represent the top three results when sharing TE and FE

| Pre-training Tasks | ITR | TIR | TGIR | SCR | OCIR | Meta-sum |
|---|---|---|---|---|---|---|
| (0) **None** | 62.50 | 68.09 | 39.67 | 84.79 | 9.90 | 265.04 |
| (1) **MVC** (use augmented image only) | 62.85 | 68.58 | 40.50 | 84.86 | 9.53 | 266.32 |
| (2) **MPFC** | 62.10 | 68.12 | 40.22 | 86.39 | 10.05 | 266.88 |
| (3) **MLM** (mask attribute words only) | 62.32 | 67.93 | 40.46 | 85.83 | 10.38 | 266.92 |
| (4) **MLM** | 62.15 | 67.43 | 40.29 | 86.72 | 10.38 | 266.97 |
| (5) **PAC** | 63.15 | 69.30 | 40.68 | 86.36 | 9.58 | 269.07 |
| (6) **MVC** | 63.30 | 68.32 | 40.94 | 85.99 | 10.83 | 269.38 |
| (7) **ITC** | 64.63 | 70.61 | 43.13 | 86.25 | 10.69 | 275.31 |
| (8) **ITC + MLM + MPFC** | 64.28 | 70.02 | 43.31 | 87.21 | 11.12 | 275.94 |
| (9) **ITC + MLM + MPFC + ITM** | 64.37 | 70.44 | 43.56 | 87.17 | 11.08 | 276.62 |
| (10) **ITC + MLM + MPFC + ITM + MVC** | 64.88 | 70.34 | 43.94 | 87.12 | 11.56 | 277.84 |
| (11) **ITC + MLM + MPFC + ITM + MVC + PAC** | 65.00 | 70.63 | 44.12 | 87.63 | 11.98 | 279.36 |
| (12) **ITC + MLM + MPFC + ITM + MVC + PAC** (w/o sharing TE and FE) | 64.16 | 69.15 | 42.87 | 86.22 | 11.31 | 273.71 |

Second, we validate the effectiveness of each pre-training task by their standalone performance, *i.e.*, each time we pick only one task for pre-training. We show the results of MPFC, MLM, PAC, MVC, ITC in L2, L4, L5, L6 and L7. It is clear from Table 7 that all of these pre-training tasks can benefit the downstream tasks. However, we found that a pre-training task tends to be relatively more helpful to downstream tasks of its similar type. For example, both MPFC (L2) and MLM (L4) are focusing on modeling the cross-modal interaction, thus they bring more gain to SCR but contribute relatively less to ITR and TIR. In contrast, since ITC (L7) has the same objective with ITR and TIR, it significantly boosts the cross-modal performance. As for TGIR, it requires not only high-quality compositional representation but also high-quality unimodal representations, thus each of the 5 pre-training tasks have a positive impact.

Third, we validate the effectiveness of the proposed PAC (L5) and MVC (L6). For PAC, we implement a comparative experiment: MLM only on those pre-defined pseudo-attribute words (L3). The main difference between L3 and L5 is whether the multi-label supervision is performed on each masked text token or the global representation. L3 leads to much lower performance than L5, indicating that the supervision of pseudo attributes on the global representation is a better choice. Interestingly, L3 achieves a comparable result to L4, where each word (including those other than the pseudo attributes) can also be masked. This means merely masking the fine-grained words is as effective as masking all the words uniformly, which indicates the most important text cues lie in those fine-grained concept words. We then verify the superiority of MVC. To this end, we add an ablation study that does not utilize multi-angle images (L1), *i.e.*, replacing the sampled different angle image with an augmented version of the original image. Comparing L1 and L6, we confirm that the improvement of MVC mainly comes from the contrastive learning on the images from different angles.

