FaD-VLP: Fashion Vision-and-Language Pre-training towards Unified Retrieval and Captioning

Suvir Mirchandani
Stanford University
suvir@cs.stanford.edu

Licheng Yu
Meta AI
lichengyu@fb.com

Mengjiao Wang
Meta AI
mengjiaow@fb.com

Animesh Sinha
Meta AI
animeshsinha@fb.com

Wenwen Jiang
Meta AI
wenwenj@fb.com

Tao Xiang
Meta AI
txiang@fb.com

Ning Zhang
Meta AI
ningzhang@fb.com

Abstract

Multimodal tasks in the fashion domain have significant potential for e-commerce, but involve challenging vision-and-language learning problems—*e.g.*, retrieving a fashion item given a reference image plus text feedback from a user. Prior work on multimodal fashion tasks have been limited by the training data for individual benchmarks, or have leveraged generic vision-and-language pre-training methods but have not fully taken advantage of the specialty of fashion data. Additionally, these works have mainly focused on multimodal understanding tasks. To address these gaps, we make two key contributions. First, we propose a novel fashion-specific pre-training framework based on weakly-supervised triplets, constructed from paired fashion image-text data. We show the triplet-based tasks are an effective addition to standard multimodal pre-training tasks. Second, we propose a flexible decoder-based model architecture capable of both fashion retrieval and captioning tasks. Together, our model design and pre-training approach achieves high marks on a diverse set of fashion tasks, including cross-modal retrieval, image retrieval with text feedback, image captioning, relative image captioning, and multimodal categorization.

1 Introduction

Artificial intelligence has taken the fashion industry by storm in recent years. Significant advances have been made in tasks like recommendation (McAuley et al., 2015; Deldjoo et al., 2022) and virtual try-on (Han et al., 2018; Yang et al., 2022). In addition to these primarily visual tasks, *multimodal* tasks are of particular interest in fashion for e-commerce applications: for example, text-to-image retrieval

Figure 1: We present FaD-VLP, a flexible architecture and pre-training method that can support retrieval-based and captioning-based tasks in the fashion domain.

enables a shopper to identify a desired clothing item via a language query (Zhuge et al., 2021).

A key opportunity to enhance customers' shopping experiences is in the development of *interactive* multimodal shopping assistants, whereby a user could converse with a system to identify a desired product (Yuan and Lam, 2021; Han et al., 2022). As in Figure 1, a smart assistant is desired to perform multiple diverse tasks, *e.g.*, cross-modal retrieval, image retrieval with text feedback, multimodal categorization, image captioning, and relative image captioning. Among them, perhaps the most notable task in fashion is image retrieval with text feedback, where the goal is to retrieve a target image given a reference image coupled with a user's language feedback (*e.g.*, *show me a similar shirt in light blue with no print*) (Wu et al., 2021; Lee et al., 2021; Kim et al., 2021). In addition to retrieval-based tasks, a central capability of conversational shopping assistants is in captioning-based tasks: describing items in detail (Yang et al., 2020) or the differences among them. However, existing works on image retrieval with text feedback have almost exclusively studied that task in isolation, focusing on specialized architectures and fusion methods, with data limited by particular benchmarks (Lee et al., 2021; Kim et al., 2021).

To train a model that can perform well on several fashion-specific multimodal use cases, we observe an opportunity in the vast availability of multimodal fashion data on e-commerce platforms. While vision-language pre-trained (VLP) models have been highly successful for the general domain (Lu et al., 2019; Li et al., 2020; Su et al., 2020), prior work has suggested that general VLP models are helpful but sub-optimal for the fashion domain (Zhuge et al., 2021; Liu et al., 2021; Goenka et al., 2022). Fashion images represent a domain shift from the pre-training data (Liu et al., 2021), and fashion tasks often require fine-grained representations rather than coarse representations from general VLP models (Zhuge et al., 2021).

To this end, we propose a domain-specific fashion pre-training procedure that takes advantage of fashion image-text data from multiple fashion catalogues. Our approach is inspired by the way that users might shop, via *comparisons*: a user may first identify a product, express a desired change in language, and then look for a new product that better matches their preferences. Given that data in this triplet form—reference product, modification, target product—is not nearly as common as the paired image-text data, we propose a lightweight method for constructing weakly-supervised pseudo-triplet data from image-text pairs. Additionally, we propose a unified, decoder-based model architecture for both retrieval-based and captioning-based fashion tasks. Together, we refer to our architecture and pre-training approach as FaD-VLP: Fashion Decoder with Vision-and-Language Pre-training.

To summarize, we make the following contributions. We propose a unified architecture for retrieval- and captioning-based fashion tasks (Section 3.1) and a fashion pre-training framework, including 2 novel pre-training tasks based on weakly-supervised pseudo-triplets (Section 3.2). Our approach achieves high marks on 7 downstream fashion tasks: image-to-text retrieval, text-to-image retrieval, image retrieval with text feedback, image captioning, relative image captioning, category recognition, and subcategory recognition (Sections 4 and 5.1). Finally, we conduct a thorough ablation study to analyze the effects of our pre-training procedure (Section 5.2).

2 Related Work

The success of the Transformer architecture (Vaswani et al., 2017) in NLP (Devlin et al., 2019) and CV (Dosovitskiy et al., 2021) has been accompanied by a substantial body of work in using transformers for vision-and-language pre-training (VLP) in the generic domain (Li et al., 2019; Chen et al., 2020; Li et al., 2021a; Radford et al., 2021a). Recent works have begun to focus on the fashion domain specifically (Yu et al., 2022b; Gao et al., 2020; Zhuge et al., 2021; Zhu et al., 2021; Dong et al., 2021; Zhang et al., 2021). VLP works mainly differ along two axes: model architecture and pre-training objectives.

