# OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data

Giuseppe Cartella[1[0000−0002−5590−3253]], Alberto Baldrati[2,3[0000−0002−5012−5800]], Davide Morelli[1,3[0000−0001−7918−6220]], Marcella Cornia[1[0000−0001−9640−9385]], Marco Bertini[2[0000−0002−1364−218X]], and Rita Cucchiara[1[0000−0002−2239−283X]]

[1] University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it
[2] University of Florence, Florence, Italy
{name.surname}@unifi.it
[3] University of Pisa, Pisa, Italy

**Abstract.** The inexorable growth of online shopping and e-commerce demands scalable and robust machine learning-based solutions to accommodate customer requirements. In the context of automatic tagging classification and multimodal retrieval, prior works either defined a low generalizable supervised learning approach or more reusable CLIP-based techniques while, however, training on closed source data. In this work, we propose OpenFashionCLIP, a vision-and-language contrastive learning method that only adopts open-source fashion data stemming from diverse domains, and characterized by varying degrees of specificity. Our approach is extensively validated across several tasks and benchmarks, and experimental results highlight a significant out-of-domain generalization capability and consistent improvements over state-of-the-art methods both in terms of accuracy and recall. Source code and trained models are publicly available at: https://github.com/aimagelab/open-fashion-clip.

**Keywords:** Fashion Domain · Vision-and-Language Pre-Training · Open-Source Datasets

## 1 Introduction

In the era of digital transformation, online shopping, and e-commerce have experienced an unprecedented surge in popularity. The convenience, accessibility, and variety offered by these platforms have revolutionized the way consumers engage with retail. Such digital shift creates an immense volume of data, therefore, the need for scalable and robust machine learning-based solutions to accommodate customer requirements becomes increasingly vital [42,48,49]. In the fashion domain, this includes tasks such as cross-modal retrieval [18, 27], recommendation [10,21,39], and visual product search [2–4,32,44], which play a crucial role in

enhancing user experience, optimizing search functionality, and enabling efficient product recommendation systems.

To address these challenges, innovative solutions that combine vision-and-language understanding have been proposed [19, 30, 50]. Although prior works have made noteworthy contributions in the fashion domain, they still suffer from some deficiencies. Approaches like [4] are able to well fit a specific task but struggle to adapt to unseen datasets and exhibit sub-optimal performance when faced with domain shifts. This results in poor zero-shot capability.

On the contrary, other techniques have employed CLIP-based methods [26, 47], which offer better generalization capabilities thanks to the pre-training on large-scale datasets. Some works as [8], have often relied on closed-source data, limiting their applicability and hindering the ability to reproduce and extend results. Therefore, there remains a need for a scalable and reusable method that can leverage open-source fashion data with varying levels of detail while demonstrating improved generalization and performance. In response to the aforementioned challenges, in this paper, we propose OpenFashionCLIP, a vision-and-language contrastive learning method that stands out from previous approaches in several ways. We adopt open-source fashion data from multiple sources encompassing diverse styles and levels of detail. Specifically, we adopt four publicly available datasets for the training phase, namely FashionIQ [44], Fashion-Gen [36], Fashion200K [20], and iMaterialist [17]. We believe this approach not only enhances transparency and reproducibility but also broadens the accessibility and applicability of our technique to a wider range of users and domains.

The contrastive learning framework employed in OpenFashionCLIP enables robust generalization capabilities, ensuring consistent performance even in the presence of domain shifts and previously unseen data. Our method adopts a fashion-specific prompt engineering technique [6, 16, 35] and is able to effectively learn joint representations from multiple domains. OpenFashionCLIP overcomes the limitations of supervised learning approaches and closed-source data training, facilitating seamless integration between visual and textual modalities.

Extensive experiments have been conducted to evaluate the effectiveness of OpenFashionCLIP across diverse tasks and benchmarks. We provide a comparison against CLIP [35], OpenCLIP [43] and a recent CLIP-based method fine-tuned on closed-source fashion data, namely FashionCLIP [8]. The experimental results highlight the significant out-of-domain generalization capability of our method. Notably, our fine-tuning strategy on open-source fashion data yields superior performance compared to competitors in several metrics, thus underscoring the benefits of leveraging open-source datasets for training.