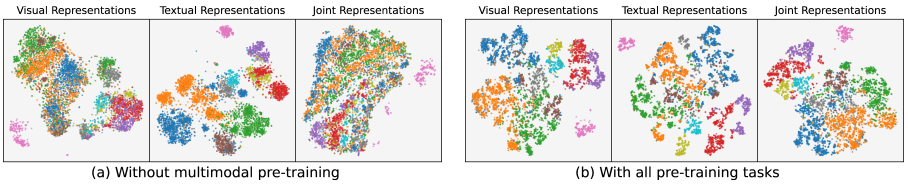


**Fig. 4.** T-sne of the learned visual/textual/joint representations from FashionViL

Next, we study the effect of different combinations of those tasks. When we add MLM and MPFC to ITC (L8), we observe a gain on Meta-sum, while the performance of ITR and TIR slightly drops. This is expected as different tasks may provide different update directions for the same parameters, which causes some tasks to overshadow the effects of others. However, minor conflicts between different tasks can be largely alleviated by employing more tasks. As shown in L9, the overall performance can be further boosted by adding ITM. The same happens when we add MVC into them (L10). When all six tasks are jointly trained (L11), we observe a significant performance gain across all benchmarks. Notably, the two new fashion-specific tasks of MVC and PAC play the most important roles to achieve the SOTA performance.

Finally, we demonstrate the superiority of sharing TE and FE. We implement a comparative model (L12) with the same pre-training tasks as L11 but using separate TE and FE. We observe a clear performance drop when breaking the parameter sharing. This indicates our modality-agnostic sharing strategy not only reduces the number of parameters but also performs far better.

### 4.4 Visualization

We visualize the representations from the image encoder, text encoder, and fusion encoder via t-SNE [45] in Fig. 4. Specifically, we feed all image-text pairs from FashionGen's test split into our model. We visualize the most popular 10 categories using different colors. We compare the t-SNE of the model without multimodal pre-training (initialized with ResNet+BERT) and the model with the full 6 pre-training tasks. We found the clusters become more discriminative when more pre-training tasks are added, indicating that FashionViL learns to acquire more fine-grained concepts. See more in the supplementary file.

## 5 Conclusions

We have introduced FashionViL, a novel end-to-end large-scale pre-training framework for V+L representation learning in the fashion domain. We proposed two effective fashion-specific pre-training tasks and introduced a novel modality-agnostic text/fusion encoder for a flexible and versatile multimodal architecture. Our FashionViL achieves new SOTA performance with superior efficiency on 5 popular fashion-related tasks.

# FashionViL: Fashion-Focused Vision-and-Language Representation Learning

—

## Supplementary Material

This supplementary material includes three sections. Sec. A describes our implementation details for the pre-training pipeline and each downstream task. Sec. B shows more experiments to demonstrate the effectiveness of FashionViL. Sec. C provides the additional visualization examples.

## A Implementation details

### A.1 Pre-training

**Image tokenizer.** As discussed in the main paper, we adopt the Masked Patch Feature Classification (MPFC) as one of our pre-training tasks. An image tokenizer is used to convert the raw pixel values into discrete labels. While previous works like BEiT [4] applied the off-the-shelf image tokenizer pre-trained on the large-scale generic image data [51], we train the image tokenizer by ourselves on the four available fashion datasets [52,70,60,23] as focusing more on the fashion domain.

Specifically, we implement a vector-quantized VAE (VQVAE) [59] with similar Encoder and Decoder architectures as VQGAN [15]. The model details are listed in Table 8. We apply the perceptual loss [30] to learn the codebook, but disregard the adversarial loss which was used in VQGAN [15] as it has been shown to be trivial for the representation learning [12]. We adopt the same training objective as PeCo [12] to learn our VQVAE with the hyper-parameters listed in Table 9. Some reconstruction samples can be found in Fig. 5.

**Pre-training.** FashionViL is end-to-end pre-trained on 6 tasks as mentioned in the main paper. Previous fashion V+L works, *i.e.* FashionBERT [18] and KaleidoBERT [79], perform all the pre-training tasks in one iteration, which is memory demanding. In this work, we follow UNITER [8] to sample one task per iteration and train it with one objective.

We implement FashionViL pre-training with MMF [54] on 4 RTX 3090 GPUs. All hyper-parameters are listed in Table 10.