Model Architecture. There are two typical classes of multimodal architectures: (i) single-stream early fusion (Li et al., 2019; Su et al., 2020; Chen et al., 2020; Li et al., 2020), and (ii) two-stream late fusion (Tan and Bansal, 2019; Lu et al., 2019; Jia et al., 2021; Radford et al., 2021a; Fei et al., 2021). The choice of single-stream vs dual-stream is usually influenced by the downstream tasks; *e.g.*, cross-modal retrieval is most practical with late fusion architectures which can have faster inference. Most existing models use encoder-style modules for both image and text, focusing on multimodal understanding tasks (*e.g.*, image-text retrieval, multimodal classification, etc.) In this work, we propose a flexible decoder-based model architecture, which embraces the advantage of both early and late fusion mechanisms, and is capable of not only multimodal understanding tasks, but also captioning tasks (*e.g.*, image captioning and relative image captioning).

Pre-training Objectives. Several pre-training tasks have been proposed for VLP. Masked Language Modeling, Image-Text Matching, and Masked Image Modeling are some of the most popular tasks (Devlin et al., 2019; Li et al., 2019; Lu et al., 2019; Su et al., 2020; Chen et al., 2020). Others include object tag prediction (Li et al., 2020), caption generation (Zhou et al., 2020; Wang et al., 2022), and image-text contrastive learning (Li et al., 2021a; Radford et al., 2021a; Li et al., 2021b). However, these tasks do not specifically take advantage of the unique aspects of fashion data. For example, there are more structured attributes in fashion captions, which entitles people to naturally do comparisons choosing their desired shopping items, *e.g.*, "I want a blue version of this t-shirt."

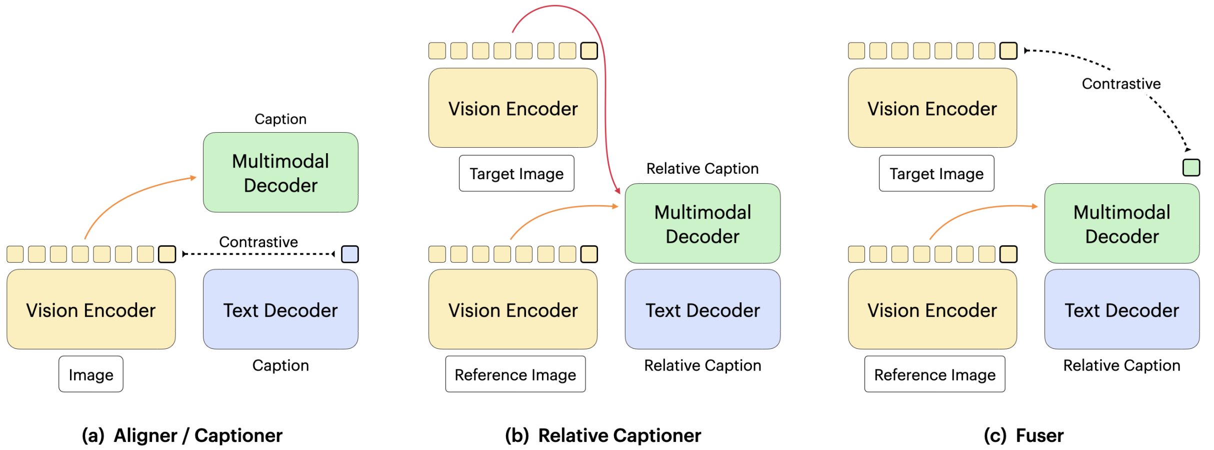

Figure 2: Our proposed FaD-VLP architecture consists of an image encoder, a text decoder, and a multimodal decoder, with three configurations that conform to various retrieval and captioning tasks. Shared colors indicate shared parameters, curved arrows represent cross attention, and tokens with a bold border denote pooled representations.

Inspired by this, we propose that weak triplet-based comparison is used as the basis for another effective pre-training task.

3 Method

We introduce FaD-VLP, our architecture and pre-training method for fashion tasks. We first detail our architecture design (Figure 2), which unifies several retrieval and captioning settings. We then describe our pre-training approach.

3.1 Model Overview

To motivate our model architecture, we enumerate three desired properties:

i. *Dual Image & Text Encoders.* As referenced in Section 2, two-stream / dual-encoder architectures are more efficient for cross-modal retrieval than single-stream architectures. With dual encoders, candidate embeddings can be retrieved using a lightweight similarity function (*e.g.*, dot product) with a particular query embedding.

ii. *Dual Multimodal & Text Encoders.* Key to our pre-training procedure is the alignment of multimodal representations with image representations. This setup is useful for the downstream task of image retrieval with text feedback: a target image is retrieved given an image with text feedback. We desire an architecture that is dual-stream with respect to a hybrid-modal input (image and text) and another image.

iii. *Multimodal Decoder for Text Generation.* For captioning tasks, we need to generate text given image input. Thus, we desire that the architecture contains a multimodal decoder.

To satisfy (i) and (iii), prior work (Li et al., 2022) has used a mixture of unimodal encoders and encoder-decoders; more recently, Yu et al. (2022a) demonstrated the effectiveness of using single decoder-based model; a decoder can be used for generation, but can also provide global representations given a whole sequence.

Building upon this result, our architecture is decoder-based, and consists of three modules: a visual encoder $\mathcal{V}$, a text decoder $\mathcal{T}$, and a multimodal decoder $\mathcal{M}$. For $\mathcal{V}$, we use a convolutional network. We obtain image token representations from the intermediate outputs of the convolutional network (*i.e.*, the output of layers 3 and 4 in a ResNet-50, following Kim et al. (2021)). We obtain pooled representations from $\mathcal{V}$ using average pooling over the final feature map. We use a multi-layer transformer architecture for $\mathcal{T}$ and $\mathcal{M}$. Each layer consists of a causal multi-headed self-attention module followed by a feed-forward network and layer normalization. For $\mathcal{M}$, we also include a cross-attention layer between the image representation and the outputs of the causal self-attention. We extract pooled representations from $\mathcal{T}$ or $\mathcal{M}$ using the output of corresponding to an `[EOS]` token (which has attended to all prior tokens).

Our architecture has the following modes:

(a) **Aligner / Captioner.** This mode can align cross-modal representations or caption an image. For alignment, we input a caption to $\mathcal{T}$ and an image to $\mathcal{V}$, extracting the pooled representations. For captioning, we pass the outputs of $\mathcal{T}$ to $\mathcal{M}$ and condition $\mathcal{M}$ on the image via cross attention.