## 2 Related Work

The ever-growing interest of customers in e-commerce has made the introduction of innovative solutions essential to enhance the online experience. On this basis, recommendation systems play a crucial role and numerous works have been introduced [10, 11, 21, 39]. An illustrative example is the automatic creation of

capsule wardrobes proposed in [21], where given an ensemble of garments and accessories the proposed method provided some possible visually compatible outfits. One of the most significant challenges of this task is the understanding of what visual compatibility means. To this aim, Cucurull *et al.* [10] addressed the compatibility prediction problem by exploiting the context information of fashion items, whereas Sarkar *et al.* [39] exploited a Transformer-based architecture to learn an outfit-level token embedding which is then fed through an MLP network to predict the compatibility score. In addition, De Divitiis *et al.* [11] introduced a more fine-grained control over the recommendations based on shape and color.

Generally, users desire to seek a specific article in the catalog with relative ease, therefore, designing efficient multimodal systems represents another important key to success for the fashion industry. A considerable portion of user online interactions fall into the area of multimodal retrieval, the task of retrieving an image corresponding to a given textual query, and vice versa. Prior works range from more controlled environments [23,27] to in-the-wild settings [18], where the domain shift between query and database images is a challenging problem.

Beyond recommendations and retrieval, another research line that is currently attracting attention is the one of virtual try-on, both in 3D [29, 37, 38] and 2D [13–15, 24, 31, 33, 46]. Virtual try-on aims to transfer a given in-shop garment onto a reference person while preserving the pose and the identity of the model. A related area is the one marked by fashion image editing [5, 12, 34]. While Dong *et al.* [12] conditioned the fashion image manipulation process on sketches and color strokes, other approaches [5, 34] introduced for the first time a multimodal fashion image editing conditioned on text.

Specifically, Pernuš *et al.* [34] devised a GAN-based iterative solution to change specific characteristics of the given image based on a textual query. Baldrati *et al.* [5], instead, focused on the creation of new garments exploiting latent diffusion models and conditioning the generation process on text, sketch, and model's pose. Solving the aforementioned downstream tasks has been made possible due to large-scale architectures explicitly trained on fashion data which effectively combine vision-and-language modalities to learn more powerful representations [19, 50]. Recent approaches exploit CLIP embeddings [35] to obtain more scalable and robust solutions able to generalize to different domains without supervision [8], but the closed source data training represents the main flaw.

## 3 On the Adaptation of CLIP to the Fashion Domain

### 3.1 Fashion-Oriented Contrastive Learning

Despite the significant scaling capability of large vision-and-language models such as CLIP, such a property comes at a cost. The pre-training of these models is usually conducted on datasets that contain million [35], or even billion [40] image-text pairs that, however, are gathered from the web and thus very noisy. Unfortunately, such coarse-grained annotations have been shown to lead to suboptimal performance for vision-and-language learning [9,25]. Moreover, the adaptation of CLIP to the specific domain of fashion is far from trivial. Indeed, a

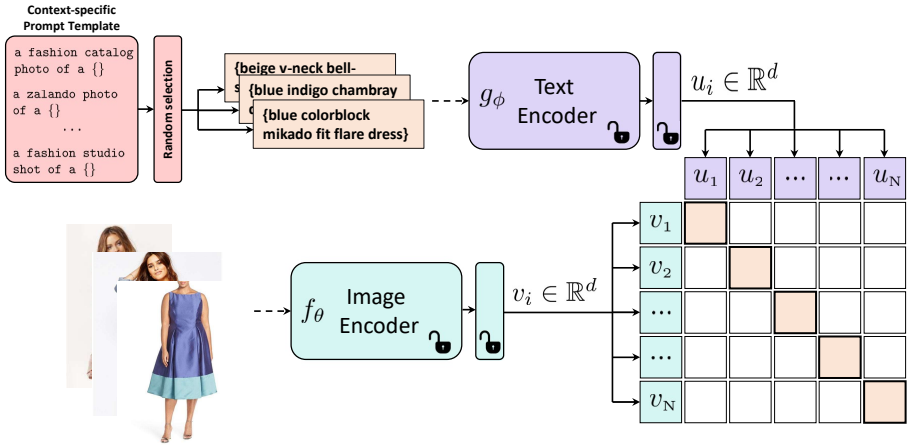


**Fig. 1.** Overview of our proposed method. We fine-tune both encoders and the linear projection layers toward the embedding space.

significant part of the images contained in these datasets is associated with incomplete captions or even worse, with simple and basic tags collected exploiting posts uploaded on the web by general and non-fashion-expert users. Considering these flaws, an adaptation of CLIP to a specific domain, uniquely relying on a vanilla pre-trained version, would not enable the attainment of optimal results. In our context, training on fashion-specific datasets containing fine-grained descriptions of garments and fashion accessories becomes crucial to obtain powerful representations while guaranteeing generalization and robustness to solve the tasks demanded by the fashion industries.