**Table 8.** High-level architecture of the encoder and decoder of our VQVAE

| **Encoder** |
|---|
| $x \in \mathbb{R}^{224\times224\times3}$ |
| Conv2D $\rightarrow \mathbb{R}^{224\times224\times128}$ |
| 6× {Res Block, Res Block, Downsample Block} $\rightarrow \mathbb{R}^{7\times7\times512}$ |
| 2× {Non-local Block, Res Block} $\rightarrow \mathbb{R}^{7\times7\times512}$ |
| GroupNorm, Swish, Conv2D $\rightarrow \mathbb{R}^{7\times7\times256}$ |

| **Decoder** |
|---|
| $z_{\mathbf{q}} \in \mathbb{R}^{7\times7\times256}$ |
| Conv2D $\rightarrow \mathbb{R}^{7\times7\times512}$ |
| 2× {Res Block, Non-local Block} $\rightarrow \mathbb{R}^{7\times7\times512}$ |
| 6× {Res Block, Res Block, Upsample Block} $\rightarrow \mathbb{R}^{7\times7\times128}$ |
| GroupNorm, Swish, Conv2D $\rightarrow \mathbb{R}^{224\times224\times3}$ |

**Table 9.** Hyper-parameters for training our VQVAE

| | | |
|---|---|---|
| **Data augmentation** | RandomResizedCrop | (224, 224) |
| **Model configuration** | Codebook size | 1024 |
| | Latent feature dimension | 256 |
| | EMA decay | 0.99 |
| **Training setting** | Number of iterations | 500,000 |
| | Batch size | 32 |
| | Initial LR | 1.44e-4 |
| | Optimizer | Adam (0.5, 0.9) |
| **Hardware** | GPU | 4 x RTX 3090 |
| | Training duration | 96h |

**Table 10.** Hyper-parameters for pre-training FashionViL

| | | |
|---|---|---|
| **Image encoder** | | ResNet50 |
| **Text/Fusion encoder** | | BERT-base-uncased |
| **Text tokenizer** | Sequence length | 75 |
| | Mask probability | 15% |
| | Whole word mask | ✓ |
| **Image tokenizer** | Min masked patches | 4 |
| | Max masked patches | 8 |
| | Aspect ratio of mask | (1/3, 3) |
| **Data augmentation** | Resize | (256, 256) |
| | RandomCrop | (224, 224) |
| | RandomHorizontalFlip | ✓ |
| **Training setting** | Number of iterations | 120,000 |
| | Batch size | 256 |
| | Initial LR of TE/FE | 1e-5 |
| | Initial LR of IE | 2e-4 |
| | LR schedule | Multi-step |
| | LR steps | 45,000 and 90,000 |
| | LR decrease ratio | 0.1 |
| | Warmup iterations | 15,000 |
| | Warmup factor | 0.25 |
| | Optimizer | AdamW (0.9, 0.999) |
| | Weight decay | 1e-4 |
| **Hardware** | GPU | 4 × RTX 3090 |
| | Training duration | 28.5h |

**Table 11.** Hyper-parameters for fine-tuning FashionViL on cross-modal retrieval

| | | |
|---|---|---|
| **Image encoder** | | ResNet50 |
| **Text/Fusion encoder** | | BERT-base-uncased |
| **Text tokenizer** | Sequence length | 75 |
| **Data augmentation** | Resize | (256, 256) |
| | RandomCrop | (224, 224) |
| | RandomHorizontalFlip | ✓ |
| **Training setting** | Number of iterations | 75,120 |
| | Batch size | 64 |
| | Initial LR of TE | 1e-5 |
| | Initial LR of IE | 2e-4 |
| | LR schedule | Multi-step |
| | LR steps | 28,170 and 56,340 |
| | LR decrease ratio | 0.1 |
| | Warmup iterations | 9,390 |
| | Warmup factor | 0.25 |
| | Optimizer | AdamW (0.9, 0.999) |
| | Weight decay | 1e-4 |
| **Hardware** | GPU | 1 x RTX 3090 |
| | Training duration | 9h |

### A.2 Fine-tuning

**Cross-modal retrieval (ITR & TIR).** As ITR and TIR have the same objective as image-text contrastive learning (ITC), we directly fine-tune FashionViL with $\mathcal{L}_{\text{ITC}}$ on the FashionGen dataset [52], where the learnable temperature $\tau$ is initialized as 0.625. All hyper-parameters are listed in Table 11.