(b) **Relative Captioner.** In this mode, we can input a text (*e.g.*, a relative caption comparing two images) into $\mathcal{T}$ and two image representations into $\mathcal{M}$. We use a gated, second cross-attention mechanism for the second image input. The second cross attention module immediately follows the first. The multimodal decoder can be trained to generate a relative caption conditioned on images.

(c) **Fuser.** This mode can fuse an image and text, and align the results with another image. We input a relative caption to $\mathcal{T}$, a reference image representation to $\mathcal{M}$ via cross attention, and a target image into $\mathcal{V}$. The pooled representation of the target can be aligned with the fused representation from $\mathcal{M}$.

We prepend text inputs with a special token indicating which of the three modes the architecture is operating in. Following prior VLP work (Lu et al., 2019; Li et al., 2020; Su et al., 2020; Li et al., 2021a, 2022), we initialize our model with BERT encoder weights (Devlin et al., 2019). As BERT does not have cross-attention parameters, they are learned from scratch.

3.2 Pre-training Objectives

Image-text pairs are a common choice for pre-training vision-language systems; these pairs can be mined or repurposed from existing datasets, such as captioning datasets. As described in Section 4.1, we repurpose a set of fashion datasets to form a pre-training dataset $\mathcal{D}$ consisting of fashion image-text pairs. We use $\mathcal{D}$ for domain-specific pre-training. This section describes our pre-training tasks, which include two tasks based on paired data (Section 3.2.1), as well as two novel tasks based on triplet data (Section 3.2.2). Implementation details are given in Appendix A.

3.2.1 Pre-training with Pairs

Cross-Modal Contrastive Learning (CMC). To align the representations of images with their corresponding texts, we use a cross-modal contrastive loss. Given an image I and a text T, we extract the pooled feature vectors $\mathbf{i}$ and $\mathbf{t}$ from the visual encoder $\mathcal{V}(I)$ and the text encoder $\mathcal{T}(T)$. We then project $\mathbf{i}$ and $\mathbf{t}$ to a normalized lower-dimensional joint embedding space using two learned linear transformations, f and g. We measure the similarity between $\mathbf{i}$ and $\mathbf{t}$ as

$$\kappa(\mathbf{i}, \mathbf{t}) = f(\mathbf{i})^T g(\mathbf{t}).$$

We can push the embeddings of matched images and texts together according to this similarity metric, and unmatched embeddings further apart, by applying the following bidirectional InfoNCE loss (van den Oord et al., 2018; Zhang et al., 2020):

$$\mathcal{L}_{\text{CMC}} = -\frac{1}{B}\sum_{j=1}^{B}\Big(\log\frac{\exp(\kappa(\mathbf{i}^{(j)}, \mathbf{t}^{(j)}))}{\sum_{k=1}^{B}\exp(\kappa(\mathbf{i}^{(j)}, \mathbf{t}^{(k)}))} + \log\frac{\exp(\kappa(\mathbf{i}^{(j)}, \mathbf{t}^{(j)}))}{\sum_{k=1}^{B}\exp(\kappa(\mathbf{i}^{(k)}), \mathbf{t}^{(j)}))}\Big).$$

where B is a sample of indices from our pretraining dataset $\mathcal{D}$. We use the Aligner / Captioner mode of our architecture for this task.

Image Caption Language Modeling (ICLM). In addition to alignment, we encourage the model to gain image-grounded text generation capabilities. We use a language modeling loss that maximizes the conditional probability of a caption $T^{(j)}$ given an image $I^{(j)}$:

$$\mathcal{L}_{\text{ICLM}} = -\frac{1}{B}\sum_{j=1}^{B}\sum_{k=1}^{|T^{(j)}|}\log P_{\mathcal{M}}(T_k^{(j)} \mid T_{<k}^{(j)}, I^{(j)})$$

where $T_k^{(j)}$ refers to the kth token, and $T_{<k}^{(j)}$ refers to the context history of tokens. As with the CMC task, we use the Aligner / Captioner mode of our architecture.

3.2.2 Pre-training with Triplets

Taking inspiration from the way users utilize *comparisons* when browsing products (e.g., looking at a product, having a desired change in mind, and identifying a new product), we hypothesize that (image, text, image) *triplets* can be used to build additional multimodal capabilities into the model beyond cross-modal tasks like CMC and ICLM.

Below, we describe two pre-training objectives that assume access to triplets $(I_{ref}, T_{rel}, I_{tgt})$, where I_{tgt} is a target image and T_{rel} is a relative caption describing the difference between I_{tgt} and some reference image I_{ref}. Note we only have access to image-text pairs in $\mathcal{D}$; we describe how we construct these triplets from pairs in Section 3.3.

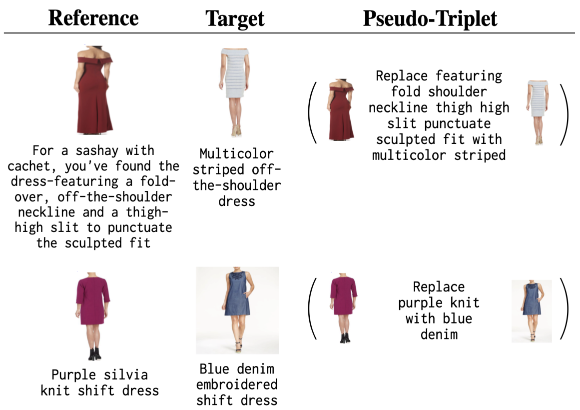

Figure 3: Two examples of pseudo-triplets. From a reference image-text pair, we find a target image-text pair and then construct a relative caption for the two images.

Hybrid-Modal Contrastive Learning (HMC). We propose a pre-training task, using the Fuser mode of the architecture, that aligns the fused representation of I_{ref} and T_{rel} with the unimodal representation of I_{tgt}. Intuitively, this would imbue the model with the ability to modify I_{ref} in embedding space, as specified by T_{rel}.

