computers

MDPI

Article

Learning Explainable Disentangled Representations of E-Commerce Data by Aligning Their Visual and Textual Attributes

Katrien Laenen * and **Marie-Francine Moens**

Human-Computer Interaction group, KU Leuven, 3000 Leuven, Belgium

* Correspondence: katrien.laenen@kuleuven.be

Abstract: Understanding multimedia content remains a challenging problem in e-commerce search and recommendation applications. It is difficult to obtain item representations that capture the relevant product attributes since these product attributes are fine-grained and scattered across product images with huge visual variations and product descriptions that are noisy and incomplete. In addition, the interpretability and explainability of item representations have become more important in order to make e-commerce applications more intelligible to humans. Multimodal disentangled representation learning, where the independent generative factors of multimodal data are identified and encoded in separate subsets of features in the feature space, is an interesting research area to explore in an e-commerce context given the benefits of the resulting disentangled representations such as generalizability, robustness and interpretability. However, the characteristics of real-word e-commerce data, such as the extensive visual variation, noisy and incomplete product descriptions, and complex cross-modal relations of vision and language, together with the lack of an automatic interpretation method to explain the contents of disentangled representations, means that current approaches for multimodal disentangled representation learning do not suffice for e-commerce data. Therefore, in this work, we design an explainable variational autoencoder framework (E-VAE) which leverages visual and textual item data to obtain disentangled item representations by jointly learning to disentangle the visual item data and to infer a two-level alignment of the visual and textual item data in a multimodal disentangled space. As such, E-VAE tackles the main challenges in disentangling multimodal e-commerce data. Firstly, with the weak supervision of the two-level alignment our E-VAE learns to steer the disentanglement process towards discovering the relevant factors of variations in the multimodal data and to ignore irrelevant visual variations which are abundant in e-commerce data. Secondly, to the best of our knowledge our E-VAE is the first VAE-based framework that has an automatic interpretation mechanism that allows to explain the components of the disentangled item representations with text. With our textual explanations we provide insight in the quality of the disentanglement. Furthermore, we demonstrate that with our explainable disentangled item representations we achieve state-of-the-art outfit recommendation results on the Polyvore Outfits dataset and report new state-of-the-art cross-modal search results on the Amazon Dresses dataset.

Keywords: explainability; disentangled representation; multimodal representation; cross-modal search; outfit recommendation

check for updates

Citation: Laenen, K.; Moens, M.-F. Learning Explainable Disentangled Representations of E-Commerce Data by Aligning Their Visual and Textual Attributes. *Computers* **2022**, *11*, 182. https://doi.org/10.3390/computers11120182

Academic Editor: Paolo Bellavista

Received: 17 October 2022
Accepted: 6 December 2022
Published: 10 December 2022

1. Introduction

Product understanding for multimedia content is a core problem in fashion e-commerce search and recommendation. To be able to retrieve and recommend suitable products we require representations of those products that capture all their relevant fine-grained product attributes. The challenge lies in the fact that relevant fine-grained product attributes are scattered across the visual and textual e-commerce data and must be correctly detected, recognized and fused into multimodal representations. Current methods for multimodal representation learning such as visual-linguistic transformers usually create representations that are entangled and suffer from feature correlation and duplication. Furthermore, as visual-linguistic transformers are purely likelihood-based they lead to spurious correlations which pose several problems. Firstly, spurious correlations result in models that can be unfair to particular subgroups in a population. Secondly, models suffering from spurious correlations are inexplicable because they are often right for the wrong reason. In addition, spurious correlations harm the generalization ability and can lead to errors in out-of-distribution settings. Because of these problems such models can be perceived as untrustworthy.