**Text-guided image retrieval (TGIR).** Previous works [42,53] found TGIR is a sensitive task (or dataset). Even a small change in the training setting can result in a quite different model performance. For a fair and stable comparison, we keep the same experimental setting for all the experiments in Table 4 in the main paper. Specifically, we removed tricks like ensemble learning and only keep the composition module implementation. For methods with lightweight text encoders (C5, C6, C7), we use CLIP embeddings [50] as the initialization of the word embeddings, which is shown to be effective in [22]. We apply batch-based classification (BBC) loss [62] for TGIR. All experiments are conducted using the hyper-parameters in Table 12.

**Category / Subcategory recognition (CR / SCR).** For CR and SCR, we directly follow the setting of KaleidoBERT [79] with the cross entropy (CE) as the loss function. All the hyper-parameters are listed in Table 13.

**Outfit complementary item retrieval (OCIR).** We follow CSA-Net [40] for the task of OCIR. We tried hard but cannot get the proposed data splits and reproduction code in CSA-Net [40]. We thus reorganize Polyvore Outfits [60] and reproduce CSA-Net by ourselves according to the paper. As a result, our results differ from the original paper, but we will release our splits and reproduction code for the convenience of future research. All the experiments implemented by us follow the same hyper-parameters listed in Table 14. Contrastive loss is applied as the training objective.

**Table 12.** Hyper-parameters for fine-tuning FashionViL on TGIR

| | | |
|---|---|---|
| **Image encoder** | | ResNet50 |
| **Text/Fusion encoder** | | BERT-base-uncased |
| **Text tokenizer** | Sequence length | 75 |
| **Data augmentation** | Resize | (256, 256) |
| | RandomCrop | (224, 224) |
| | RandomHorizontalFlip | ✓ |
| **Training setting** | Number of iterations | 44,960 |
| | Batch size | 32 |
| | Initial LR of FE | 1e-5 |
| | Initial LR of IE | 2e-4 |
| | LR schedule | Multi-step |
| | LR steps | 16,860 and 28,100 |
| | LR decrease ratio | 0.1 |
| | Warmup iterations | 2,810 |
| | Warmup factor | 0.25 |
| | Optimizer | AdamW (0.9, 0.999) |
| | Weight decay | 1e-4 |
| **Hardware** | GPU | 1 x RTX 3090 |
| | Training duration | 5.5h |

**Table 13.** Hyper-parameters for fine-tuning FashionViL on (S)CR

| | | |
|---|---|---|
| **Image encoder** | | ResNet50 |
| **Text/Fusion encoder** | | BERT-base-uncased |
| **Text tokenizer** | Sequence length | 75 |
| **Data augmentation** | Resize | (256, 256) |
| | RandomCrop | (224, 224) |
| | RandomHorizontalFlip | ✓ |
| **Training setting** | Number of iterations | 37,580 |
| | Batch size | 32 |
| | Initial LR of FE | 1e-5 |
| | Initial LR of IE | 2e-4 |
| | Optimizer | AdamW (0.9, 0.999) |
| | Weight decay | 1e-4 |
| **Hardware** | GPU | 1 x RTX 3090 |
| | Training duration | 2.5h |

**Table 14.** Hyper-parameters for fine-tuning FashionViL on OCIR

| | | |
|---|---|---|
| **Image encoder** | | ResNet50 |
| **Data augmentation** | Resize | (256, 256) |
| | RandomCrop | (224, 224) |
| | RandomHorizontalFlip | ✓ |
| **Training setting** | Number of iterations | 8,000 |
| | Batch size | 64 |
| | Initial LR of IE | 1e-4 |
| | LR schedule | Multi-step |
| | LR steps | 1,500 and 5,000 |
| | LR decrease ratio | 0.1 |
| | Warmup iterations | 1,000 |
| | Warmup factor | 0.25 |
| | Optimizer | AdamW (0.9, 0.999) |
| | Weight decay | 1e-4 |
| **Hardware** | GPU | 1 x RTX 3090 |
| | Training duration | 1.5h |

**Table 15.** Results of multi-image subcategory recognition on FashionGen [52]

| SCR w/o pt | | SCR w/ pt | | M-SCR w/o pt | | M-SCR w/ pt | |
|---|---|---|---|---|---|---|---|
| 91.45 | 78.13 | 92.33 | 83.02 | 90.33 | 72.16 | 93.39 | 84.30 |

# B Additional quantitative results

## B.1 Performance on multi-image subcategory recognition

Our model can be easily extended to support multi-image input by concatenating all image tokens together. However, there is no existing downstream task taking multiple images for direct comparison with published results, thus such experiments are omitted. We have now simulated a new one – multi-image subcategory recognition (M-SCR), which takes multiple images as input. Table 15 shows that our pre-training (pt) can yield even larger gain (Acc & Macro$\mathcal{F}$). More interestingly, SCR outperforms M-SCR w/o pre-training, but the comparison is reversed after pre-training, indicating (a) the fusion of multiple images and text is not trivial, and (b) our FashionViL is effective in the fusion task.