### 3.2 CLIP Preliminaries

Contrastive learning is a self-supervised machine learning technique that aims to learn data representations by constructing a powerful embedding space where semantically related concepts are close while dissimilar samples are pushed apart. On this line, the vision-and-language domain has already capitalized on such a learning technique. The CLIP model [35] represents the most common and illustrative method for connecting images and text in a shared multimodal space. The CLIP architecture consists of a text encoder $g_\phi$ and an image encoder $f_\theta$, trained on image-caption pairs $\mathcal{S} = \{(x_i, t_i)\}_{i=1}^{N}$.

The image encoder $f_\theta$ embeds an image $x \in \mathcal{X}$ obtaining a visual representation $\boldsymbol{v} = f_\theta(x)$. In the same manner, the text encoder $g_\phi$ takes as input a tokenized string $\tilde{t}$ and returns a textual embedding $\boldsymbol{u} = g_\phi(\tilde{t})$. For each batch $\mathcal{B}$ of image-caption pairs $\mathcal{B} = \{(x_i, t_i)\}_{i=1}^{L}$, where $L$ is the batch size, the objective is to maximize the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$ while minimizing the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{u}_j$, $\forall j \neq i$. The CLIP loss can be formally expressed as the sum of two symmetric terms:

$$\mathcal{L}_{contrastive} = \mathcal{L}_{T2I} + \mathcal{L}_{I2T}, \tag{1}$$

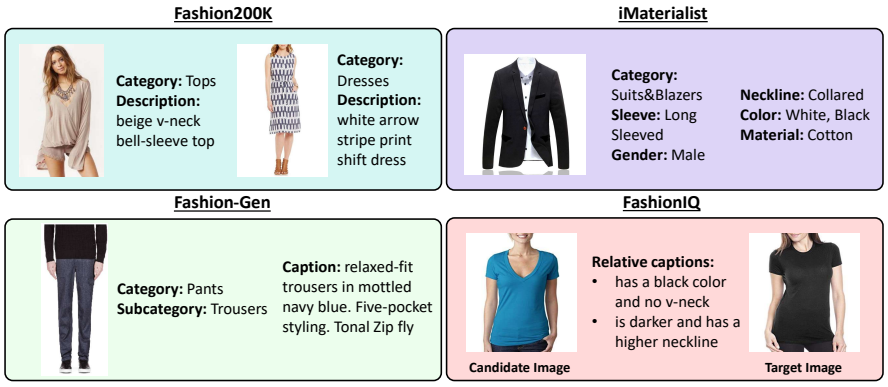


**Fig. 2.** Qualitative samples from the training datasets.

$$\mathcal{L}_{T2I} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i)}{\sum_{j=1}^{L}\exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j)}, \tag{2}$$

$$\mathcal{L}_{I2T} = -\frac{1}{L}\sum_{i=1}^{L}\log\frac{\exp(\tau \boldsymbol{v}_i^T \boldsymbol{u}_i)}{\sum_{j=1}^{L}\exp(\tau \boldsymbol{v}_i^T \boldsymbol{u}_j)}, \tag{3}$$

where $\tau$ represents a temperature parameter.

### 3.3 Open Source Training

In the fashion domain, several datasets, characterized by multimodal annotations from human experts, have been introduced. Differently from prior work [8] that fine-tuned CLIP on a private dataset, we devise a contrastive learning strategy entirely based on open-source data. An overview of the proposed CLIP-based fine-tuning is shown in Fig. 1. In detail, we adopt four publicly available datasets:

**Fashion-Gen [36].** The dataset contains a total of $325,536$ high resolution images ($1360 \times 1360$) with $260,480$ samples for the training set and $32,528$ images both for validation and test set. In addition, 48 main categories and 121 fine-grained categories (*i.e. subcategory*) are defined.

