



OPEN

# Contrastive language and vision learning of general fashion concepts

Patrick John Chia<sup>1</sup>✉, Giuseppe Attanasio<sup>2</sup>, Federico Bianchi<sup>3</sup>, Silvia Terragni<sup>4,7</sup>, Ana Rita Magalhães<sup>5</sup>, Diogo Goncalves<sup>5</sup>, Ciro Greco<sup>6</sup> & Jacopo Tagliabue<sup>6</sup>

The steady rise of online shopping goes hand in hand with the development of increasingly complex ML and NLP models. While most use cases are cast as specialized supervised learning problems, we argue that practitioners would greatly benefit from general and transferable representations of products. In *this work*, we build on recent developments in contrastive learning to train *FashionCLIP*, a CLIP-like model adapted for the fashion industry. We demonstrate the effectiveness of the representations learned by *FashionCLIP* with extensive tests across a variety of tasks, datasets and generalization probes. We argue that adaptations of large pre-trained models such as CLIP offer new perspectives in terms of scalability and sustainability for certain types of players in the industry. Finally, we detail the costs and environmental impact of training, and release the model weights and code as open source contribution to the community.

**Generalization and scalability in machine learning.** The extraordinary growth of online retail—as of 2020, 4 trillion dollars per year<sup>1</sup>—has profoundly impacted the fashion industry, with 1 out of 4 transactions now happening online<sup>2</sup>. The combination of large amounts of data and a variety of use cases has made e-commerce fertile for cutting-edge machine learning (ML) models, with Natural Language Processing (NLP) involved in recommendations<sup>3–5</sup>, information retrieval (IR)<sup>6</sup>, product classification<sup>7</sup> and many other use cases<sup>8–10</sup>.

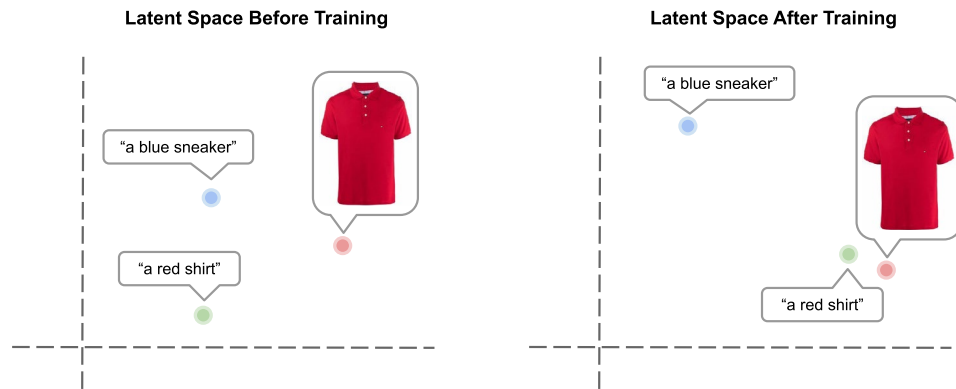
However, as the community starts to address the huge operational costs of training and developing models<sup>11</sup>, it is becoming clear that the value of ML innovations has been mostly captured by a few players<sup>12</sup>. Where the rest of the retail industry is making concrete efforts to adapt promptly, companies offering ML products as a service recently gained traction, creating a new multi-billion dollar market<sup>13–16</sup>. The need for ML capabilities that can be applied across entire industries and verticals raises the stakes for an age-old question in ML: *can we build models that can be reused on different tasks and datasets?*

While generalization is a theoretical virtue, real-world models often succeed by (over)fitting to a specific dataset and task<sup>17,18</sup>. In practice, generalization has been considered both hard to achieve *and* economically undesirable for large-scale use cases. In this context, the advent of large-scale, self-supervised models such as *Contrastive Language-Image Pre-training* (CLIP)<sup>17</sup> is particularly interesting both from a theoretical and a practical point of view. Building upon large pre-trained models to learn *general* concepts in specific verticals/industries (e.g., Fashion, Electronics, DIY, etc.) may provide a new and sustainable way to bring the benefits of ML capabilities to a broader set of practitioners, especially outside of large tech companies. The idea would be to fine-tune general foundational models<sup>19</sup> to learn concepts that are specific to a domain (e.g., fashion), but general enough to be applicable to all the use cases within that domain.

In *this work*, we show through extensive testing and open-source code that multi-modal training can be successfully used to learn general concepts in a specific domain, namely fashion. In fact, we will argue that it is not only technically possible, but also economically viable, and practically advantageous, since moving away from the traditional setting where single supervised models are trained specifically per use case reduces annotation and maintenance costs while providing solutions transferable across tasks.

**Self-supervised contrastive learning of fashion concepts.** Contrastive learning has recently become a predominant approach to learn meaningful representations of concepts in ML. The learning framework builds on the idea that semantically related *concepts* (e.g., two pictures of the same object from different viewpoints)

<sup>1</sup>Coveo, Montreal, Canada. <sup>2</sup>Bocconi University, Milan, Italy. <sup>3</sup>Stanford University, Stanford, CA, USA. <sup>4</sup>Telepathy Labs, Zurich, Switzerland. <sup>5</sup>Farfetch, Porto, Portugal. <sup>6</sup>South Park Commons, New York, USA. <sup>7</sup>University of Milano-Bicocca, Milan, Italy. ✉email: pchia@coveo.com



**Figure 1.** Two-dimensional representation of images and text in FashionCLIP vector space **before** and **after** training. Images and their corresponding textual descriptions are embedded closer to each other in the latent vector space after training.

should have *similar* representations, while unrelated ones should be *dissimilar*. Initially devised for self-supervised image representation learning<sup>20,21</sup>, contrastive learning has recently been applied to language as well<sup>22,23</sup>. Recent work has used contrastive training to bridge different modalities, e.g., vision and language<sup>24,25</sup>, audio and language<sup>26,27</sup>, or a combination of the three<sup>28,29</sup>. These models learn concept representations from different modalities (e.g., a textual excerpt such as “a dog running on a field” and a picture depicting the scene) and optimize them to be close in a shared latent space. Crucially, the typical pipeline is self-supervised: since no manual annotation is involved (e.g., in the previous example, one can gather image-text pairs from the web), human intervention is limited to deciding which pre-training task shall be used.

CLIP<sup>17</sup> is a vision-language multi-modal neural network trained via CL to associate vision concepts with text. The model comprises a vision and text encoder, each followed by a linear layer to project the image and text representations to the same latent space. CLIP is trained to *position* images and matching descriptions (e.g. an image of a red shirt and its description “a red shirt”) close together in the vector space (see Fig. 1 for an example). When trained on 400 million <image, text> pairs collected from the internet, CLIP has demonstrated competitive zero-shot or few-shot transfer to downstream tasks such as OCR and fine-grained object classification<sup>17</sup>.