# C Additional qualitative results

We provide more visualization results in this section to better understand the performance of our FashionViL in a qualitative way.

## C.1 VQVAE reconstruction

We show some reconstruction results generated by our VQVAE in Figure 5. The overall quality of the reconstructed images is satisfactory with those basic semantic information (*e.g.*, the outline and color of the object) well preserved.

## C.2 Additional t-sne visualization

We provide more t-sne visualizations for FashionViL's joint representations on the fine-grained categories in Figure 6. In each column, we visualize all t-sne embeddings belonging to the same category (*e.g.*, `TOPS`) and color them according to their subcategory labels (*e.g.*, `BLOUSES` and `T-SHIRTS`). With the help of our pre-training tasks, the multimodal representations are better clustered in the latent space at both category-level and subcategory-level, which further proved the effectiveness of our pre-training.

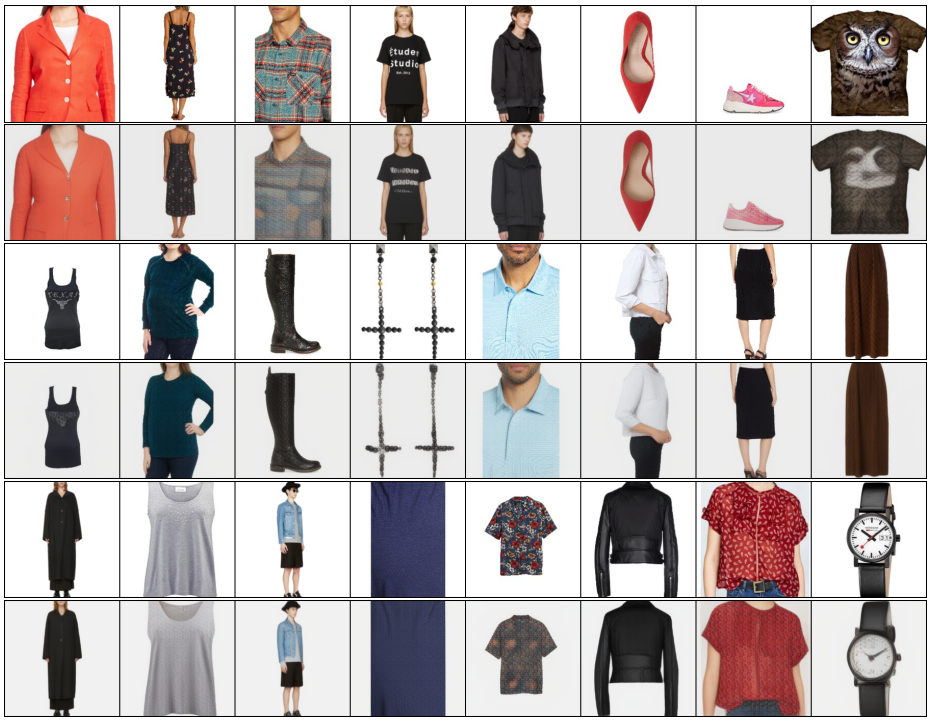

**Fig. 5.** Some reconstruction results generated by our VQVAE. Odd rows are the original images, and even rows are the reconstructed images from the previous row