**Fashion IQ [44].** There are $77,684$ images, divided into three main categories (dresses, shirts, and tops&tees), with product descriptions and attribute labels.

**Fashion200K [20].** It contains $209,544$ clothing images from five categories (dresses, tops, pants, skirts, and jackets) and an associated textual description.

**iMaterialist [17].** It is a multi-label dataset containing over one million images and 8 groups of 228 fine-grained attributes.

These datasets are characterized by different levels of detail of the image annotations. FashionIQ has been proposed to accomplish the task of interactive image retrieval, therefore, the captions are relative to what should be modified in the source image to retrieve the target image. On the contrary, iMaterialist

only contains attributes while Fashion-Gen and Fashion200K present more semantically rich descriptions. As a pre-processing step, we apply lemmatization and extract the noun chunks from the textual descriptions. Noun chunks are sequences of words that include a noun and any associated word that modifies or describes that noun (*e.g.* an adjective). In particular, we adopt the spaCy[§] NLP library to extract noun chunks. Data pre-processing is performed for all datasets, except for iMaterialist which only contains simple attributes, thus making such an operation unnecessary. For the sake of clarity, from now on we refer to $t_i$ as the pre-processed caption after noun chunks extraction. Examples of image-caption pairs from the training datasets are reported in Fig. 2.

Compared to FashionCLIP which was trained on approximately $700k$ images, our training set is much larger and sums to $1,147,929$ image-text pairs. In detail, during fine-tuning, we construct each batch so that it contains image-text pairs from all the different data sources. Considering the great number of pairs of the complete training dataset, we fine-tune all the pre-trained weights of the CLIP model. Indeed, only training the projections toward the embedding space would not allow to fully effectively capture the properties of the data distribution.

### 3.4 Prompt Engineering

Prompt engineering is the technique related to the customization of the prompt text for each task. Providing context to the model has been shown to work well in a wide range of settings. Following prior works [6, 16, 35], we provide our model with a fashion-specific context defining a template of prompts related to our application domain. Specifically, given a template of prompts $\mathcal{P} = \{(p_i)\}_{i=1}^{|\mathcal{P}|}$, at each training step, we select a random $p_i \in \mathcal{P}$ for each image-caption pair $(x_i, t_i) \in \mathcal{B}$. The caption $t_i$ is concatenated to $p_i$ obtaining the final CLIP input.

The complete fashion-specific template includes the following prompts: `"a photo of a"`, `"a photo of a nice"`, `"a photo of a cool"`, `"a photo of an expensive"`, `"a good photo of a"`, `"a bright photo of a"`, `"a fashion studio shot of a"`, `"a fashion magazine photo of a"`, `"a fashion brochure photo of a"`, `"a fashion catalog photo of a"`, `"a fashion press photo of a"`, `"a zalando photo of a"`, `"a yoox photo of a"`, `"a yoox web image of a"`, `"an asos photo of a"`, `"a high resolution photo of a"`, `"a cropped photo of a"`, `"a close-up photo of a"`, `"a photo of one"`.

## 4 Experimental Evaluation

In this section, we describe the open-source datasets used as benchmarks together with the tasks performed to assess the scalability and robustness of our approach.

### 4.1 Benchmark Datasets

We validate our approach across three different datasets:

[§] `https://github.com/explosion/spaCy`

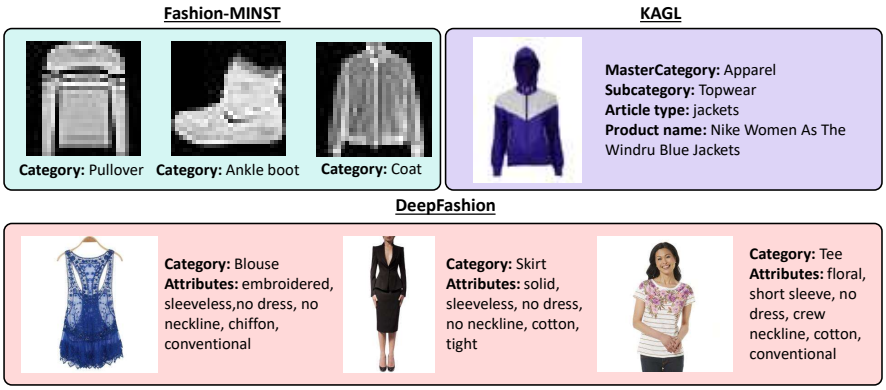


**Fig. 3.** Samples from the benchmark datasets.

**DeepFashion** [27] contains over 800,000 images and is divided into several benchmarks. In our experiments, we employ the attribute prediction subset which contains 40,000 images and 1,000 different attributes.