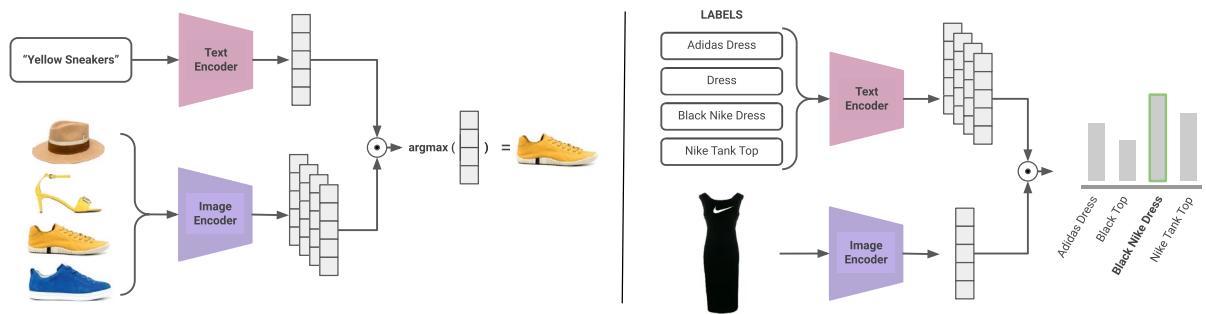
More formally, CLIP is a multi-modal model that makes use of an image ( $I_{\theta^I}$ ) and a text ( $T_{\theta^T}$ ) encoder. Both encoders are deep neural networks that map raw representations (i.e., an image and a text) to a 512-dimensional dense vector (e.g. given an image  $i$ ,  $I_{\theta^I}(i) \in \mathbb{R}^{512}$ ). During training,  $N$  pairs of matching images and texts  $\langle i, t \rangle$  are selected (e.g., as in Fig. 1, the image of a red shirt and the description “a red shirt”), encoded using  $I_{\theta^I}$  and  $T_{\theta^T}$ ,  $L_2$ -normalized, and compared pairwise. CLIP minimizes cross-entropy loss such that  $\bar{I}_{\theta^I}(i_j) \cdot \bar{T}_{\theta^T}(t_k)$  for  $j, k = 1, \dots, N$  is highest when the caption is paired with the correct image ( $j = k$ ), and low otherwise ( $j \neq k$ ), where  $\bar{I}_{\theta^I}(\cdot) / \bar{T}_{\theta^T}(\cdot)$  are the  $L_2$ -normalized outputs of the image and text encoders. We summarize the optimization objective for CLIP in Eq. (1) and (2).

$$\mathcal{L}(\theta^I, \theta^T) = -\frac{1}{2N} \left( \sum_j \log \frac{e^{\bar{I}_{\theta^I}(i_j) \cdot \bar{T}_{\theta^T}(t_j)}}}{\sum_{k'} e^{\bar{I}_{\theta^I}(i_j) \cdot \bar{T}_{\theta^T}(t_{k'})}} + \sum_k \log \frac{e^{\bar{I}_{\theta^I}(i_k) \cdot \bar{T}_{\theta^T}(t_k)}}}{\sum_{j'} e^{\bar{I}_{\theta^I}(i_{j'}) \cdot \bar{T}_{\theta^T}(t_k)}} \right) \quad (1)$$

$$\theta^{I*}, \theta^{T*} = \arg \min_{\theta^I, \theta^T} \mathcal{L}(\theta^I, \theta^T) \quad (2)$$

Here,  $\theta^I$  and  $\theta^T$  are the learnable parameters of the image and text encoder neural networks, and the  $(\cdot)$  operator represents the dot product. The first addition operand of Equation 1 is the cross-entropy on the image axis, while the second addition operand is on the text axis.

Recently, industry practitioners have begun to recognize the importance and utility of contrastive pre-training for their target domain, with several works presenting successful downstream applications starting from the CLIP model<sup>30</sup>. In fashion, the multi-modal nature of CLIP has been found helpful in recent discriminative<sup>31,32</sup> models, which have been developed under the standard paradigm of task-specific, supervised models. In the generative setup, CLIP often complements a larger framework: for example, CLIP is used to learn linguistically grounded codebooks in Variational Auto Encoders<sup>33</sup> or to guide image synthesis and manipulation in diffusion generative models.<sup>34,35</sup> While interesting for grounding (see below Fig. 8), the target use case (image generation) and more narrow focus (single task, single dataset) are not readily comparable to FashionCLIP, but instead suggest a possible complementary application in generative use cases. However, no recent CLIP application has been developed to produce industry-wide representations across multiple use cases and datasets. In other words, CLIP has been used only as a pre-trained model, with no attempt to overcome single-task supervised models' operational and conceptual problems.



**Figure 2.** Schematic overview of multi-modal retrieval (left) and zero-shot classification tasks (right).

In *this* work, we introduce FashionCLIP, a CLIP-based model explicitly trained and tested to produce general product representations for fashion concepts. We train FashionCLIP on a large, high-quality novel fashion dataset: as discussed in the next section, our goal is to establish whether such fine-tuning is sufficient to produce product representations that are transferable in a zero-shot fashion to entirely new datasets.

**Research question and methodology.** Standard supervised models for vertical-specific applications such as fashion are costly to train and operate, providing a large barrier to entry for SaaS providers and smaller players<sup>12</sup>. For example, a product classification model might be trained on  $\langle \text{product description}, \text{category} \rangle$  pairs derived from catalog data<sup>36</sup> while optimizing for classification accuracy: if the labels change, or the model is deployed on a different catalog, accuracy would drop. It is important to note that moving to CLIP-based architectures, such as CMA-CLIP<sup>37</sup>, does not *ipso facto* solve the problem: if CLIP is used as a per-task model, it will raise the same scalability issues as traditional supervised methods.

After training FashionCLIP, we set out to answer a broader, and potentially more impactful question: given the right dataset and fine-tuning procedure, can we learn multi-modal concepts that are general enough for the entire fashion domain? We proceed with a mixture of quantitative benchmarks – inspired both by existing literature and problems known to be important in the industry – and qualitative probes to answer it: since obtaining general concepts is our goal, it is important to verify that FashionCLIP does not only learn a dataset (e.g., an “Armani collection”), but genuine transferable concepts, such as “skirt”, “sleeves”, etc. Taking inspiration from CLIP, our two initial benchmarks will test how FashionCLIP goes from text to image, and vice versa (see Fig. 2):

1. *Text to image* Product search is one of the main channels of interactions and revenues between a shop and its users, accounting on average for 30% to 60% of the total online revenues<sup>38,39</sup>. Historically, product search has been performed chiefly with textual features by first matching queries and product descriptions in an index<sup>40–42</sup> and then re-ranking the candidate results<sup>43</sup>. However, there are good reasons to believe that including visual features can bring significant improvements since images are often the most curated aspect of the catalog. In contrast, text quality varies throughout verticals, languages, and specific product feeds. Our extensive tests show that FashionCLIP learns fashion concepts, and successfully applies them to unseen products and incomplete or ambiguous descriptions.
2. *Image to text* Product classification is the task of predicting a product category given its meta-data. Classification (even for CLIP-based models, such as CMA-CLIP) is cast as a supervised learning problem where one extracts golden labels from the catalog itself or collects them through crowd-sourcing<sup>7,44</sup>. Generalizing classification to arbitrary labels without constant retraining is again crucial for making ML feasible across numerous players in the fashion industry: transferable concepts help with the interoperability of overlapping, and yet different, fashion taxonomies<sup>45</sup>, a challenge increasingly recognized as central by both practitioners and commentators<sup>46</sup> (this includes the case of catalogs in less represented languages, for which an English classification is still desirable). Our extensive tests show that FashionCLIP zero-shot capabilities, based on learned associations between vision and textual concepts, allow for quick classification of products in target classes of interest, *irrespective of the specific labeling schemes of individual suppliers*.

We also perform specific probing to understand whether the concepts learned by the model are robust and (somehow) aligned with human semantic intuitions, as opposed to picking up spurious correlations in the dataset<sup>47</sup>:

1. *Grounding*. We probe FashionCLIP for grounding capabilities through localization maps and apply them to the task of zero-shot semantic segmentation for fashion concepts (e.g., sleeve length, texture patterns) and attributes (e.g., *laced shoes*, *glittered shirts*).
2. *Compositionality*. We propose a novel way to probe whether the model can go from semantic segmentation to inferential abilities by composing such concepts to generate new linguistic expressions. We do that through the device of “improbable object”, where a linguistic expression is meant to describe an odd combination of concepts that have never been observed before (e.g., a pair of shoes with handles).