Explanations are a promising mechanism for promoting fairness, transparency and trust. Therefore, there is an increasing interest in developing eXplainable artificial intelligence (XAI) systems that are interpretable and explainable while maintaining the same level of performance [1]. One promising technique is multimodal disentangled representation learning which produces disentangled representations that are less sensitive to misleading correlations in the training data and have interesting properties such as generalizability, robustness and interpretability. In multimodal disentangled representation learning, the multimodal product data are unraveled to discover the small set of independent generative factors from which a clothing item is assembled. These independent generative factors or so-called factors of variation are groups of product attributes such as colors (e.g., red, navy, khaki), neckline shapes (e.g., crew neck, V-neck, boat neck) or fits (e.g., loose, relaxed, regular). Disentangled representations are interpretable by design since each feature or subset of features corresponds with exactly one factor of variation. However, an automatic method to identify which factor of variation is encoded in which subset of features is still lacking. Currently, a feature dimension is given an interpretation by gradually altering the value of that dimension while keeping other dimensions fixed and then checking which product aspect is changing. However, this lack of an automatic interpretation method prevents disentangled representations from being used to their fullest potential such as for explainable search and recommendation. In addition, disentangling real-world multimodal e-commerce data is very challenging. First, the relevant product attributes are scattered across the visual and textual modalities. Some are visible only in the product image or are mentioned only in the product description, while others are expressed through both modalities. Secondly, on the image side, whether product attributes are discovered as factors of variation depends on their saliency and granularity. Often, product attributes are much less salient than other visual variations in e-commerce images such as the visual appearance or pose of the human models which are irrelevant when modelling fashion items. It is not straightforward how to learn to ignore these salient but irrelevant visual variations in favor of less salient but relevant visual variations. Furthermore, the factors of variation for clothing items are product attribute groups with varying granularity. Some attribute groups such as color or fit usually take up a larger portion of the image, while others such as neckline shape or fasteners are very fine-grained and only have subtle differences. Overall, the visual variation in e-commerce images is immense resulting in a huge search space that should be explored to discover the relevant factors of variation. Thirdly, on the text side the product descriptions are incomplete, meaning they refer to some but not all product attributes that are instances of relevant factors of variation to be discovered. In addition, they contain noise such as washing instructions which should be ignored. Finally, for real-world e-commerce data, we usually do not have any ground truth data except for the correspondence of the product images and descriptions. Hence, we do not know how to segment the product images and descriptions into relevant product attributes nor are we aware of the semantic correspondences of the visual and textual product attributes.

In this work we propose E-VAE, an explainable variational autoencoder (VAE) which infers a multimodal disentangled space where the visual and textual item data are aligned at two levels. At a coarse-grained level, we align product images with textual attributes from the corresponding product description that refer to factors of variation. The alignment of the fashion images and textual product attributes steers the disentanglement process towards discovering generative factors that correspond with natural concepts that humans use to describe and distinguish clothing items and helps to ignore irrelevant factors of variation. At a fine-grained level, we infer the latent semantic correspondences of the textual attributes with the image regions in the multimodal disentangled space. This results in a fine-grained visual contextualization of the textual attributes and facilitates the alignment at the coarse-grained level. Through the coarse-grained alignment of the product images and textual attributes, textual information seeps through to the disentangled representations of the product images, making them multimodal. Hence, we consider the disentangled representations of the product images as our multimodal item representations. In addition, a byproduct of the coarse-grained alignment is that it provides a method to explain the contents of the disentangled multimodal item representations with text. We compare the proposed E-VAE model with state-of-the-art systems for outfit recommendation and cross-modal search. We evaluate their performance on these two tasks as well as their ability to explain the contents of the representations they produce. Our contributions are:

- To the best of our knowledge, we are the first to propose an automatic method to explain the contents of disentangled multimodal item representations with text.
- We disentangle real-world multimodal e-commerce data which is challenging because (i) some attributes are shared and some are complementary between the product image and description which makes effective fusion complex, (ii) the saliency, granularity and visual variation of the product attributes complicates detection and recognition, (iii) the noise and incompleteness of product descriptions makes alignment difficult, and (iv) we lack ground truth data at the product attribute level.
- We show how the weak supervision of the two-level alignment steers the disentanglement process towards discovering factors of variation that humans use to organize, describe and distinguish fashion products and to ignore others which is essential when the visual search space is huge and noisy.
- We demonstrate that our E-VAE creates representations that are explainable while maintaining the same level of performance as the state-of-the-art or surpassing it. More precisely, we achieve state-of-the-art outfit recommendation results on the Polyvore Outfits dataset and new state-of-the-art cross-modal search results on the Amazon Dresses dataset.

The remainder of the paper is structured as follows. In Section 2 we describe related work. Next, Section 3 provides a detailed description of our E-VAE and how we steer and interpret the disentanglement. Then, Section 4 gives an overview of our experimental setup. Results are presented in Section 5. Finally, Section 6 lists the main conclusions and directions for future work.

2. Related Work

In this work, we aim to learn explainable disentangled representations of fashion items from multimodal e-commerce data. This is challenging because relevant product attributes are scattered across the product images and descriptions, the visual variation in the product images is huge, and the product descriptions are noisy and incomplete. Furthermore, explaining the contents of disentangled representations is not straightforward since an automatic interpretation method is still lacking. We review some current works on unimodal and multimodal disentangled representation learning as well as on explainability in e-commerce search and recommender systems.