**Fashion-MNIST** [45] is based on the Zalando catalog and consists of 60,000 training images, a test set of 10,000 examples, and 10 categories. All images are in grayscale and have a $28 \times 28$ resolution. Following [8], we apply image inversion, thus working on images with a white background.

**KAGL** is a subset of [1] and contains 44,441 images equipped with textual annotations including the master category, the sub-category, the article type, and the product description. In detail, we filter out all those images not belonging to the *'apparel'* master category and kept the images depicting humans, resulting in a total of 21,397 samples, 8 sub-categories, and 58 article types. Qualitative examples of the adopted benchmarks are reported in Fig. 3.

### 4.2 Implementation Details

We train the final model for 60 epochs using a batch size of 2048. To save memory, we adopt the gradient checkpointing technique [7]. AdamW [28] is employed as optimizer, with $\beta_1$ set to 0.9 and $\beta_2$ equal to 0.98, epsilon of $1e-6$, and weight decay equal to 0.2. A learning rate of $5e-7$ and automatic mixed precision are applied. For a fair comparison with competitors, we select the ViT-B/32 backbone as the image encoder. During training, we apply the prompt engineering strategy described in Sec. 3.4.

As a pre-trained CLIP model, we refer to the OpenCLIP implementation [22] trained on LAION-2B [40] composed of 2 billion image-text pairs. In the evaluation phase, following the pre-processing procedure of [35], we resize the image along the shortest edge and apply center crop.

### 4.3 Zero-shot Classification

In our context, zero-shot classification refers to the task of classification on unseen datasets characterized by different data distributions compared to the training datasets. The task is crucial to assess the transfer capability of the model to adapt to new and

**Table 1.** Category prediction results on the Fashion-MNIST and the KAGL datasets.

| | | | | **F-MNIST** | | **KAGL** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Backbone** | **Pre-Training** | **Fine-tuned** | Acc@1 | F1 | Acc@1 | Acc@5 | Acc@10 | F1 |
| CLIP | ViT-B/16 | OpenAI WIT | ✗ | 69.29 | 67.88 | 31.54 | 70.08 | 90.09 | 36.04 |
| CLIP | ViT-B/32 | OpenAI WIT | ✗ | 69.51 | 66.56 | 21.44 | 66.13 | 84.97 | 27.70 |
| OpenCLIP | ViT-B/32 | LAION-400M | ✗ | 81.62 | 81.16 | 33.69 | 76.60 | 89.23 | 37.89 |
| OpenCLIP | ViT-B/32 | LAION-2B | ✗ | 83.69 | 82.75 | 46.18 | 84.49 | 95.44 | 51.23 |
| FashionCLIP | ViT-B/32 | LAION-2B | ✓ | 82.23 | 82.03 | **52.90** | 85.41 | 93.40 | **54.48** |
| **OpenFashionCLIP** | ViT-B/32 | LAION-2B | ✓ | **84.33** | **84.19** | 45.97 | **88.30** | **96.46** | 53.85 |

**Table 2.** Attribute recognition results on the DeepFashion dataset.

| | | | | **Overall Recall** | | | **Per-Class Recall** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Backbone** | **Pre-Training** | **Fine-tuned** | R@3 | R@5 | R@10 | R@3 | R@5 | R@10 |
| CLIP | ViT-B/16 | OpenAI WIT | ✗ | 8.00 | 11.40 | 17.54 | 13.31 | 17.42 | 24.54 |
| CLIP | ViT-B/32 | OpenAI WIT | ✗ | 7.35 | 10.30 | 16.60 | 11.39 | 15.13 | 21.67 |
| OpenCLIP | ViT-B/32 | LAION-400M | ✗ | 12.58 | 17.22 | 25.64 | 17.9 | 22.81 | 30.71 |
| OpenCLIP | ViT-B/32 | LAION-2B | ✗ | 13.07 | 17.70 | 26.13 | 19.35 | 24.31 | 32.51 |
| FashionCLIP | ViT-B/32 | LAION-2B | ✓ | 15.19 | 20.83 | 32.37 | 17.30 | 22.27 | 30.56 |
| **OpenFashionCLIP** | ViT-B/32 | LAION-2B | ✓ | **24.47** | **32.97** | **45.77** | **28.67** | **36.07** | **47.28** |