We summarize our contributions as follows:

1. while other researchers have independently developed CLIP-based solutions for individual fashion problems, FashionCLIP is the first explicit attempt to produce general multi-modal concepts for industry: the breadth and nature of our testing methodology make FashionCLIP appealing as a general fashion model, applicable to situations where supervised systems are not practical or viable. Our model is trained on over 700k  $\langle \text{image}, \text{text} \rangle$  pairs from the inventory of Farfetch, one of the largest fashion luxury retailers in the world, and is shown to be useful in important use cases in a vast global market;
2. we evaluate FashionCLIP in various tasks, showing that fine-tuning helps capture domain-specific concepts and generalizes them in zero-shot scenarios; we supplement quantitative tests with qualitative analyses and offer insights into how concepts grounded in a visual space unlock linguistic generalization. These results would not be possible without the flexibility provided by natural language as a supervision signal and the domain-specific accuracy achieved through fine-tuning;
3. we transparently report training time, costs, and emissions. We additionally release to the community, under an open-source license, training code, a demo app, and plug-and-play checkpoints to help leverage our findings while facilitating ROI considerations<sup>48,49</sup>; the large and unique dataset is also scheduled to be released directly by Farfetch. Taken together, FashionCLIP artifacts (model, demo, data) are a foundational toolkit for practitioners in the space and a template for other verticalized CLIP models (<https://github.com/patrick-johnnyh/fashion-clip>).

We believe that our methods and results are interesting not just for the fashion industry but broadly speaking for the ever-expanding industry of online retail, as our artifacts, use cases and benchmarks might serve as a blueprint for other vertical-specific applications of large multi-modal models. Finally, adding to the industry significance of the work, the evaluation in “Grounding and Compositionality” section is new in the context of CLIP-like models, and we believe it may be of independent interest for future work in NLP. As a matter of fact, showcasing a practical thread connecting generalization and latent space interpretation to industrial scalability may be the most interesting contribution of FashionCLIP.

## Results

In this section, we detail the performance of FashionCLIP over a range of tasks, demonstrating the efficacy of domain adaptation and the applicability of CLIP-like models to fashion. Details on the training and on the evaluation are available in the “Methods” Section. We leverage a variety of in-domain and out of domain datasets, with varying degrees of similarity: **TEST** is the test set from Farfetch containing 20k products; **HOUT-C** is the dataset containing a category which we excluded from training; **HOUT-B** is the dataset containing two brands which were excluded from training; **STLE** is a merchandising dataset from Farfetch; **KAGL** is a subset of<sup>50</sup>, where each product has a white background image, a caption, and a category; **F-MNIST**<sup>51</sup> contains 10,000 gray-scale images from 10 product classes; **DEEP**<sup>52</sup> contains 4000 product images that are non-standardized (i.e., contain humans) from 50 categories. An overview of image and textual data offered by Farfetch (**TEST**, **HOUT-C**, **HOUT-B**, **STLE**), **KAGL**, **F-MNIST** and **DEEP** can be found in Fig. 3. Our extensive benchmarks and evaluations answer two research questions quantitatively: can domain-specific knowledge improve CLIP understanding of an industry (Fig. 5) and, if yes, does that knowledge translate across different use cases and datasets?





**Multi-modal retrieval.** The Multi-modal Retrieval task is described as follows: given a textual description and a set of images, we ask the model to find the image related to that description. For example, a product retrieval task entails matching a product description (e.g., “a red polo for men”) and a photo of it in a catalog.

Multi-modal retrieval is possible due to the optimization objective of FashionCLIP which aligns the language and image latent spaces (see Fig. 1). We test FashionCLIP on multi-modal retrieval to assess the benefits of domain-specific fine-tuning on real-world product search.

Our benchmark takes as input a product description from the catalog’s *test set* and asks models to rank product images corresponding to the caption—the gold standard is the image associated with the product. We extract the ranking using embedding similarities: FashionCLIP performs the dot product between the input caption embedding and each image vector embedding obtained via  $T_{\theta_T}(\cdot)$  and  $I_{\theta_I}(\cdot)$  respectively and returns a rank based on descending order. We use  $HITS@5$ <sup>53</sup> (Hit Rate @  $k = 5$ ) and  $MRR$ <sup>54</sup> (Mean Reciprocal Rank) as our metrics. Table 1 compares FashionCLIP against non-domain specific CLIP on different heldout test sets and shows how fine-tuning significantly improves the understanding of our target domain.

We also perform extensive qualitative tests comparing FashionCLIP with the production search engine presently employed in the catalog. Fig. 4 shows a case of particular interest for product search: in this example, visual concepts do not belong to the fashion domain and are not available in the caption. The first comparison (*left*) shows that FashionCLIP can recover the concept of *tiger* when prompted with “t-shirt with tiger”; for the same query, the search engine retrieves items matching the category, unable to interpret *tiger* based solely on text. The second comparison (*right*) shows that FashionCLIP can interpret *a cat* from a stylized, partially occluded drawing. In contrast, the search engine fails to generalize beyond the captions explicitly containing the string “cat”. Finally, visualizing the learnt embeddings (Fig. 5) also helps to build an intuition of FashionCLIP’s better conceptual resolution when it comes to the target domain.

**Zero-shot classification.** We replicate CLIP’s original zero-shot classification setup<sup>17</sup>, which allows us to quantitatively assess the transferability of FashionCLIP’s fine-tuned representations to different data distribu-

	Image	Textual Data	
<b>Farfetch</b>		Short Description	Cropped knit cardigan
		Gender	Women
		Brand	M Missoni
		Highlights	light brown/multicolour; mix print; ribbed-knit edge; V-neck; three-quarter length sleeves; ribbed hem; cropped
<b>KAGL</b>		Product Name	Nike Women As Element Ja Pink
		Category	Topwear
		Sub-category	T-Shirts
<b>F-MNIST</b>		Category	T-shirt / Top
<b>DEEP</b>		Product Description	Breton Stripe Henley
		Category	Henley
		Attributes	Long sleeve; Sleeve

**Figure 3.** Sample of data from various datasets used. We observe a range of distributions on both the image and textual modalities. For the image modality, we see a range from “Low resolution, B &W” to “High resolution, In-the-Wild”. For the textual modality, Farfetch offers the best “textual resolution”, while DEEP also has very fashion specific terminology. The **KAGL**, **F-MNIST**, and **DEEP** datasets are publicly available. For more details regarding the data, see the Data Availability Section.



**Figure 4.** Retrieval with non-fashion concepts. Sample results for “t-shirt with tiger” and “t-shirt with cat” from FashionCLIP (green) vs Farfetch production search engine (red).

Model	Dataset	HITS@5	MRR
F-CLIP	TEST	<b>0.66</b>	<b>0.50</b>
CLIP		0.28	0.21
F-CLIP	HOUT-C	<b>0.62</b>	<b>0.47</b>
CLIP		0.33	0.23
F-CLIP	HOUT-B	<b>0.58</b>	<b>0.41</b>
CLIP		0.31	0.22

**Table 1.** Comparing FashionCLIP (F-CLIP) vs CLIP on the multi-modal retrieval task. Best performing models are in bold.