2.1. Disentangled Representation Learning

The term disentangled representation was first introduced in [2]. The underlying assumption is that, to create generalizable and interpretable semantic representations of data observations, we should unravel (disentangle) the underlying structure of those data. More precisely, given data observations (e.g., images of clothing items or faces) we should find the set of independent generative factors from which these observations are generated (e.g., hair length, hair color, skin tone, eye shape, eye color, etc. for a face). The generative factors or so-called factors of variation are independent, meaning we can change one factor without changing others (e.g., changing the hair color does not affect other facial factors of variation such as hair length or skin tone). A disentangled representation is then a low-dimensional vector that contains all the information about each factor of variation, with each coordinate (or subset of coordinates) containing information about only one such factor. The VAE framework, which was originally proposed in [3,4], is typically used for disentangled representation learning. It is a neural network architecture consisting of an encoder that maps an input observation to a low-dimensional latent space and a decoder that reproduces the input observation given the latent representation. To organize the latent space in such a way that each coordinate (or subset of coordinates) captures only one factor of variation, the VAE encoder learns to output a factorized standard Gaussian distribution over all the latent space dimensions from which the disentangled representation is sampled. VAEs [5,6] have shown promising results for visual data such as the CelebA [7], Faces [8] and Chairs [9] datasets but where the amount, granularity and visual variation of the generative factors is more limited and where there is much less noise compared to fashion e-commerce data. For textual data, VAEs have been less successful due to the problem of posterior collapse, although there are some works on how to mitigate this problem [10,11].

Also in e-commerce VAEs have shown promising results. For instance in [12], the authors learn disentangled representations for users and items based on user behaviour data. They disentangle the user behaviour data into macro latent factors that govern user intentions and micro-latent factors that describe the preference about these intentions. While they solely disentangle user behaviour data, our goal is to disentangle multimodal data, in our case visual and textual item data. We propose to steer the disentanglement process towards finding relevant factors of variation in the multimodal item data by jointly learning to disentangle and learning a two-level alignment of the vision and language in the disentangled space, that is, of the full images and textual attributes and of the image regions and textual attributes. Furthermore, the alignment allows us to interpret the components of the disentangled representations with text, whereas [12] do not allow to interpret the disentangled representations or to generate explanations.

Other neural architectures can be used for disentangled representation learning as well. In [13], the authors use a CNN-based architecture to learn disentangled item representations for item retrieval and outfit recommendation. Their method uses a pre-defined ground truth attribute hierarchy to determine the disentanglement. This requires immense labeling effort from e-retailers as new products arrive every day and fashion seasons, trends and styles change over time. In contrast, E-VAE learns to disentangle without having access to the actual factors of variation nor to complete ground truth attribute labels.

Finally, our work belongs to a recent line of research that integrates weak supervision in disentangled representation learning [14,15]. More precisely, we steer the disentanglement of fashion e-commerce data towards discovering relevant factors of variation through the alignment of the disentangled neural representations with discrete labels in the form of textual attributes from the product descriptions that are instances of generative factors of interest. Furthermore, our approach has the additional advantage that it makes the disentangled representations explainable.

2.2. Multimodal Disentangled Representation Learning

Recently, there has been an increasing interest in multimodal disentangled representation learning. For instance in [16], the authors propose a neural architecture that learns disentangled representations from multi-feedback, that is, both positive (i.e., click) and negative feedback (i.e., unclick and dislike). More precisely, they propose a co-filtering dynamic routing mechanism to capture the complex relations in multi-feedback and also deal with the noise hidden in multi-feedback. In contrast, instead of disentangling multimodal user behaviour data our focus is on disentangling multimodal item data but which are also characterized by complex cross-modal relations and noise. In [17], content-collaborative disentangled user and item representations are inferred from user behaviour data and multimodal item data. First, they try to extract as much of the relevant features from the item content data, which are a concatenation of a coarse-grained visual and textual representation. Next, they discover the remaining necessary features from the user-item interactions. In our work, we disentangle multimodal item data by aligning disentangled representations of product images with sparse representations of textual attributes from e-commerce descriptions in a multimodal disentangled space. Since we operate at the attribute level, our feature disentanglement is more fine-grained than that of [17]. In [18], a deep generative model embeds user content features and user collaborative features in a latent space both with their own encoder module. The collaborative encoder module learns to disentangle the user collaborative features into uncertainty and semantic information about the user. The uncertainty information is passed to a user-dependent channel which determines the amount of information from the latent user content features to fuse with the latent user collaborative features to produce the final latent user embedding. However, their embeddings are not explainable.