unseen domains. Following the standard CLIP evaluation setup [35], we perform classification by embedding the image and all $k$ categories. Regarding prompt engineering, we always append every category `{label}` to the same generic prompt `"a photo of a"`. We feed each category prompt through the CLIP textual encoder $g_\phi$ obtaining a set of feature vectors $\mathcal{U} = \{(\boldsymbol{u}_i)\}_{i=1}^{N}$. In the same manner, we feed the image $x_i$ through the CLIP image encoder to get the embedded representation $\boldsymbol{v} = f_\phi(x_i)$. To classify the image we compute the cosine similarity between $\boldsymbol{v}$ and each text representation $\boldsymbol{u}_i$. The predicted category is the one with the highest similarity. Experiments have been conducted on the test splits of Fashion-MNIST, KAGL, and on the attribute prediction benchmark of DeepFashion. Our model is compared against the original CLIP model [35], which was trained on the private WIT dataset, OpenCLIP [43] pre-trained on LAION-400M [41] and LAION-2B [40], and FashionCLIP [8] that was fine-tuned on closed source data from *Farfetch*. Note that we have reproduced the results of FashionCLIP by exploiting the source code released by the authors and adapting it to our tasks and settings. The task is evaluated considering three well-known metrics, namely accuracy@$k$, recall@$k$, and weighted $F1$ score. The accuracy@$k$ computes the number of times the correct label is among the top $k$ labels predicted by the model. The recall@$k$, instead, measures the number of relevant retrieved items with respect to the total number of relevant items for a given query. The weighted F1 score accounts for the class distribution in the dataset by calculating the F1 score for each class individually and then averaging based on the class frequencies.

Quantitative results on Fashion-MNIST and KAGL are summarized in Table 1. The first aspect to mention is the improvement against CLIP and OpenCLIP on all metrics and both datasets, indicating the effectiveness of our fine-tuning strategy enabling a strong generalization and adaptation of our model to the specific fashion domain. Compared to FashionCLIP, our model shows better performance on Fashion-MNIST, while when tested on the 58 article types of KAGL, the results are comparable. OpenFashionCLIP performs better with the increase of the number of considered categories.

Table 2 shows the results on the attribute prediction benchmark of the DeepFashion dataset. Categories of this dataset are attributes describing different garment charac-

**Table 3.** Cross-modal retrieval results on the KAGL dataset.

| | | | | Image-to-Text | | | Text-to-Image | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Backbone** | **Pre-Training** | **Fine-tuned** | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| FashionCLIP | ViT-B/32 | LAION-2B | ✓ | 6.61 | 19.23 | 28.66 | 6.97 | 19.14 | 27.49 |
| **OpenFashionCLIP** | ViT-B/32 | LAION-2B | ✓ | **7.57** | **20.72** | **30.38** | **7.73** | **20.58** | **28.56** |

**Table 4.** Ablation study to assess the validity of the prompt engineering technique.

| | **F-MNIST** | | **KAGL** | | **DeepFashion** | | **KAGL** | |
|---|---|---|---|---|---|---|---|---|
| **Model** | Acc@1 | F1 | Acc@1 | F1 | R@3 | R@3 (cls) | R@1 (I2T) | R@1 (T2I) |
| w/o prompt engineering | 83.21 | 82.99 | **47.51** | 47.3 | 20.34 | 25.21 | 7.47 | **7.73** |
| **OpenFashionCLIP** | **84.33** | **84.19** | 45.97 | **53.85** | **24.47** | **28.67** | **7.57** | **7.73** |

teristics (*e.g.* v-neck, sleeveless, etc.), therefore we leverage the recall metric in this setting to account for the multi-label nature of the dataset. In particular, we evaluate both the per-class recall@$k$ and the overall recall@$k$ among all attributes. In this case, our solution outperforms FashionCLIP by a consistent margin, highlighting the effectiveness of our training strategy with data of different annotation detail granularity.