Model	Dataset	F1
F-CLIP	TEST	<b>0.39</b>
CLIP		0.31
F-CLIP	KAGL	<b>0.67</b>
CLIP		0.63
F-CLIP	F-MNIST	<b>0.71</b>
CLIP		0.66
F-CLIP	DEEP	<b>0.47</b>
CLIP		0.45

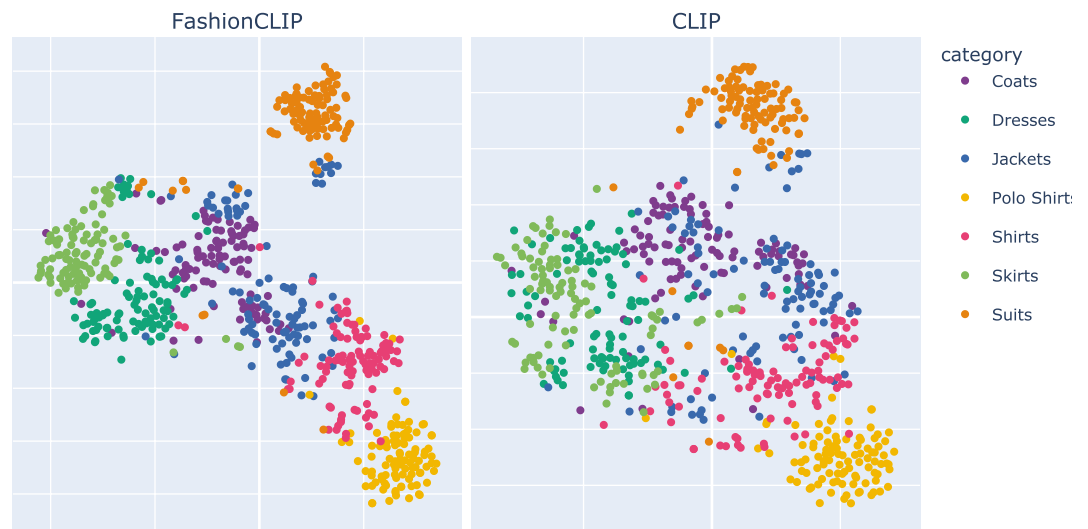
**Table 2.** Comparing the performance of FashionCLIP (F-CLIP) on product classification task over several datasets (F1 is *weighted macro F1*). Best performing models are in bold.

Dataset	F-CLIP	LINEAR	$\Delta F1$
TEST <sub>S</sub>	0.746	0.900	+ 0.154
KAGL <sub>S</sub>	0.764	0.881	+ 0.117
DEEP <sub>S</sub>	0.411	0.444	+ 0.033
F - MNIST <sub>S</sub>	0.781	0.602	- 0.179

**Table 3.** LINEAR classification performance relative to zero-shot on F-CLIP (F1 is *weighted macro F1*).

Model	Man	Woman
Prior	0.24	0.20
F-CLIP	<b>0.36</b>	<b>0.27</b>
CLIP	0.33	0.17

**Table 4.** F1 macro on STLE; Prior classifies using empirical class probabilities. Best performing models are in bold.



**Figure 5.** Comparison of F-CLIP and CLIP Image Space T-SNE projection. We observe better clustering (0.115 vs 0.0745 silhouette score<sup>58</sup>) in F-CLIP for categories such as Shirts, Skirts and Dresses, where the products form a denser cluster with less overlap between categories, suggesting that the F-CLIP latent space is better tuned for fashion concepts.

tions from the same vertical (i.e. Fashion). The model generates *one* image embedding for the product image, and *k* text embeddings, one for each of the labels in the classification scheme (e.g., “shoes”, “shirt”). The predicted label is the one that is closer (measured via dot product) to the image in the model’s vector space. We use *weighted macro F1*<sup>55</sup> as the performance metric. Table 2 summarizes the results of different SOTA benchmarks. On all the tested benchmarks, FashionCLIP is superior to CLIP, a result which suggests that domain-specific fine-tuning is indeed useful in-domain and that it generalizes to other, completely unseen datasets.

Furthermore, we set out to investigate the “cheating hypothesis” on our domain-specific model, i.e., the hypothesis that supervised models do not generalize as well as CLIP because they fit spurious features unique to each dataset. We freeze the image encoder from FashionCLIP and fine-tune a linear classifier, LINEAR, over the embeddings generated on a subset of categories (47) from the validation set from Farfetch. We run benchmarks on TEST<sub>S</sub>, KAGL<sub>S</sub>, F-MNIST<sub>S</sub> and DEEP<sub>S</sub>, sub-sampled versions of the respective datasets. Where labels are different, we adapt LINEAR to the labels by pooling the scores from relevant classes. We compare this to zero-shot performance, using the original labels to generate the text embeddings.

Table 3 reports our findings, which are partially similar to those from CLIP<sup>17</sup>. Given that F-MNIST is very different from TEST—comparable, for example, to CIFAR-100<sup>56</sup> vs. ImageNet<sup>57</sup>—the decrease in performance may be an indication of cheating. However, LINEAR performs well on the other datasets, with the biggest gain for KAGL, whose product image most resembles those in TEST (i.e., high-resolution items on a white background). Compared to the original setting<sup>17</sup>, one may argue that the supervised model has an easier job in our case: much fewer categories ( $10^1$  vs.  $10^3$ ) and relatively homogeneous items, F-MNIST aside.

While we leave the investigation of fashion classification in more ecological settings as future work, our results contain actionable insights for real-world deployments. In particular, supervised classifiers still require a good deal of manual intervention even for similar datasets, and they are utterly unusable on neighboring yet different problems. Table 4 reports performance on STLE divided by Man- and Woman-related items. Products in the dataset still come from Farfetch, but labels are manually assigned by merchandisers and are orthogonal to the taxonomy (*classic, streetwear, edgy* vs. *shoes, hats, bags*). The versatility afforded by language supervision allows zero-shot models to tackle the challenge by simple prompt engineering (“an item in *classic* style”); in contrast, supervised models would require a new training and evaluation pipeline. As emphasized above, learning general fashion concepts is the main motivation behind *this* work: while specific, supervised pipelines may still be the best choice for specific problems, they are no longer the only viable option in multi-tasks scenarios thanks to the advent of large-scale models such as FashionCLIP. Although no single answer can fit all the use cases, we wish to encourage data-driven decision-making by charting all the options and providing cost and performance assessments.

**Grounding and compositionality.** As argued in the *Introduction*, given that we are interested in establishing a connection between generality and scalability through large multi-modal models, it is important to further evaluate the quality of the learned representations. While the question of whether FashionCLIP *learns* fashion has been addressed quantitatively above, we are also interested in evaluating the model from a broader theoretical perspective of language understanding, offering a glimpse of the extent of FashionCLIP’s “true” generalization capabilities, *ala* “infinite use of finite means”<sup>59</sup>.

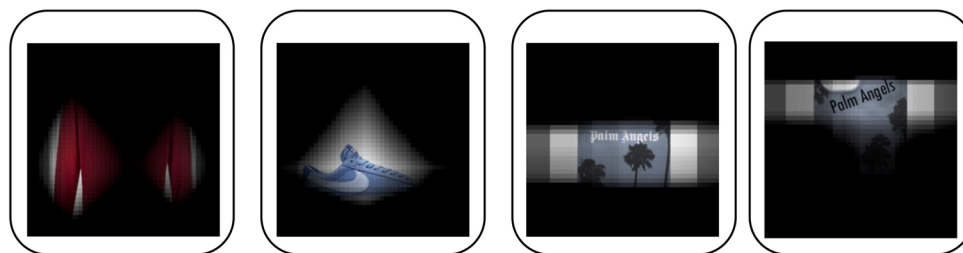
The literature on language compositionality spans centuries: limiting ourselves only to recent work, *grounding* has been explored in connection with efficient learning<sup>60,61</sup>, and “true understanding”<sup>62,63</sup>. Using combinatorial principles to test generalization abilities is a known strategy in toy world<sup>64,65</sup>: we exploit insights from our target domain to operationalize similar principles on *real-world* objects.