Closely related to our work is the work of [19], where a neural architecture is proposed that consists of three VAEs. Given instances of a first modality (e.g., image), the first VAE learns to reconstruct a second modality (e.g., text). The second VAE learns to reconstruct the second modality (e.g., text) based on this modality itself. The third VAE, called the mapper, learns to align the latent distributions of the two VAEs in order to effectively combine the information extracted by both. As a result, the mapper produces disentangled representations of the first modality (e.g., image) that are close to the disentangled representations of the second modality (e.g., text). The proposed architecture is used for image-to-text retrieval. Like us, they also use a form of alignment to align images and texts in a disentangled space. However, their alignment is more coarse-grained than ours and their disentangled representations are not explainable. Also closely related is the work of [20]. They propose a VAE framework to disentangle visual and textual data by inferring a single shared multimodal disentangled space and a private unimodal disentangled space per modality. The shared multimodal space focuses on factors of variation that can be expressed through both modalities, while the private disentangled spaces focus on modality-specific factors of variation. They demonstrate the effectiveness of their method on cross-modal search. Similar to us, they consider the substitutability and complementarity relations of vision and language in the disentanglement process. However, they do not try to infer the latent semantic correspondences at the level of visual and textual attributes to improve the disentanglement and explainability.

2.3. Explainability

For model explainability, existing search and recommender systems most often use attention to highlight certain words or regions that were important to produce the search or recommendation results [21–24]. In [21], an encoder-decoder-based architecture is designed which applies hierarchical co-attention on user and item reviews to capture the deep user-item interactions and output a rating as well as a natural language explanation for a given user-item pair. In [22], natural language explanations are generated based on the user and item features as well as visual features of the item and sentiment features. In both approaches, the textual explanations are generated based on attention on entangled feature representations that are not interpretable and suffer from feature correlation and duplication which negatively affect the explanation. In contrast, our textual explanations are generated for explainable item representations that are disentangled in a small set of independent factors of variation. In [23], users and item image regions are projected into a fine-grained semantic space where the user's preferences towards each visual attribute of the item can be calculated with attention. In [24], the authors propose a neural architecture that learns to attend to the regions in an image showing product attributes that a particular user is interested in, enabling it to visually explain a recommendation. Furthermore, the model uses the user review text as a weak supervision signal to infer which image regions show product attributes the user cares about. While these works provide visual explanations based on entangled item representations that may contain duplicated and highly correlated features obtained with attention-based neural architectures, we provide textual explanations of the contents of disentangled item representations obtained with a VAE-based generative deep neural network. In [25], the authors propose to visually explain VAEs by means of gradient-based attention and show how the attention maps can be used to localize anomalies in images. In this work, we provide a method to interpret the components of the disentangled space by jointly training our E-VAE to learn to disentangle images and to align the disentangled representations with textual phrases.

3. Methodology

Our goal in this work is twofold. First, we want to obtain multimodal disentangled representations of items that we can interpret and explain. Second, we aim to design a method to steer the disentanglement process of real-world fashion e-commerce images towards discovering the relevant factors of variation which are often fine-grained and less salient than irrelevant visual variations. Therefore we propose E-VAE, an explainable VAE-based framework that infers a multimodal disentangled space under the weak supervision of the two-level alignment of the visual and textual item data. The fashion-related phrases in the textual item data denote attributes of different groups (e.g., colors, necklines, sleeve lengths) that humans use to organize, describe and distinguish fashion products and can thus be considered instances of relevant factors of variation. Hence, the alignment steers the disentanglement process in the desired direction. Additionally, it has the advantage that it provides a method to interpret the disentangled representations with text.

Our E-VAE is an extension of β-VAE [5] but with three major changes. First, E-VAE infers disentangled representations that are not only interpretable but also explainable with text. Second, E-VAE is designed to disentangle not unimodal data but instead multimodal item data where some factors of variation are shared and some are complementary between the vision and language data. Third, E-VAE learns to disentangle real-world e-commerce data where the amount, granularity and visual variation of the factors of variation is huge and where there is a lot of noise in both the vision and language data.