We project the pooled features $\mathbf{m}$ and $\mathbf{i_t}$ of the multimodal and target image embeddings $\mathcal{M}(I_{ref}, T_{rel})$ and $\mathcal{V}(I_{tgt})$ using two learned linear transformations, h and f respectively. We measure the similarity between $\mathbf{m}$ and $\mathbf{i_t}$ as

$$\kappa'(\mathbf{m}, \mathbf{i_t}) = h(\mathbf{m})^T f(\mathbf{i_t}).$$

We then apply the following contrastive loss:

$$\mathcal{L}_{\text{HMC}} = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp(\kappa'(\mathbf{m}^{(j)}, \mathbf{i_t}^{(j)}))}{\sum_{k=1}^{B} \exp(\kappa'(\mathbf{m}^{(j)}, \mathbf{i_t}^{(k)}))}.$$

Relative Caption Language Modeling (RCLM). We additionally apply a language modeling loss, using the Relative Captioner mode of our architecture, that maximizes the conditional probability of a relative caption given a reference and a target image:

$$\mathcal{L}_{\text{RCLM}} = -\frac{1}{B}\sum_{j=1}^{B}\sum_{k=1}^{|T_{rel}^{(j)}|} \log P_{\mathcal{M}}(T_{rel,k}^{(j)} \mid T_{rel,<k}^{(j)}, I_{ref}, I_{tgt}).$$

3.3 Constructing Pseudo-Triplets

Collecting additional pre-training data for the triplet-based tasks would be expensive, thus we

Name	# Pairs
FACAD (Yang et al., 2020)	888K
Fashion-Gen (Rostamzadeh et al., 2018)	260K
Fashion200K (Han et al., 2017)	172K
Shopping100K (Ak et al., 2018)	100K
DeepFashion (Liu et al., 2016)	26K

Table 1: Breakdown of the pre-training dataset.

aim to construct triplets purely from the paired-image data. We propose a simple approach for generating weakly supervised pseudo-triplets.

We iterate through each image-text pair in $\mathcal{D}$, treating it as a reference (I_{ref}, T_{ref}). As in Figure 3, for each reference, we select a target image-text pair (I_{tgt}, T_{tgt}), then construct a relative caption as a function rel of the two captions, T_{ref} and T_{tgt}. We use $(I_{ref}, rel(T_{ref}, T_{tgt}), I_{tgt})$ as our pseudo-triplet.

Selecting Targets. Intuitively, we want I_{ref} and I_{tgt} to be related (such that there are some shared properties of the reference and the target item) but not identical (so the relative caption would be meaningful). To do this, for a given reference (I_{ref}, T_{ref}), we find

$$\text{argmin}_{j \in S} \Delta((I_{ref}, T_{ref}), (I^{(j)}, T^{(j)}))$$

where Δ is a similarity metric and S is a set of indices sampled from the indices in $\mathcal{D}$, and set $(I_{tgt}, T_{tgt}) = (I^{(j)}, T^{(j)})$. We use the following metric:

$$\begin{aligned}\Delta((I, T), (I', T')) = &- \lambda_1 \cdot \kappa(\phi_I(I), \phi_I(I'))) \\ &- \lambda_2 \cdot \kappa(\phi_I(T), \phi_T(T')) \\ &+ \lambda_3 \cdot d(T, T'),\end{aligned}$$

where ϕ_I and ϕ_T are feature extractors for image and text, d is the token-wise Hamming distance (for nouns, adjectives, and participles), κ is cosine similarity, and λ_1, λ_2, and λ_3 are scalar weights. We use frozen feature extractors: ResNet-50 (He et al., 2016) for ϕ_I and all-MiniLM-L12-v2 (Reimers and Gurevych, 2019) for ϕ_T. The intuition behind our choice for Δ is that pairings should be visually similar, and texts should be semantically similar and have token overlap.

Constructing Relative Captions. To construct a relative caption T_{rel}, we design a function rel of the reference and target captions T_{ref} and T_{tgt}. Our goal is to describe the difference between both images such that HMC and RCLM are empirically helpful for our downstream tasks. By training $\mathcal{M}$ to

Method	Image-to-Text			Text-to-Image			Average
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
VSE (Kiros et al., 2014)	4.01	11.03	22.14	4.35	12.76	20.91	12.53
VSE++ (Faghri et al., 2018)	4.59	14.99	24.10	4.60	16.89	28.99	15.69
SCAN (Lee et al., 2018)	4.59	16.50	26.60	4.30	13.00	22.30	14.55
PFAN (Wang et al., 2019)	4.29	14.90	24.20	6.20	20.79	31.52	16.98
ViLBERT (Lu et al., 2019)	20.97	40.49	48.21	21.12	37.23	50.11	36.35
VLBERT (Su et al., 2020)	19.26	39.90	46.05	22.63	36.48	48.52	35.47
FashionBERT (Gao et al., 2020)	23.96	46.31	52.12	26.75	46.48	55.74	41.89
ImageBERT (Qi et al., 2020)	22.76	41.89	50.77	24.78	45.20	55.90	40.22
OSCAR (Li et al., 2020)	23.39	44.67	52.55	25.10	49.14	56.68	41.92
Kaleido-BERT (Zhuge et al., 2021)	27.99	60.09	68.37	33.88	60.60	68.59	53.25
FaD-VLP (Ours, *w/o* CLIP init.)	59.88	83.64	91.52	55.24	83.18	91.30	77.46
FaD-VLP (Ours)	**64.30**	**86.78**	**93.48**	**58.66**	**84.92**	**91.58**	**79.95**

Table 2: Results for Image-Text / Text-Image Retrieval (ITR / TIR) on Fashion-Gen (Rostamzadeh et al., 2018).