In this section, we provide evidence of semantic grounding in FashionCLIP and build on that to offer a preliminary investigation of its compositional abilities. Our analysis starts from two lessons from previous research. First, *localization maps*<sup>66,67</sup> are an effective way to probe the model for *referential* knowledge<sup>68</sup> (we borrow here the referential/inferential distinction from the classic work by Marconi<sup>69</sup>) and visually grounded lexical knowledge. Second, from a linguistic point of view most search queries in fashion have the form of Noun Phrases (NPs)—e.g. “armani dress”. Therefore, the semantics of NP can be considered a good real-world generalization<sup>9,70</sup> for studying FashionCLIP *compositional* and *inferential* abilities.

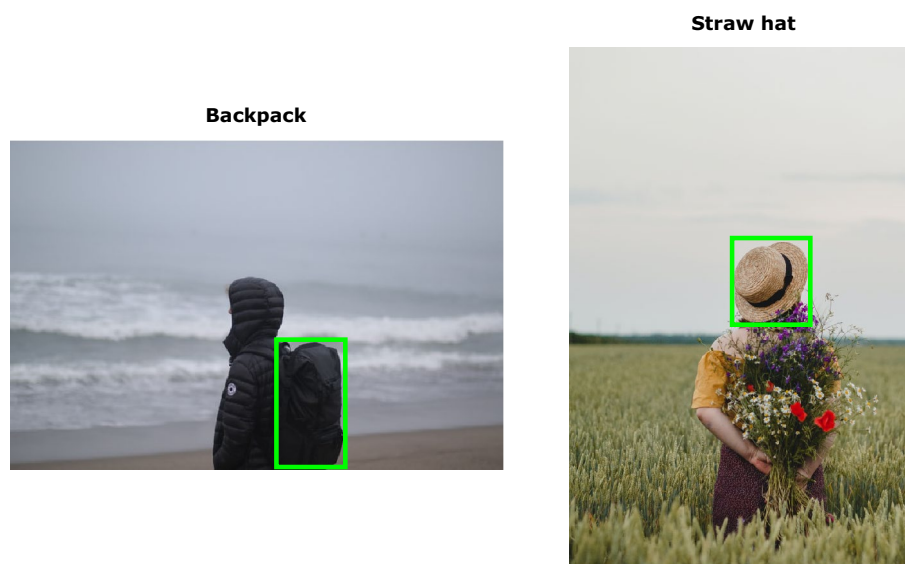
**Grounding.** We probe FashionCLIP for evidence of referential knowledge and investigate its grounding capabilities by utilizing localization maps. We further apply localization maps to the task of zero-shot fashion parsing—a crucial open problem in the industry<sup>71</sup>.

Localization maps are obtained by repeatedly occluding different parts of the image. We then encode each occluded version and measure its distance from the target text in the contrastive space. Intuitively, the farther the image is pushed away by the occlusion, the stronger the linkage was between the removed visual concept and the text and, in turn, the higher its score on the map. Fashion parsing is a specific case of semantic segmentation where bounding box annotations contain clothing items. We extract bounding box annotations (as an approximation of fine-grained segmentation) from localization maps by finding the minimum bounding rectangle of highly activated areas.

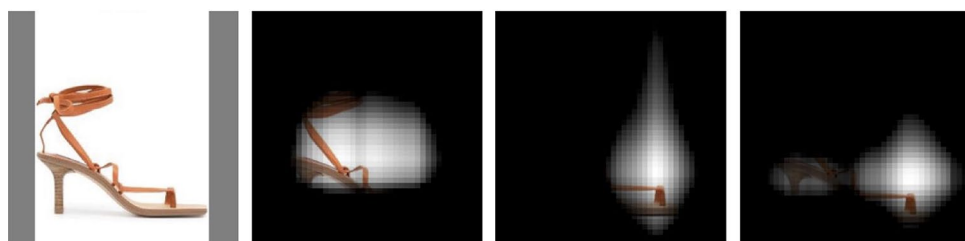
As shown in Figs. 6 and 8, features such as “high heels”, “ankle strap”, “long sleeves” are well represented in FashionCLIP; the model also seems to be very aware of brands, in more or less explicit form. FashionCLIP picks up the abstract logo on *sneakers* (Fig. 6), as well as showing (similar to CLIP) good OCR capabilities, when recognizing a logo as an explicit text string. Fig. 7 shows zero-shot bounding box annotations of some samples in the previously unseen ModaNet<sup>72</sup> dataset. While it is unlikely that zero-shot models could replace specialized segmentation training, we believe that models such as FashionCLIP could provide a cheap way to generate probabilistic labels for weak supervision pipelines.



**Figure 6.** Grounded lexical knowledge. Maps are easy-to-use probes into the model fashion knowledge. *Left to right*: localization map for “long sleeves” on a red polo; sneakers and the map for “Nike”, a phone cover and the map for “Palm Angels”; the same phone cover and map, when the logo is written with an out-of-distribution font in a new spot.



**Figure 7.** Item bounding-box detection. Localization maps can be easily extended to provide zero-shot bounding boxes for items of interest. Green bounding boxes show the predicted locations for fashion concepts “Backpack” (left) and “Straw hat” (right). Images above are taken from the publicly available Unsplash Lite Dataset 1.2.0: FashionCLIP was tested extensively on ModaNet - please reach out to authors for links to those images.



**Figure 8.** Grounding and compositionality. Localization maps for a product retrieved with the query “ankle strap sandals with high heels”: left-to-right, the product, “ankle strap”, “sandals”, “high heels”.

*Compositionality.* Given the preliminary evidence that isolated concepts reliably map onto visual regions, our working hypothesis is that FashionCLIP should exhibit true *inferential* abilities by *composing* such concepts to generate new NPs.

We build on domain knowledge, previous literature<sup>52</sup> and Farfetch’s inventory to probe the model for knowledge of *brands* (e.g. “nike”), *features* (“high heels”), and *drawings* (“keyboard”), manually verifying the text-to-region mapping for each of these concepts via localization maps. Given that these single concepts are grounded





**Figure 9.** Improbable products. By combining fashion features, brands, and items in new ways, we obtain visually realistic products with clear, zero-shot compositional semantics. From left to right: “Nike long dress”, “converse with handles”, “red shoes with black high heel”, “keyboard pochette”.

in regions (Fig. 8), we could leverage this knowledge to generate new images and NPs *systematically*. Crucially, we can assign a defined semantics to a new *brand + object* NP that describes an “improbable object” that has never been seen before (Fig. 9). Improbable objects vary: they may portray odd combinations of concepts, such as a *Nike long dress*, a surreal item, *sneakers with handles*, or an unlikely extension of existing fashion items, such as the *keyboard pochette* (which generalizes the theme first found in J. Mugatu’s *keyboard tie*). A new NP such as “nike dress” would require the visual region corresponding to the word *dress* to contain the visual region of the logo corresponding to the word *nike*.

We supplement our analysis by re-purposing our classification and retrieval pipeline: in the classification task, FashionCLIP achieves an accuracy of 0.74 when asked to pick the improbable label out of a set of credible distractors. The following are examples of test cases:

- *target: NIKE DRESS* (as seen in Fig. 9), *labels*: Nike dress, an Armani dress, a shirt, the flag of Italy, a Gucci dress, a Nike t-shirt;
- *target: BLACK SHOES WITH RED HEEL*, *labels*: black shoes with red heel, black shoes, red shoes with red heel, red shoes with black heel, red shoes, fuchsia shoes, the flag of Italy, sneakers, black sneakers, a bag.
- *target: RED SHOES WITH BLACK HEEL* (as seen in Fig. 9), *labels*: black shoes with red heel, black shoes, red shoes with red heel, red shoes with black heel, red shoes, fuchsia shoes, the flag of Italy, sneakers, black sneakers, a bag.