### 4.4 Cross-modal Retrieval

Cross-modal retrieval refers to the task of retrieving relevant contents from a multi-modal dataset using multiple modalities such as text and images. Different modalities should be integrated to enable an effective search based on the user's input query. Cross-modal retrieval can be divided into two sub-tasks: image-to-text and text-to-image retrieval. In the first setting, given a query image $x$, we ask the model to retrieve the first $k$ product descriptions that better match the image. On the opposite, in text-to-image retrieval, given a text query, the first $k$ images that better correlate with the input query are returned. In Table 3, we evaluate our fine-tuning method on the KAGL dataset in terms of recall@$k$ with $k = 1, 5, 10$. OpenFashionCLIP performs better compared to FashionCLIP on both settings and according to all recall metrics, thus further confirming the effectiveness of our proposal.

### 4.5 Effectiveness of Prompt Engineering

Finally, in Table 4, we evaluate the individual contribution of prompt engineering in our fine-tuning method. We present the ablation study on all considered benchmarks. The first line of the table (*i.e.* *w/o* prompt engineering) refers to the case where we perform fine-tuning without using the fashion-specific template described in Sec. 3.4 but employing a fixed prompt (*i.e.* `"a photo of a"`). Notably, Fashion-MNIST is used for the classification task, DeepFashion for retrieval, and KAGL for both. As the results demonstrate, the idea to construct a fashion-specific set of prompts clearly performs well across all cases except for the KAGL classification benchmark. We argue that in general, domain-specific prompt engineering represents a key factor to obtain greater domain adaptation of the CLIP model.

## 5 Conclusion

In this paper, we introduced OpenFashionCLIP, a vision-and-language contrastive learning method designed to address the scalability and robustness challenges posed by

the fashion industry for online shopping and e-commerce. By leveraging open-source fashion data from diverse sources, OpenFashionCLIP overcomes limitations associated with closed-source datasets and enhances transparency, reproducibility, and accessibility. Our strategy, characterized by the fine-tuning of all pre-trained weights across all CLIP layers together with the adoption of a context-specific prompt engineering technique, effectively enables better adaption to our specific domain. We evaluated our strategy on three benchmarks and demonstrated that the proposed solution led to superior performance over the baselines and competitors achieving better accuracy and recall in almost all settings.

**Acknowledgements** This work has partially been supported by the European Commission under the PNRR-M4C2 (PE00000013) project "FAIR - Future Artificial Intelligence Research" and the European Horizon 2020 Programme (grant number 101004545 - ReInHerit), and by the PRIN project "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content" (CUP B87G22000460001), co-funded by the Italian Ministry of University.

## References

1. Aggarwal, P.: Fashion Product Images (Small), `https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small`
2. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-Shot Composed Image Retrieval with Textual Inversion. arXiv preprint arXiv:2303.15247 (2023)
3. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned Image Retrieval for Fashion using Contrastive Learning and CLIP-based Features. In: ACM Multimedia Asia (2021)
4. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features. In: CVPR Workshops (2022)
5. Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. arXiv preprint arXiv:2304.02051 (2023)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language Models are Few-Shot Learners. In: NeurIPS (2020)
7. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training Deep Nets with Sublinear Memory Cost. arXiv preprint arXiv:1604.06174 (2016)
8. Chia, P.J., Attanasio, G., Bianchi, F., Terragni, S., Magalhães, A.R., Goncalves, D., Greco, C., Tagliabue, J.: Contrastive language and vision learning of general fashion concepts. Scientific Reports **12**(1), 18958 (2022)
9. Cornia, M., Baraldi, L., Fiameni, G., Cucchiara, R.: Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. arXiv preprint arXiv:2111.12727 (2022)
10. Cucurull, G., Taslakian, P., Vazquez, D.: Context-aware visual compatibility prediction. In: CVPR (2019)
11. De Divitiis, L., Becattini, F., Baecchi, C., Del Bimbo, A.: Disentangling features for fashion recommendation. ACM TOMM **19**(1s), 1–21 (2023)
12. Dong, H., Liang, X., Zhang, Y., Zhang, X., Shen, X., Xie, Z., Wu, B., Yin, J.: Fashion editing with adversarial parsing learning. In: CVPR (2020)