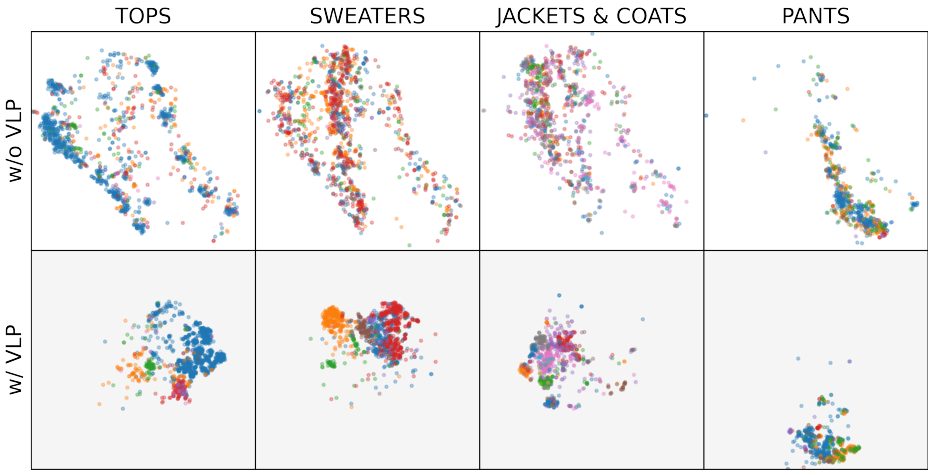


**Fig. 6.** T-sne of the multimodal representations from not pre-trained and pre-trained FashionViL. Different colors represent subcategories of the categories mentioned in each column header

## References


1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In: NeurIPS (2021) 4
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015) 3
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 5
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022) 7, 15
5. Bird, S., Klein, E., Loper, E.: Natural language processing with python: analyzing text with the natural language toolkit. https://www.nltk.org (2009) 7
6. Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D.: Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. TACL (2021) 4
7. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: CVPR (2020) 9, 11
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020) 1, 3, 4, 5, 6, 15
9. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP (2014) 11
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019) 4, 5
11. Dong, X., Zhan, X., Wu, Y., Wei, Y., Wei, X., Lu, M., Liang, X.: M5product: A multi-modal pretraining benchmark for e-commercial product downstream tasks. arXiv preprint arXiv:2109.04275 (2021) 4
12. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Peco: Perceptual codebook for bert pre-training of vision transformers. arXiv preprint arXiv:2111.12710 (2021) 7, 15
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020) 4
14. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Liu, Z., Zeng, M., et al.: An empirical study of training end-to-end vision-and-language transformers. arXiv preprint arXiv:2111.02387 (2021) 7
15. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 7, 15
16. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: BMVC (2018) 10
17. Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., et al.: Wenlan 2.0: Make ai imagine via a multimodal foundation model. arXiv preprint arXiv:2110.14378 (2021) 4
18. Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., Wang, H.: Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In: SIGIR (2020) 2, 4, 6, 9, 10, 12, 15

19. Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I., Gurevych, I.: Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. arXiv preprint arXiv:2103.11920 (2021) 4, 6, 9
20. Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., Feris, R.S.: Dialog-based interactive image retrieval. In: NeurIPS (2018) 9
21. Han, X., He, S., Zhang, L., Song, Y.Z., Xiang, T.: Uigr: Unified interactive garment retrieval. In: CVPR workshops (2022) 3, 9
22. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data. In: BMVC (2021) 17
23. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: ICCV (2017) 2, 9, 15
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 8
25. Hoe, J.T., Ng, K.W., Zhang, T., Chan, C.S., Song, Y.Z., Xiang, T.: One loss for all: Deep hashing with a single cosine similarity based learning objective. In: NeurIPS (2021) 7
26. Hou, Y., Vig, E., Donoser, M., Bazzani, L.: Learning attribute-driven disentangled representations for interactive fashion retrieval. In: ICCV (2021) 9, 12
27. Hu, X., Yin, X., Lin, K., Zhang, L., Gao, J., Wang, L., Liu, Z.: Vivo: Visual vocabulary pre-training for novel object captioning. In: AAAI (2021) 5
28. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020) 3, 4
29. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 3, 4, 9
30. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) 15
31. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML (2021) 1, 3, 4, 7
32. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: CVPR (2021) 9, 11
33. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021) 1, 4, 5, 8
34. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019) 1, 3, 4, 5, 6
35. Li, L.H., You, H., Wang, Z., Zareian, A., Chang, S.F., Chang, K.W.: Unsupervised vision-and-language pre-training without parallel images and captions. In: NAACL-HLT (2021) 4
36. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: ACL-IJCNLP (2021) 5
37. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020) 1, 4, 5, 10, 11, 12
38. Liao, L., He, X., Zhao, B., Ngo, C.W., Chua, T.S.: Interpretable multimodal retrieval for fashion products. In: ACM MM (2018) 2
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 2