For the retrieval task, we add the new images to **TEST**, and use the NPs as queries: out of 20k products, the model’s top choice is correct half the time ( $HITS@1 = 0.53$ ), a percentage that quickly rises to 0.82 with  $k = 5$  (as a comparison, CLIP scored  $HITS@1 = 0.51$  and  $HITS@5 = 0.73$ ).

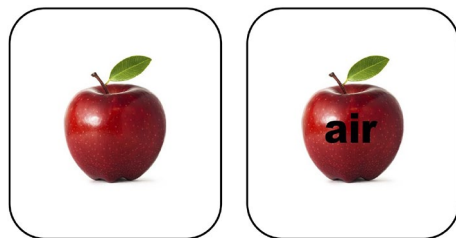
While a full-fledged investigation of compositional abilities is beyond the scope of *this* contribution, FashionCLIP inferences on improbable products suggest the presence of *some* degree of compositionality: important fashion concepts are “identifiable” in the latent space and can be singled out and re-combined into unseen concepts, exhibiting on a small scale the creative generalization we usually associate with symbolic systems<sup>73</sup>. In addition, the ability to distinguish “red shoes with black heel” from “black shoes with red heel” implies knowledge beyond a bag-of-words semantics<sup>74</sup>.

Recent research suggests that CLIP’s compositional capabilities are limited<sup>75</sup>. As shown by our results, restricted domains allow for direct manipulation, without the risk of confounding; indeed, restricted domains may be easier to explore but further investigation is needed to confirm compositional capabilities. Furthermore, as suggested by the use of the MASKClip objective introduced in the ARMANI model<sup>33</sup>, adding explicit visual segmentation may induce better discrimination for certain fashion concepts. While more costly losses are an interesting area at the intersection of grounding and compositionality, given both the narrow generative focus and the magnitude of the improvements in the original paper<sup>33</sup>, their conclusions cannot be readily applied to FashionCLIP. We look forward to performing future research combining insights from generative and discriminative use cases.

## Discussion

FashionCLIP is a domain-adaptation of CLIP, motivated by central use cases in fashion<sup>71</sup>: differently from *task-specific supervised* methods, FashionCLIP does not need a specialized architecture, labeling, and tuning. We extensively verified the flexibility afforded by language supervision, and investigated FashionCLIP’s semantic capabilities on new tasks. Our focus on a specific industry allows not just practical gains but also opens up theoretical possibilities by constraining the domain, which is still large, but also easy to manipulate. By providing quantitative and qualitative evidence that contrastive learning, coupled with a large and diverse dataset, can indeed produce general multi-model industry concepts, we connect theoretical virtues with significant practical gains, and open new possibilities for scaling the horizontal deployment of machine learning systems in an effective way.

As a truly general system, FashionCLIP concepts could be used for many more tasks: for example, multi-modal representations can be features in downstream systems, or directly used for zero-shot recommendations in item-to-item scenarios<sup>76</sup>; classification over arbitrary labels could be used as a fast and scalable labeling mechanism, supporting probabilistic labeling<sup>77</sup> or data generation for multi-modal IR models<sup>78</sup>. While leaving



**Figure 10.** Typographical attack. FashionCLIP correctly identifies the object to the left as an “apple”, but misclassifies the one to the right as “nike air”, as the text acts as a confounder<sup>79</sup>.

LR	Loss	Time(m)	USD	kgCO <sub>2</sub> eq
1e-4	16.0	618	31\$	0.77
1e-5	1.73	617	31\$	0.77
<b>1e-6</b>	<b>2.83</b>	<b>621</b>	<b>31\$</b>	<b>0.78</b>

**Table 5.** Comparing training time, performance, costs, and carbon emission on variants of the FashionCLIP architecture on the *Farfetch* catalog. Cost is calculated with the AWS pricing for a *p3.2xlarge*; estimations were conducted using the Machine Learning Impact calculator<sup>84</sup>. Model used for testing in bold.

this (and many other themes) to future iterations, we do believe *this* work—with its artifacts and methodology—to be a first, rounded assessment of the great potential of general, transferable, multi-modal concepts for digital commerce.

The authors are aware of the risks of multi-modal CLIP-like models in production associated with their limited robustness, as well as general issues with biases in large language models pre-trained at scale. In particular, we acknowledge that the risk of adversarial attacks on multi-modal models is an area of active research<sup>80,81</sup>. To the limits of our knowledge, we have no reason to believe that FashionCLIP introduces any *additional* risk when compared to the original CLIP. As with the original model, it should be noted that FashionCLIP appears to be susceptible to “typographical attacks” (Fig. 10). No datasets used for training or testing contain PII and/or other sensitive user data.

## Methods

**Training dataset.** *Farfetch* made available for the first time an English dataset comprising over 800 k fashion products, with more than 3k brands across dozens of object types. Compared to other large fashion datasets, our dataset is significantly more complete than DeepFashion<sup>52</sup>, which lacks detailed text descriptions, and even larger than CM-Fashion<sup>33</sup>, which has been collected without any direct involvement by *Farfetch*. Items are organized in hierarchical trees, producing a three-layer taxonomy: for example, *trees* could be something like *Clothing*> *Dresses*> *Day Dresses* or *Clothing*> *Coats*> *Parkas*, for a total of 800+ trees. As input for the image encoder, we use the standard product image, which is a picture of the item over a white background, with no humans (images follow a specific set of rules regarding the placement of the item, lights of the photo, etc., designed to highlight the item’s features); as for the text, *Farfetch* has two types of text, *highlight* (e.g., “stripes”, “long sleeves”, “Armani”) and a *short description* (“80s styled t-shirt”). See Fig. 3 for an example.

We create a training, validation, and test set from the catalog by random sampling products. Our final training and validation sets comprise 700 k and 50 k products respectively from 188 categories.

**Training pipeline.** We apply fine-tuning starting from the pre-trained CLIP with the following parameters: we use Adam Optimizer with betas in (0.9, 0.98), epsilon of 1e-6 and weight decay equal to 0.2 and three different learning rates [1e-4, 1e-5, 1e-6]. We train the models for 4 epochs, evaluate every 500 steps and select the model with the lowest validation loss for each configuration (Table 5, model selected in bold). In our preliminary tests, the model with the lowest validation loss overall did not generalize the best in the zero-shot setting. This poses an interesting question, left for future work, of how to fine-tune these large pre-trained models without losing in generalization. The pipeline has been implemented with Metaflow<sup>82</sup>, with training executed remotely on cloud GPUs; experiment tracking was provided by Comet<sup>83</sup>.

**Testing datasets.** We prepare the following datasets for testing purposes and to further gauge the potential impact of the model in production at scale. **TEST** is the test set from *Farfetch* containing 20k products; **HOUT-C** is the dataset containing a category which we excluded from training (*Performance Tops*), for a total of 1.5k items; **HOUT-B** is the dataset containing two brands which were excluded from training, for a total of 1.7k items; **STLE** is a merchandising dataset from *Farfetch*, completely independent from the catalog, that classifies 7749 items across 6 styles for gender women and 4 styles for gender men; example of styles are *Classic* and

Streetwear and each item may belong to more than one style; **KAGL** is a subset of<sup>50</sup>, where each product has a white background image, a caption, and a category, for a total of 9990 items over 62 categories; **F-MNIST**<sup>51</sup> contains 10,000 gray-scale images from 10 product classes, with pixel intensity inverted to obtain images with white background (note that these images have a size of  $24 \times 24$  thus showing much less details than the images on which the models have been trained on). **DEEP**<sup>52</sup> contains 4000 product images that are non-standardized (i.e. contains humans) from 50 categories.