Method	Dress		Shirt		Toptee		Average
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	
TIRG (Vo et al., 2019)	14.13	34.61	13.10	30.91	14.79	34.37	23.66
CIRPLANT (Liu et al., 2021)	17.45	40.41	17.53	38.81	21.64	45.38	30.20
CoSMo (Lee et al., 2021)	21.39	44.45	16.90	37.49	21.32	46.02	31.25
FashionVLP (Goenka et al., 2022)	26.77	53.20	22.67	46.22	28.51	57.47	39.14
DCNet (Kim et al., 2021)	28.95	56.07	23.95	47.30	30.44	58.29	40.84
FaD-VLP (Ours, *w/o* CLIP init.)	29.15	55.97	23.45	46.61	30.85	57.57	40.60
FaD-VLP (Ours)	32.08	57.96	25.22	49.71	33.20	60.84	43.17
Prog. Lrn. - RN-50 (Zhao et al., 2022)†	29.00	53.94	35.43	58.88	39.16	64.56	46.83
Prog. Lrn. - ViT-B/32 (Zhao et al., 2022)†	**33.60**	**58.90**	**39.45**	**61.78**	**43.96**	**68.33**	**51.01**

Table 3: Results for Image Retrieval with Text Feedback (IRTF) on Fashion IQ (Wu et al., 2021). † refers to concurrent work on fashion vision-language pre-training.

match the fused representation of an image I_{ref} and relative caption T_{rel}, we would be adding fusion capability to $\mathcal{M}$ (with no extra data) and potentially getting stronger representations of image and text.

We use a lightweight procedure for *rel*. We extract the first sentence from T_{ref} and T_{tgt}, and perform part-of-speech tagging. We filter out all tokens that do not function as nouns and adjectives in the caption, and all tokens that occur less than 500 times over the the entire dataset. This leaves us with a noisy list of "attributes" for each image.

Next, we remove tokens that overlap between reference tokens and target tokens. This step removes some of the redundancy (at the token level) between the reference and target captions, as relative captions would not need to mention aspects in the target image that are unchanged from the reference. We use the remaining reference tokens and target tokens to fill a randomly selected template of the form "change <ref_tokens> to <tgt_tokens>," "<tgt_tokens> instead of <ref_tokens>," etc. While our goal is not to produce fully grammatical or complete relative captions, we found most constructed sentences are meaningful, as in Figure 3. More importantly, they are effective at improving pre-training.

3.3.1 Bootstrapping with Relative Caption Generations

The RCLM pre-training task enables the model to produce relative captions given two images. This gives us the ability to sample new relative captions at training time as a form of data augmentation. We use nucleus sampling (with $p = 0.9$) to generate more diverse relative captions. Functionally, this step connects the Relative Captioner mode and the Fuser mode of our architecture.

4 Experiments

In this section, we provide the details of our pre-training dataset and downstream fashion tasks. Implementation details are given in Appendix A.

4.1 Datasets

Our pre-training dataset is comprised of image-text pairs from five fashion datasets: FACAD (Yang et al., 2020), Fashion-Gen (Rostamzadeh et al., 2018), Fashion200K (Han et al., 2017), Shopping100k (Ak et al., 2018), and DeepFashion (Liu et al., 2016). Each of these datasets contains data

Method	Category			Subcategory		
	Acc.	F1	Avg.	Acc.	F1	Avg.
FashionBERT (Gao et al., 2020)	91.25	70.5	80.9	85.27	62.0	77.9
ImageBERT (Qi et al., 2020)	90.77	69.9	80.3	80.11	57.5	75.0
OSCAR (Li et al., 2020)	91.79	72.7	82.2	84.23	59.1	78.5
Kaleido-BERT (Zhuge et al., 2021)	95.07	71.4	83.2	88.07	63.6	79.7
FaD-VLP (Ours) [*w/o* CLIP init.]	97.90	89.3	93.6	**93.53**	**83.2**	91.4
FaD-VLP (Ours)	**98.32**	**89.5**	**93.9**	93.37	**83.2**	**91.5**

Table 4: Results for Category / Subcategory Recognition (CR / SR) on Fashion-Gen (Wu et al., 2021).

Method	B	M	R	C	Sum
FashionBERT (Gao et al., 2020)	3.3	9.8	29.7	0.3	45.8
OSCAR (Li et al., 2020)	4.5	10.9	30.1	0.31	48.6
Kaleido-BERT (Zhuge et al., 2021)	5.7	12.8	32.9	0.33	54.7
FaD-VLP (Ours) [*w/o* CLIP init., PT]	29.3	28.6	54.5	1.41	126.6
FaD-VLP (Ours) [*w/o* PT]	31.0	29.6	55.8	1.55	131.9
FaD-VLP (Ours)	**31.1**	**29.7**	**56.0**	**1.56**	**132.4**

Table 5: Results for Image Captioning (IC) on Fashion-Gen (Rostamzadeh et al., 2018).

Method	B	M	R	C	Sum
Decoder Baseline					76.5
w/o CLIP init, Triplet PT, Pair PT	13.3	17.4	39.0	0.71	76.7
w/o Triplet PT, Pair PT	13.4	17.6	39.2	0.73	77.5
w/o Triplet PT	13.8	18.4	39.6	0.78	79.5
FaD-VLP (Ours)	**14.5**	**18.6**	**40.8**	**0.8**	**82.0**

Table 6: Results for Relative Image Captioning (RIC) on Fashion-Gen (Wu et al., 2021).

sourced from fashion catalogues. In total, our pre-training dataset consists of 1.4M image-text pairs, with the breakdown as listed in Table 1. Further description is given in Appendix A.

Following (Gao et al., 2020) and (Zhuge et al., 2021), we use the Fashion-Gen dataset for our cross-modal retrieval, captioning, and multimodal categorization tasks. We use Fashion IQ (Wu et al., 2021) for our image retrieval with text feedback task and our relative captioning task. Fashion IQ contains 18K (reference image, text feedback, target image) training triplets and 6016 validation triplets over three categories: Dress, Shirt, and Toptee. Each (reference image, target image) pair is human-annotated with two relative captions, which are concatenated together (Wu et al., 2021).