40. Lin, Y.L., Tran, S., Davis, L.S.: Fashion outfit complementary item retrieval. In: CVPR (2020) 9, 10, 12, 17
41. Liu, H., Yu, T., Li, P.: Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval. In: EMNLP (2021) 4, 9
42. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV (2021) 9, 11, 17
43. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019) 1, 3, 4, 5, 10
44. Ma, Y., Jia, J., Zhou, S., Fu, J., Liu, Y., Tong, Z.: Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In: AAAI (2017) 2
45. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR (2008) 14
46. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 6
47. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019) 8
48. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015) 2
49. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020) 10, 12
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 3, 4, 5, 9, 17
51. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021) 7, 15
52. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317 (2018) 2, 4, 9, 10, 11, 12, 15, 17, 19
53. Shin, M., Cho, Y., Ko, B., Gu, G.: Rtic: Residual learning for text and image composition using graph convolutional network. arXiv preprint arXiv:2104.03015 (2021) 9, 17
54. Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D.: Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf (2020) 8, 15
55. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: pre-training of generic visual-linguistic representations. In: ICLR (2020) 1, 3, 4, 5, 8, 10, 11
56. Sun, S., Chen, Y.C., Li, L., Wang, S., Fang, Y., Liu, J.: Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In: NAACL-HLT (2021) 3, 4, 6, 9
57. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: EMNLP-IJCNLP (2019) 1, 3, 4
58. Tan, R., Vasileva, M.I., Saenko, K., Plummer, B.A.: Learning similarity conditions without explicit supervision. In: ICCV (2019) 12
59. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017) 7, 15

60. Vasileva, M.I., Plummer, B.A., Dusad, K., Rajpal, S., Kumar, R., Forsyth, D.: Learning type-aware embeddings for fashion compatibility. In: ECCV (2018) 2, 4, 9, 12, 15, 17
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 4
62. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing Text and Image for Image Retrieval - an Empirical Odyssey. In: CVPR (2019) 3, 6, 9, 11, 17
63. Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., Wang, L.: Ufo: A unified transformer for vision-language representation learning. arXiv preprint arXiv:2111.10023 (2021) 4
64. Wang, W., Bao, H., Dong, L., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358 (2021) 4
65. Wang, Z., Wang, W., Zhu, H., Liu, M., Qin, B., Wei, F.: Distilled dual-encoder model for vision-language understanding. arXiv preprint arXiv:2112.08723 (2021) 4, 9
66. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. In: ICLR (2021) 1, 5
67. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion iq: A new dataset towards retrieving images by natural language feedback. In: CVPR (2021) 2, 3, 4, 6, 9, 11
68. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016) 5
69. Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F.: E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In: ACL-IJCNLP (2021) 4
70. Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C.H., Tan, J., Xie, D., Wang, J., Wang, X.: Fashion captioning: Towards generating accurate descriptions with semantic rewards. In: ECCV (2020) 2, 9, 15
71. You, H., Zhou, L., Xiao, B., Codella, N.C., Cheng, Y., Xu, R., Chang, S.F., Yuan, L.: Ma-clip: Towards modality-agnostic contrastive language-image pre-training. OpenReview (2021) 4
72. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016) 3
73. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019) 3
74. Zhang, L., Wu, H., Chen, Q., Deng, Y., Li, Z., Kong, D., Cao, Z., Siebert, J., Han, Y.: Vldeformer: Learning visual-semantic embeddings by vision-language transformer decomposing. arXiv preprint arXiv:2110.11338 (2021) 9
75. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR (2021) 4
76. Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., Tang, J., Zhou, J., Yang, H.: Ufc-bert: Unifying multi-modal controls for conditional image synthesis. In: NeurIPS (2021) 4
77. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: AAAI (2020) 5
78. Zhu, Y., Zhao, H., Zhang, W., Ye, G., Chen, H., Zhang, N., Chen, H.: Knowledge perceived multi-modal pretraining in e-commerce. In: ACM MM (2021) 4

79. Zhuge, M., Gao, D., Fan, D.P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-bert: Vision-language pre-training on fashion domain. In: CVPR (2021) 2, 4, 9, 10, 12, 15, 17