13. Fenocchi, E., Morelli, D., Cornia, M., Baraldi, L., Cesari, F., Cucchiara, R.: Dual-Branch Collaborative Transformer for Virtual Try-On. In: CVPR Workshops (2022)
14. Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Transform, Warp, and Dress: A New Transformation-Guided Model for Virtual Try-On. ACM TOMM **18**(2), 1–24 (2022)
15. Fincato, M., Landi, F., Cornia, M., Cesari, F., Cucchiara, R.: VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In: ICPR (2021)
16. Gao, T., Fisch, A., Chen, D.: Making Pre-trained Language Models Better Few-shot Learners. In: ACL (2021)
17. Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M.R., Belongie, S.: The iMaterialist Fashion Attribute Dataset. In: ICCV Workshops (2019)
18. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: ICCV (2015)
19. Han, X., Yu, L., Zhu, X., Zhang, L., Song, Y.Z., Xiang, T.: FashionViL: Fashion-Focused Vision-and-Language Representation Learning. In: ECCV (2022)
20. 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)
21. Hsiao, W.L., Grauman, K.: Creating capsule wardrobes from fashion images. In: CVPR (2018)
22. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (2021), `https://doi.org/10.5281/zenodo.5143773`
23. Kuang, Z., Gao, Y., Li, G., Luo, P., Chen, Y., Lin, L., Zhang, W.: Fashion retrieval via graph reasoning networks on a similarity pyramid. In: ICCV (2019)
24. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In: ECCV (2022)
25. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML (2022)
26. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In: ICLR (2022)
27. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: CVPR (2016)
28. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
29. Majithia, S., Parameswaran, S.N., Babar, S., Garg, V., Srivastava, A., Sharma, A.: Robust 3D Garment Digitization from Monocular 2D Images for 3D Virtual Try-On Systems. In: WACV (2022)
30. Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., Cucchiara, R.: Fashion-Oriented Image Captioning with External Knowledge Retrieval and Fully Attentive Gates. Sensors **23**(3), 1286 (2023)
31. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. arXiv preprint arXiv:2305.13501 (2023)
32. Morelli, D., Cornia, M., Cucchiara, R.: FashionSearch++: Improving consumer-to-shop clothes retrieval with hard negatives. In: CEUR Workshop Proceedings (2021)

33. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress Code: High-Resolution Multi-Category Virtual Try-On. In: ECCV (2022)
34. Pernuš, M., Fookes, C., Štruc, V., Dobrišek, S.: FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion. arXiv preprint arXiv:2301.02110 (2023)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
36. 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)
37. Santesteban, I., Otaduy, M., Thuerey, N., Casas, D.: ULNeF: Untangled Layered Neural Fields for Mix-and-Match Virtual Try-On. In: NeurIPS (2022)
38. Santesteban, I., Thuerey, N., Otaduy, M.A., Casas, D.: Self-supervised collision handling via generative 3d garment models for virtual try-on. In: CVPR (2021)
39. Sarkar, R., Bodla, N., Vasileva, M.I., Lin, Y.L., Beniwal, A., Lu, A., Medioni, G.: OutfitTransformer: Learning Outfit Representations for Fashion Recommendation. In: WACV (2023)
40. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
41. Schuhmann, C., Kaczmarczyk, R., Komatsuzaki, A., Katta, A., Vencu, R., Beaumont, R., Jitsev, J., Coombes, T., Mullis, C.: LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In: NeurIPS Workshops (2021)
42. Shiau, R., Wu, H.Y., Kim, E., Du, Y.L., Guo, A., Zhang, Z., Li, E., Gu, K., Rosenberg, C., Zhai, A.: Shop The Look: Building a Large Scale Visual Shopping System at Pinterest. In: KDD (2020)
43. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust Fine-Tuning of Zero-Shot Models. In: CVPR (2022)
44. 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)
45. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv preprint arXiv:1708.07747 (2017)
46. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning. In: CVPR (2023)
47. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained Interactive Language-Image Pre-Training. In: ICLR (2022)
48. Zhai, A., Wu, H.Y., Tzeng, E., Park, D.H., Rosenberg, C.: Learning a unified embedding for visual search at pinterest. In: KDD (2019)
49. Zhang, Y., Pan, P., Zheng, Y., Zhao, K., Zhang, Y., Ren, X., Jin, R.: Visual Search at Alibaba. In: KDD (2018)
50. 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)