4.2 Downstream Tasks

Our 7 downstream tasks are defined as follows.
Image-to-Text Retrieval (ITR) and Text-to-Image Retrieval (TIR). Given a gallery $\Theta_{I,T}$ of fashion image-text pairs (I, T), and a query image I_q (or query text T_q), retrieve the corresponding text T_q (or image I_q). We train these tasks with the CMC loss.
Image Retrieval with Text Feedback (IRTF). Given a gallery Θ_I of fashion images I, a reference image I_{ref}, and a relative caption T_{rel}, retrieve the target image I_{tgt} that most closely applies T_{rel} to I_{ref}. We train this task with the HMC loss.
Category Recognition (CR) and Subcategory Recognition (SR). Given an image I_q and a list of categories (or subcategories) C, predict the category (or subcategory) c into which I_q falls. Example categories include {`SNEAKERS, JEANS`} and and example subcategories include {`SILKS & CASHMERES, HEELED SANDALS`}. We train these tasks with categorical cross entropy loss. We train with the ICLM loss.
Image Captioning (IC). Given a fashion image I_q, generate a descriptive caption T_q.
Relative Image Captioning (RIC). Given a reference image I_{ref} and a target image I_{tgt}, generate a text T_{rel} describing I_{tgt} relative to I_{ref}. We train this task with the RCLM loss.

Evaluation Metrics For ITR, TIR, and IRTF, we follow prior work (Kim et al., 2021; Zhuge et al., 2021) and evaluate the retrieval performance using Recall@K, or the percentage of queries for which the correct target is retrieved within the top K results. For CR and SR, we follow (Zhuge et al., 2021) and report accuracy and the macro-F1 score over all categories (or subcategories). For IC and RIC, we report BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), abbreviated B, M, R, and C respectively.

5 Results & Discussion

This section describes our results on our downstream tasks, and analyzes our method through an ablation study.

	CLIP	CMC	ICLM	HMC	RCLM	Boot	Sum	IRTF	ITR	TIR	CR	SR	IC	RIC
(1)							528.73	31.93	54.97	56.97	93.67	87.86	126.59	76.74
(2)		✓	✓				546.09	38.00	58.18	61.14	94.50	89.38	125.52	79.37
(3)		✓	✓	✓			546.48	39.83	58.38	61.48	93.03	88.73	125.39	79.64
(4)		✓	✓	✓	✓		551.51	39.96	59.26	62.04	93.95	88.73	125.70	81.87
(5)		✓	✓	✓	✓	✓	552.07	40.60	58.82	61.83	93.58	89.45	126.07	81.72
(6)	✓						546.62	35.62	58.55	60.58	93.74	88.77	131.87	77.49
(7)	✓	✓					551.86	37.20	59.68	63.00	93.20	88.20	131.97	78.61
(8)	✓	✓	✓				562.12	40.66	61.18	64.42	94.25	89.29	132.78	79.54
(9)	✓	✓	✓	✓			564.38	42.20	62.34	66.37	93.83	88.64	132.05	78.95
(10)	✓	✓	✓	✓	✓		566.17	42.52	61.03	65.48	93.93	88.55	132.54	82.12
(11)	✓	✓	✓	✓	✓	✓	569.81	43.17	62.87	67.05	94.05	88.30	132.39	81.98

Table 7: Ablation on our pre-training objectives, with the aggregation metrics for each of our 7 downstream tasks and a meta-sum. CLIP refers to CLIP initialization; CMC, ICLM, HMC, and RCLM are pre-training tasks; Boot refers to bootstrapping.

5.1 Comparison with SOTA Models

We compare FaD-VLP to existing work on our set of 7 downstream tasks: ITR/TIR (Table 2), IRTF (Table 3), CR/SR (Table 4), IC (Table 5), and RIC (Table 6). We find that FaD-VLP sees improvement on all evaluations. For ITR and TIR, FaD-VLP outperforms prior methods by a large margin, including methods with generic VLP (Li et al., 2020). We show results of FaD-VLP with and without CLIP visual encoder initialization, indicating that the gain from fashion domain-specific pre-training is amplified when used on top of generic VLP. We see a similar trend for IRTF, where we see a gain of +2.5 on the Average metric when generic and domain-specific pre-training are coupled. FaD-VLP is competitive with state-of-the-art models that use a comparable ResNet-50 visual encoder (Kim et al., 2021; Zhao et al., 2022). Zhao et al. (2022) is a concurrent work on multi-stage vision-language pre-training that sees gains on IRTF from fashion-specific pre-training on top of a CLIP visual and textual encoder, and also sees benefits in scaling up to the larger ViT encoder. We also illustrate the effectiveness of FaD-VLP on another understanding task: multimodal categorization. Results on CR and SR indicate that the global fused representations are better for predicting categories and fine-grained subcategories after domain-specific pre-training.

Previous methods that evaluated IC were encoder-based: while these models, *e.g.*, (Zhuge et al., 2021), are evaluated on both ITR/TIR and IC, the IC performance was suboptimal because these methods used sequential Masked Language Modeling predictions for generating text at inference time. Our results highlight the benefit of our decoder-based architecture (even without CLIP initialization or domain-specific pre-training), which is effective at ITR/TIR *and* IC. We see additional gains with generic VLP and domain-specific VLP. For RIC, we compare to a decoder baseline that concatenates the two input image representations and uses a single cross attention mechanism. The RIC results further highlight the effectiveness of triplet pre-training on top of paired pre-training.

5.2 Effect of Pre-training Tasks

We conduct a thorough ablation study on the pre-training tasks to analyze the impact of our various pre-training objectives (Table 7).

Effect of Paired Pre-training. Our paired pre-training tasks (CMC and ICLM) provide a clear stepwise improvement on IRTF, ITR, TIR, and RIC (see lines (1), (2) and (6)-(8)), while the individual gains are less evident in the IC, CR, and SR tasks.

Effect of Triplet Pre-training. Although the triplet pre-training tasks (HMC and RCLM) use no additional data, we find that they provide further improvement on top of the paired pre-training tasks (see lines (2)-(4) and (8)-(10)). This occurs in IRTF, where the triplet tasks lead to approximately 2 point gain, as well as ITR and TIR, where they lead to a 3 point gain (when used on top of the CLIP initialization). RIC also sees a gain with triplet pre-training of above 2.3 points. There appear to be smaller gains for IC, CR, and SR.