Section 3.1 explains how E-VAE disentangles the fashion images. Section 3.2 describes how we make our disentangled space multimodal and our representations explainable through the two-level alignment of the vision and language data.

3.1. Disentanglement

On the one hand, our E-VAE learns to disentangle fashion images. We assume the fashion images are generated from K underlying independent generative factors $f_1, ..., f_K$. We want to model each image v with a latent vector $\boldsymbol{z} \in \mathbb{R}^{d_z}$ where each coordinate z_j or subset of coordinates corresponds with exactly one generative factor f_k (i.e., $d_z \geq K$). Such a latent vector is called a disentangled representation.

3.1.1. Prior

Representations $\boldsymbol{z}$ in the latent space are assumed to follow the prior distribution $p(\boldsymbol{z})$, which is chosen to be a factorized standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Since each latent feature z_j corresponds with a generative factor f_k, this means we believe that each generative factor has a standard normal distribution and that the generative factors are independent.

3.1.2. Decoder

We assume that each image v is generated from latent variables $\boldsymbol{z}$ that correspond with the generative factors $f_1, ..., f_K$ the latter not directly observed. E-VAE defines a joint probability over the images and latent variables $p(v, \boldsymbol{z})$ that describes our data. This joint probability can be decomposed in the likelihood and the prior, i.e., $p(v, \boldsymbol{z}) = p(v|\boldsymbol{z})p(\boldsymbol{z})$. The generative process assumed for our data is thus that we first sample a latent vector $\boldsymbol{z}^{(d)}$ from the prior ($\boldsymbol{z}^{(d)} \sim p(\boldsymbol{z})$) and then the image $v^{(d)}$ from the likelihood ($v^{(d)} \sim p(v|\boldsymbol{z})$). The likelihood $p(v|\boldsymbol{z})$ is modeled by the decoder. The decoder is parameterized with a deep deconvolutional neural network $\mathcal{D}(\boldsymbol{z})$ based on the ResNet50 architecture [26].

3.1.3. Encoder

The goal of E-VAE is to actually find the relevant generative factors in the visual item data. In other words, we want to infer good values for the latent variables given the observed data when calculating the posterior $p(\boldsymbol{z}|v)$. As the computation of the true posterior is intractable, it is approximated with the variational posterior $q(\boldsymbol{z}|v)$ which is modeled by the encoder. The encoder should organize the latent space in a way compatible with the generative process described above. Therefore, the encoder is trained to return the mean and the covariance matrix that describe the Gaussian distribution of the latent features, i.e., $q(\boldsymbol{z}|v) = \prod_{j=1}^{d_z} \mathcal{N}\big(z_j|\mu_j(\hat{\boldsymbol{v}}), \sigma_j^2(\hat{\boldsymbol{v}})\big)$, and which is enforced to be close to a standard Gaussian distribution. Here, $\hat{\boldsymbol{v}}$ is an intermediate representation generated by the encoder which is parameterized with a deep convolutional neural network $\mathcal{E}(v)$ based on the ResNet50 architecture [26] commonly used for image encoding. More precisely, the encoder consists of the convolutional layers of the ResNet50 architecture with one fully-connected layer on top. The fully-connected layer takes the input from the last convolutional layer, flattens it to obtain $\hat{\boldsymbol{v}}$, and outputs the mean $\boldsymbol{\mu}(\hat{\boldsymbol{v}})$ and variance $\boldsymbol{\sigma}^2(\hat{\boldsymbol{v}})$ of the latent variables in the disentangled space.

3.1.4. ELBO

From a probabilistic perspective, encoder $q(\boldsymbol{z}|v)$, decoder $p(v|\boldsymbol{z})$, and the prior $p(\boldsymbol{z})$ interact as follows during training. First, an image is encoded as a distribution over a low-dimensional latent space by the encoder. Next, a point from this distribution is sampled and the image is reconstructed by the decoder. Then, E-VAE is trained to maximize the evidence lower bound (ELBO):

$$\begin{aligned}\mathrm{ELBO} = \frac{1}{D}\sum_{d=1}^{D}\Big(&\mathbb{E}_{q(\boldsymbol{z}^{(d)}|v^{(d)})}\big[\log p(v^{(d)}|\boldsymbol{z}^{(d)})\big] \\ &- \beta \cdot KL\big(q(\boldsymbol{z}^{(d)}|v^{(d)})||p(\boldsymbol{z}^{(d)})\big)\Big),\end{aligned} \tag{1}$$