**Training FashionCLIP.** We re-purpose the CLIP main architecture<sup>17</sup>, which we describe briefly in the *Introduction* for the sake of completeness. In the end, we obtain a multi-modal space where images and texts are jointly projected and learned: if training has been successful, we expect that, for example, the textual embedding for the string “red long dress” is actually similar (as measured by the dot product) to the image embeddings of red dresses. Table 5 shows training time, performance, and costs.

### Data availability

The **KAGL**, **F-MNIST**, and **DEEP** datasets are publicly available. The Farfetch dataset is scheduled to be released in the near future. As part of the ongoing mission to help the retail space leverage the latest A.I. techniques and to promote multidisciplinary research in data science across industries, Farfetch is working to finalize the release of the dataset used in this study under a research-friendly license. Please check <https://github.com/Farfetch> for updates on the data release, and reach out to the authors for preliminary inquiries.

Received: 26 May 2022; Accepted: 25 October 2022

Published online: 08 November 2022

### References

1. Cramer-Flood, E. Global Ecommerce 2020. Ecommerce Decelerates amid Global Retail Contraction but Remains a Bright Spot. (2020).
2. McKinsey. The state of Fashion 2019. (2019).
3. de Souza Pereira Moreira, G., Jannach, D. & da Cunha, A. M. On the importance of news content representation in hybrid neural session-based recommender systems. In *INRA@RecSys* (2019).
4. Guo, M. *et al.* Deep learning-based online alternative product recommendations at scale. In *Proceedings of The 3rd Workshop on e-Commerce and NLP* 19–23 (Association for Computational Linguistics, Seattle, WA, USA, 2020). <https://doi.org/10.18653/v1/2020.ecnlp-1.3>.
5. Goncalves, D. *et al.* The importance of brand affinity in luxury fashion recommendations. In *Recommender Systems in Fashion and Retail* (eds Dokoohaki, N. *et al.*) 3–19 (Springer International Publishing, Cham, 2021).
6. Ai, Q. & Narayanan, R. L. Model-agnostic vs. model-intrinsic interpretability for explainable product search. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management, CIKM '21* 5–15, <https://doi.org/10.1145/3459637.3482276> (Association for Computing Machinery, New York, NY, USA, 2021).
7. Chen, L., Chou, H., Xia, Y. & Miyake, H. Multimodal item categorization fully based on transformer. In *Proceedings of The 4th Workshop on e-Commerce and NLP* 111–115 (Association for Computational Linguistics, Online, 2021). <https://doi.org/10.18653/v1/2021.ecnlp-1.13>.
8. Tsagakias, M., King, T. H., Kallumadi, S., Murdock, V. & de Rijke, M. Challenges and research opportunities in ecommerce search and recommendations. *SIGIR Forum* **54**, 1–23 (2021).
9. Bianchi, F., Tagliabue, J. & Yu, B. Query2Prod2Vec: Grounded word embeddings for eCommerce. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers* 154–162 (Association for Computational Linguistics, Online, 2021). <https://doi.org/10.18653/v1/2021.naacl-industry.20>
10. Tagliabue, J. & Yu, B. Shopping in the multiverse: A counterfactual approach to in-session attribution. In *Proceedings of the SIGIR 2020 Workshop on eCommerce (ECOM 20)* (2020).
11. Sculley, D. *et al.* Hidden technical debt in machine learning systems. In *NIPS* (2015).
12. Tagliabue, J. You do not need a bigger boat: Recommendations at reasonable scale in a (mostly) serverless and open stack. In *Fifteenth ACM Conference on Recommender Systems, RecSys '21* 598–600 (Association for Computing Machinery, New York, NY, USA, 2021). <https://doi.org/10.1145/3460231.3474604>.
13. Techcrunch. Algolia finds \$110M from Accel and Salesforce (2019).
14. Techcrunch. Bloomreach raises \$150M on \$900M valuation and acquires Exponea (2021).
15. Techcrunch. Lucidworks raises \$100M to expand in AI finds (2019).
16. Marotta, S. Canada's Latest Tech Public Debut Swings Amid Soft IPOs (2021). <https://www.bloomberg.com/news/articles/2021-11-25/canada-slatest-tech-public-debut-swings-amid-slew-of-soft-ipos>. Accessed 04-Nov-2022.
17. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *ICML* (2021).
18. Shankar, S. Thoughts on MLEngineering After a Year of my PhD—shreya-shankar.com. (2022). <https://www.shreyashankar.com/phd-year-one/>. Accessed 04-Nov-2022.
19. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *CoRR arXiv:2108.07258* (2021).
20. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
21. Grill, J.-B. *et al.* Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
22. Iyer, D., Guu, K., Lansing, L. & Jurafsky, D. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4859–4870 (Association for Computational Linguistics, Online, 2020). <https://doi.org/10.18653/v1/2020.acl-main.439>.
23. Su, Y. *et al.* A contrastive framework for neural text generation. arXiv preprint [arXiv:2202.06417](https://arxiv.org/abs/2202.06417) (2022).
24. Jia, C. *et al.* Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* 4904–4916 (PMLR, 2021).
25. Mu, N., Kirillov, A., Wagner, D. & Xie, S. Slip: Self-supervision meets language-image pre-training. arXiv preprint [arXiv:2112.12750](https://arxiv.org/abs/2112.12750) (2021).
26. Schneider, S., Baevski, A., Collobert, R. & Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *Proc. Interspeech* **2019**, 3465–3469 (2019).
27. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020).