Effect of Bootstrapping. We additionally experiment specifically with the effect of bootstrapping in the generation of relative captions (in which we feed samples from the multimodal decoder in the Relative Captioner mode to the HMC task in the

Figure 4: Examples of the top 5 results for two examples from FashionIQ. For each reference image in FashionIQ, there are two pieces of text feedback.

Fuser mode). We find that bootstrapping yields a small benefit for IRTF, ITR, and TIR (see line (5) and (11)), indicating that the global representations may become more robust to noise when the Fuser is trained with more diverse relative captions.

5.3 Qualitative Results

We visualize results on IRTF in Fig. 4. These examples illustrate the trained model's ability to handle compositional changes, *e.g.*, adding a blue print as well as short sleeves. They also illustrate the model's ability to handle changes to certain mentioned attributes while not changing unmentioned attributes; for example, adding a blue print and short sleeves does not alter the length of the dress.

6 Conclusion

Our work introduces a novel fashion-specific pre-training framework based on weakly-supervised triplets, constructed from paired fashion image-text data, and flexible decoder-based model architecture capable of both retrieval and captioning tasks in fashion. Our approach outperforms baselines on a diverse set of fashion tasks and highlight the value of fashion-specific pre-training, as well as the promise of triplet-based pre-training.

Limitations

Although fashion is a huge global industry, our experiments have been limited primarily to Western fashion styles with English language descriptions. It is likely that other clothing types contain properties that affect the performance of our method (e.g., different categories of attributes and different degrees of variation between products) and that other languages are less amenable to our token-wise method for constructing relative captions. Additionally, while this system sees gains from pre-training, corresponding limitations arise as well: we require several GPUs to pre-train FaD-VLP and to fine-tune it on our downstream tasks. Building smaller, more resource-efficient models with the same performance is an open question. We discuss broader limitations in the following section.

Broader Impact

Our model is motivated by e-commerce applications, and demonstrates good performance on benchmarks, but it is not suitable for commercial use in its current form. It likely carries biases present in its training data: the text in its pre-training dataset was sourced from human-provided descriptions, and the distribution of fashion items in the datasets likely reflect societal stereotypes and expectations about characteristics such as gender. The technology presented in this paper could be used to support interactive shopping assistants, which have the potential to help users locate items that match their preferences. However, style is a personal, subjective form of self-expression, and so further research could focus on individualizing results to particular users.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank Arka Sadhu, Adina Williams, Alex Tamkin, Siddharth Karamcheti, and Benjamin Newman for helpful discussions.

References

Kenan E. Ak, Joo-Hwee Lim, Jo Yew Tham, and Ashraf A. Kassim. 2018. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

Yashar Deldjoo, Fatemeh Nazary, Arnau Ramisa, Julian J. McAuley, Giovanni Pellegrini, Alejandro Bellogín, and Tommaso Di Noia. 2022. A review of modern fashion recommender systems. *arXiv*, abs/2202.02757.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*.

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Xiaoyong Wei, Minlong Lu, and Xiaodan Liang. 2021. M5Product: A multi-modal pretraining benchmark for e-commercial product downstream tasks. *arXiv*, abs/2109.04275.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018 (BMVC)*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, Hao Sun, and Ji-Rong Wen. 2021. Wenlan 2.0: Make AI imagine via a multimodal foundation model. *arXiv*, abs/2110.14378.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashion-BERT: Text and image matching with adaptive loss for cross-modal retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. FashionVLP: Vision language transformer for fashion retrieval with feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao Han, Sen He, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2022. UIGR: Unified interactive garment retrieval. *arXiv*, abs/2204.03111.

Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic spatially-aware fashion concept discovery. In *International Conference on Computer Vision (ICCV)*.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An image-based virtual try-on network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *AAAI Conference on Artificial Intelligence*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, abs/1411.2539.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *European Conference on Computer Vision (ECCV)*, volume 11208.

Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv*, abs/1908.03557.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, volume 12375.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv*, abs/2001.07966.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-Gen: The generative fashion dataset and challenge. *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing (NeurIPS)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. In *International Joint Conference on Intelligence (IJCAI)*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple visual language model pretraining with weak supervision.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han Yang, Xinrui Yu, and Ziwei Liu. 2022. Full-range virtual try-on with recurrent tri-level transform. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision (ECCV)*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022a. CoCa: Contrastive captioners are image-text foundation models. *arXiv*, abs/2205.01917.

Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao M. J. Wang, Hugo Chen, Tamara L. Berg, and Ning Zhang. 2022b. CommerceMM: Large-scale commerce multimodal representation learning with omni retrieval. *arXiv*, abs/2202.07247.

Yifei Yuan and Wai Lam. 2021. Conversational fashion image retrieval via multiturn natural language feedback. In *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv*, abs/2010.00747.

Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. 2021. UFC-BERT: unifying multi-modal controls for conditional image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive learning for image retrieval with hybrid-modality queries. In *International ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI Conference on Artificial Intelligence*.

Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. 2021. Knowledge perceived multi-modal pretraining in e-commerce. In *MM '21: ACM Multimedia Conference*.

Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-BERT: Vision-language pre-training on fashion domain. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Training Details

This section describes our implementation and training details for pre-training and fine-tuning on each downstream task.

Implementation. In our model, we use a ResNet-50 image encoder (He et al., 2016) initialized from CLIP (Radford et al., 2021b). We initialize the text and multimodal decoder from BERT-base (Devlin et al., 2019), using the first 6 layers for the text decoder and the second 6 layers for the multimodal decoder. The dimensionality of our joint embedding space is 2048. We implement our models in PyTorch (Paszke et al., 2019) and pre-train on two 8 GPU NVIDIA A100 nodes.

Pre-training Datasets. In the table below, we show statistics for the 5 datasets we use in our pre-training dataset. For datasets that are split into train and validation, we only use the training set.

Name	# Pairs	# Tokens
FACAD	888K	21.15 ± 4.49
Fashion-Gen	260K	37.35 ± 14.70
Fashion200K	172K	4.84 ± 1.32
Shopping100K	100K	20.00 ± 3.99
DeepFashion	26K	53.03 ± 19.02

Table 8: Statistics for datasets in our pre-training dataset: the number of image-caption pairs and the mean $\pm$ std. number of tokens per caption. Token statistics are calculated on a sample of 10K entries.

Below, we describe each dataset briefly:
FACAD (Yang et al., 2020) is a dataset of fashion images and descriptions crawled from the web, with 888K pairs in its training set.

DeepFashion (Liu et al., 2016) consists of diverse clothing images with different annotation types. We use the 26K images and descriptions in the In-shop Clothes Retrieval benchmark, which is sourced from an online catalogue. We use a concatenation of the color annotation and in-shop description as the caption.

Shopping100k (Ak et al., 2018) consists of 100K pairs of fashion images and attributes that are collected from meta-data on shopping websites. Each image has at least 5 structured attributes, which we concatenate to form a caption (*e.g.*, *t-shirt with white color and jersey fabric and regular fit*).

Fashion-Gen (Rostamzadeh et al., 2018) consists of 260K fashion images and detailed descriptions, annotated by professional stylists, in its training set. The image span 48 categories and 121 fine-grained subcategories.

Fashion200K (Han et al., 2017) consists of 172K clothing images in its training set along with descriptions from shopping websites. We exclude 15 pairs (0.0087%) that have broken image links. We use a concatenation of the attribute annotations as the caption.

After assembling pre-training dataset, we hold out a random split of 3% of the data (for any validation needed on pre-training data). Our final pre-training dataset has 1.4 million image-text pairs.

Pre-training Details. We have two pre-training stages, one for paired pre-training and the second for paired pre-training plus triplet pre-training; we found the two-stage setup to be more effective than a single stage. For both stages, we sum the appropriate losses. We can calculate CMC/ICLM from a single image-text pair and HMC/RCLM from a single triplet (with the same reference image as the pair). For pre-training, we use the Adam optimizer with a learning rate of 10^{-5}. We use a batch size of 512 split over 2 nodes with a total of 16 GPUs. In the first stage of pre-training (paired pre-training tasks), we train for 82K steps, taking approximately 1.5 days. In the optional second stage (paired pre-training plus triplet pre-training tasks), we train for 55K steps, which takes another approximately 1.5 days. Our model has 1.9×10^8 parameters. We use the following preprocessing/data augmentations on the images: resize to 224×224 pixels with a random crop, that has random scaling between 0.8 and 1.0 and aspect ratio between 0.75 and 1.3, along with a random horizontal flip. We use the same preprocessing for downstream fine-tuning.

For constructed triplets, we use $|S| = 1000$, and we weight the three metrics in Δ to have approximately equal relative weight by setting set $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = \frac{1}{16}$. We use `nltk` (Bird et al., 2009) for word and sentence tokenization. We sample from the following templates when constructing relative captions: {`modify <r> to be <t>`, `<t> instead of <r>`, `change <r> to <t>`, and `replace <r> with <t>`} where `<r>` and `<t>` represent the reference and target tokens respectively from the procedure from Section 3.3. If either `<r>` or `<t>` is empty, we run the procedure again on the next sentence in the reference and target captions.

B Fine-tuning Details

We provide details on our fine-tuning stage for each downstream task below. For all tasks, we use a batch size of 128 distributed over 4 GPUs.

ITR/TIR. We use the Adam optimizer with a learning rate of 10^{-5} and fine-tune for up to 120K steps. Fine-tuning takes approximately 20 hours.

IRTF. We use the Adam optimizer with a warmup learning rate of 10^{-6} for 140 steps, followed cosine decay from 3×10^{-5} over 4,100 steps. Fine-tuning takes about 1.5 hours. In our ablation, we found that models without pre-training take longer to converge for this task; to strengthen our baseline comparison, we train these for 25K steps, which takes approximately 5 hours.

CR/SR. We use the Adam optimizer with a learning rate of 10^{-5} and fine-tune for up to 81K steps. Fine-tuning takes approximately 14 hours.

IC. We use the Adam optimizer with a learning rate of 10^{-5} and fine-tune for up to 40K steps. Fine-tuning takes approximately 7 hours.

RIC. We use the Adam optimizer with a learning rate of 10^{-5} and fine-tune for up to 3K steps. Fine-tuning takes approximately 1 hour.

C Evaluation Details

This section describes our evaluation pipeline for the 7 downstream tasks.

ITR/TIR. We evaluate cross-modal retrieval on Fashion-Gen (Rostamzadeh et al., 2018). We compare to the results reported by Zhuge et al. (2021) and for fair comparison, follow their evaluation protocol, which works as follows. For each image (or text) query in the evaluation set of 32,528 queries, the model is to pick the matched text (or image) out of a sample of 101 candidates. The 101 candidates contain 1 correct match as well as 100 sampled texts (or images) from other products in the same subcategory. If a subcategory has less than 100 products, we sample the remaining negative examples from the same category. Since the random sample of candidates for each query is not released by the authors, we generate 5 random samples of categories for our experiments and report the average. Like Zhuge et al. (2021), we report Recall@1, Recall@5, and Recall@10.

IRTF. We evaluate this task on Fashion IQ (Wu et al., 2021) using the protocol laid out by the authors and reused in other works (Kim et al., 2021). For each category (Dress, Shirt, Toptee), there is a set of evaluation queries (of size 2017, 2038, and 1961 respectively) as well as a gallery of candidates (of size 3817, 6346, 5373 respectively) that includes a correct match. We report Recall@5 and Recall@10 on each of the three categories.

CR/SR. These tasks are trained and evaluated separately on classification into 48 categories or 121 subcategories in Fashion-Gen (Rostamzadeh et al., 2018). We follow Zhuge et al. (2021) and report accuracy and macro-F1 score.

IC/RIC. For IC, we evaluate BLEU-4, METEOR, ROUGE-L, and CIDEr scores using `pycocoevalcap`[1] following Zhuge et al. (2021). We use the same evaluation metrics for RIC. As Fashion IQ data contains two relative captions for every (reference, target) pair, we predict two relative captions in RIC, and concatenate them with "`and`". We apply the same procedure to the ground-truth relative captions before calculating the evaluation metrics.

D Sample Outputs

[1] https://github.com/LuoweiZhou/coco-caption