28. Nagrani, A. *et al.* Learning audio-video modalities from image captions. arXiv preprint [arXiv:2204.00679](https://arxiv.org/abs/2204.00679) (2022).
29. Shvetsova, N. *et al.* Everything at once—multi-modal fusion transformer for video retrieval. arXiv preprint [arXiv:2112.04446](https://arxiv.org/abs/2112.04446) (2021).
30. Minderer, M. *et al.* Simple open-vocabulary object detection with vision transformers. arXiv preprint [arXiv:2205.06230](https://arxiv.org/abs/2205.06230) (2022).
31. Liu, H. *et al.* Cma-clip: Cross-modality attention clip for image-text classification. arXiv preprint [arXiv:2112.03562](https://arxiv.org/abs/2112.03562) (2021).
32. Sevegnani, K. *et al.* Contrastive learning for interactive recommendation in fashion. CoRR [arXiv:2207.12033](https://arxiv.org/abs/2207.12033), <https://doi.org/10.48550/arXiv.2207.12033> (2022).
33. Zhang, X. *et al.* Armani: Part-level garment-text alignment for unified cross-modal fashion design. arXiv preprint [arXiv:2208.05621](https://arxiv.org/abs/2208.05621) (2022).
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv preprint [arXiv:2204.06125](https://arxiv.org/abs/2204.06125) (2022).
35. Kong, C., Jeon, D., Kwon, O. & Kwak, N. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. arXiv preprint [arXiv:2210.05872](https://arxiv.org/abs/2210.05872) (2022).
36. Gupta, V., Karnick, H., Bansal, A. & Jhala, P. Product classification in e-commerce using distributional semantics. In *COLING* (2016).
37. Fu, J. *et al.* Cma-clip: Cross-modality attention clip for text-image classification. In *IEEE ICIP 2022* (2022).
38. Commerce, B. How Ecommerce Site Search Can Create a Competitive Advantage. (2021). <https://www.bigcommerce.com/articles/e-commerce/site-search/#the-effectiveness-of-e-commerce-site-search->. Accessed 04-Nov-2022.
39. Alaimo, D. 87% of shoppers now begin product searches online. (2018).
40. Robertson, S. & Zaragoza, H. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.* **3**, 333–389. <https://doi.org/10.1561/15000000019> (2009).
41. Gillick, D., Presta, A. & Tomar, G. S. End-to-end retrieval in continuous space. arXiv preprint [arXiv:1811.08008](https://arxiv.org/abs/1811.08008) (2018).
42. Izacard, G. *et al.* Towards unsupervised dense information retrieval with contrastive learning. arXiv preprint [arXiv:2112.09118](https://arxiv.org/abs/2112.09118) (2021).
43. Hu, Y., Da, Q., Zeng, A., Yu, Y. & Xu, Y. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '18*, 368–377 (Association for Computing Machinery, New York, NY, USA, 2018). <https://doi.org/10.1145/3219819.3219846>.
44. Chen, L. & Miyake, H. Label-guided learning for item categorization in e-commerce. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers 296–303* (Association for Computational Linguistics, Online, 2021). <https://doi.org/10.18653/v1/2021.naacl-industry.37>.
45. Costin, A. M., Eastman, C. & Issa, R. R. A. The Need for Taxonomies in the Ontological Approach for Interoperability of Heterogeneous Information Models 9–17 (2017). <https://ascelibrary.org/doi/pdf/10.1061/9780784480830.002>.
46. McDowell, M. Taxonomy is the new fashion-tech essential. (2020). <https://www.voguebusiness.com/technology/taxonomy-is-the-new-fashion-tech-essential-theys>. Accessed 04-Nov-2022.
47. Feizi, S., Singla, S. Salient imagenet: how to discover spurious features in deep learning? In *International Conference on Learning Representations* (2021).
48. Chia, P. J. Fashion CLIP. <https://github.com/patrickjohncyh/fashion-clip/> (2022). [Online; accessed 15-September-2022].
49. Chia, P. J. GradREC. <https://github.com/patrickjohncyh/gradient-recs/> (2022). [Online; accessed 15-September-2022].
50. Aggarwal, P. Fashion Product Images Dataset. (2020).
51. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. [arXiv:cs.LG/1708.07747](https://arxiv.org/abs/1708.07747) (2017).
52. Liu, Z., Luo, P., Qiu, S., Wang, X. & Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
53. Cremonesi, P., Koren, Y. & Turrin, R. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, 39–46 (Association for Computing Machinery, New York, NY, USA, 2010). <https://doi.org/10.1145/1864708.1864721>.
54. Craswell, N. *Mean Reciprocal Rank 1703–1703* (Springer, US, Boston, MA, 2009).
55. Manning, C. D., Raghavan, P. & Schütze, H. *An Introduction to Information Retrieval* Online. (Cambridge University Press, Cambridge, UK, 2009).
56. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images* Tech Rep, (2009).
57. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
58. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) (1987).
59. von Humboldt, W. On language: On the diversity of human language construction and its influence on the mental development of the human species. (1836/1999).
60. Yu, H., Zhang, H. & Xu, W. Interactive grounded language acquisition and generalization in a 2d world. In *ICLR* (2018).
61. Chevalier-Boisvert, M. *et al.* BabyAI: A platform to study the sample efficiency of grounded language learning. In *ICLR* (2019).
62. Bender, E. M. & Koller, A. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 5185–5198 (Association for Computational Linguistics, Online, 2020). <https://doi.org/10.18653/v1/2020.acl-main.463>.
63. Merrill, W. C., Goldberg, Y., Schwartz, R. & Smith, N. A. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand?. *Trans. Assoc. Comput. Linguist.* **9**, 1047–1060 (2021).
64. Chollet, F. On the measure of intelligence. ArXiv [arXiv:1911.01547](https://arxiv.org/abs/1911.01547) (2019).
65. Gandhi, K., Stojnić, G., Lake, B. M. & Dillon, M. R. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. [arXiv:cs.LG/2102.11938](https://arxiv.org/abs/2102.11938) (2021).
66. Fong, R. C. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).
67. Covert, I., Lundberg, S. & Lee, S. -I. Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.* **22**, 1–90 (2021).
68. Bianchi, F. *et al.* Contrastive language-image pre-training for the italian language. arXiv preprint [arXiv:2108.08688](https://arxiv.org/abs/2108.08688) (2021).
69. Marconi, D. *Lexical Competence* (MIT Press, Cambridge, MA, 1997).
70. Bianchi, F., Greco, C. & Tagliabue, J. Language in a (search) box: Grounding language learning in real-world human-machine interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4409–4415 (Association for Computational Linguistics, Online, 2021). <https://doi.org/10.18653/v1/2021.naacl-main.348>.
71. Cheng, W.-H., Song, S., Chen, C.-Y., Hidayati, S. C. & Liu, J. Fashion meets computer vision: A survey. *ACM Comput. Surv.* <https://doi.org/10.1145/3447239> (2021).
72. Zheng, S., Yang, F., Kiapour, M. H. & Piramuthu, R. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia* 1670–1678 (2018).
73. Chierchia, G. & McConnell-Ginet, S. *Meaning and Grammar: An Introduction to Semantics* 2nd edn. (MIT Press, Cambridge, MA, USA, 2000).

74. Pham, T. M., Bui, T., Mai, L. & Nguyen, A. M. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? [arXiv:cs.LG/2012.15180](https://arxiv.org/abs/cs.LG/2012.15180) (2021).
75. Thrush, T. *et al.* Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR* (2022).
76. Chia, P. J., Tagliabue, J., Bianchi, F., Greco, C. & Goncalves, D. “does it come in black?” clip-like models are zero-shot recommenders. In *Proceedings of The 5th Workshop on e-Commerce and NLP* (Association for Computational Linguistics, 2022).
77. Ratner, A. J. *et al.* Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* Vol. 11 3, 269–282 (2017).
78. Yu, B., Tagliabue, J., Greco, C. & Bianchi, F. “an image is worth a thousand features”: Scalable product representations for in-session type-ahead personalization. In *Companion Proceedings of the Web Conference 2020, WWW '20* 461–470 (Association for Computing Machinery, New York, NY, USA, 2020). <https://doi.org/10.1145/3366424.3386198>.
79. Vincent, J. OpenAI’s state-of-the-art machine vision AI is fooled by handwritten notes. (2021). <https://www.theverge.com/2021/3/8/22319173/openai-machine-visionadversarial-typographic-attack-clip-multimodal-neuron>. Accessed 04-Nov-2022.
80. Noever, D. A. & Noever, S. E. M. Reading isn’t believing: Adversarial attacks on multi-modal neurons. arXiv preprint [arXiv:2103.10480](https://arxiv.org/abs/2103.10480) (2021).
81. Yu, Y., Lee, H. J., Kim, B. C., Kim, J. U. & Ro, Y. M. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. [arXiv:2005.10987](https://arxiv.org/abs/2005.10987) (2020).
82. Berg, D. *et al.* Open-Sourcing Metaflow, a Human-Centric Framework for Data Science (2019). <https://netflixtechblog.com/open-sourcing-metaflow-a-human-centricframework-for-data-science-fa72e04a5d9>. Accessed 04-Nov-2022.
83. Comet.ML. Comet.ML home page (2021).
84. Lacoste, A., Luccioni, A., Schmidt, V. & Dandres, T. Quantifying the carbon emissions of machine learning. arXiv preprint [arXiv:1910.09700](https://arxiv.org/abs/1910.09700) (2019).

### Author contributions

FashionCLIP was started by JT and FB, PC and GA led implementation and experiments; DG and ARM prepared the dataset, performed EDA, and provided domain knowledge; ST and CG helped with fine-tuning, model evaluation, and research background. Everybody contributed to the final draft. JT and FB acted as senior PIs for the project.

### Competing interests

GA is a member of the Bocconi Institute of Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit. FB was a BIDSA and DMI member during the project. JT and CG were at Coveo Labs for the initial phase of the project. The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to P.J.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022