where D is the number of training instances and hyperparameter $\beta > 1$. The ELBO is composed of two terms. The first term is the log-likelihood of the dth image and encourages the decoder to learn to reconstruct the image. The second term tries to regularize the organisation of the latent space by enforcing that the distributions returned by the encoder are close to the prior, which is a standard Gaussian. From a neural network perspective, encoder $\mathcal{E}(v)$ takes an image v as input and outputs values for the mean $\boldsymbol{\mu}(\hat{\boldsymbol{v}})$ and variance $\boldsymbol{\sigma}^2(\hat{\boldsymbol{v}})$ of the latent variables. Then, the latent representation $\boldsymbol{z}$ of the input image is obtained from the multivariate distribution parameterized by the mean and variance using the reparameterization trick, i.e., $\boldsymbol{z} = \boldsymbol{\mu}(\hat{\boldsymbol{v}}) + \epsilon\boldsymbol{\sigma}(\hat{\boldsymbol{v}})$ with $\epsilon \sim \mathcal{N}(0,1)$. Finally, decoder $\mathcal{D}(\boldsymbol{z})$ takes the latent representation and outputs the reconstructed image $\bar{v}$. The encoder-decoder is trained by minimizing $\mathcal{L}_{ELBO}$:

$$\begin{aligned}\mathcal{L}_{ELBO} = \frac{1}{D}\sum_{d=1}^{D}\Big(&\sum_{k=1,l=1}^{W,H} v_{k,l}^{(d)} \log\big(f(\bar{v}_{k,l}^{(d)})\big) \\ &+\beta \cdot \sum_{j=1}^{d_z} \frac{1}{2} \cdot \big(-\log(\sigma_j^2(\hat{\boldsymbol{v}})) + \sigma_j^2(\hat{\boldsymbol{v}}) + \mu_j(\hat{\boldsymbol{v}})^2 - 1\big)\Big),\end{aligned} \tag{2}$$

where indices k and l range over respectively the image width W and height H, the image pixels $v_{k,l}^{(d)}$ are in the range $[0,1]$ and f is the sigmoid activation function. As mentioned above, this loss encourages to capture the most relevant features in the latent space and to create a regularized latent space where the latent features correspond to the independent generative factors.

3.2. Explainability through Two-Level Alignment

E-VAE makes the disentangled representations of the fashion images multimodal and explainable through the weak supervision of the two-level alignment. More precisely, we extract fashion-related terms and phrases from the e-commerce descriptions and align the extracted textual attributes with the visual data at two levels. First, at a coarse-grained level, we align the textual attributes with the disentangled representations of the fashion images. This encourages E-VAE to focus on encoding factors of variation that correspond with these attributes. In other words, through the alignment the disentangled representations of the fashion images will be enriched with information from the e-commerce descriptions and become multimodal. Furthermore, the alignment with the textual attributes provides a way to interpret and explain which product attributes are encoded in which coordinates of the latent disentangled space. Second, at a fine-grained level, we align the textual attributes with the image regions in the disentangled space. This will visually contextualize the textual attributes further and facilitates the alignment at the coarse-grained level. Next we describe this process in detail.

3.2.1. Textual Attribute Extraction and Representation

Given a product description, we only retain phrases $x_1, ..., x_M$ that are fashion-related based on a glossary of fashion attributes as is done in [27] (for more details see Section 4.1). We represent these attributes with word embeddings $\boldsymbol{e}_1, ..., \boldsymbol{e}_M$ trained with the Skipgram model [28] on a fashion corpus. As a result, attributes that are instances of the same factor of variation are likely to be embedded close together as they are expected to appear in similar contexts. Since the disentanglement process needs to find relevant attributes and put these together in groups that correspond with the factors of variation, these word embeddings provide a good starting point. Next, the word embeddings are projected to the disentangled space:

$$\boldsymbol{s}_j = \text{ReLU}(W_s \boldsymbol{e}_j), \tag{3}$$

with $W_s \in \mathbb{R}^{d_z \times d_t}$. As a result of using the ReLU activation function to project the textual attributes to the disentangled space and through the disentanglement which encodes relevant information related to a certain factor of variation only in a subset of dimensions of the disentangled space, we expect the textual attribute representations to be sparse.

3.2.2. Coarse-grained Alignment of Images and Textual Attributes

At a coarse-grained level, we align the textual attributes $\boldsymbol{s}_j$ with the disentangled representations of the fashion images $\boldsymbol{z}$. More precisely, the alignment score of the kth image and the lth text is computed as the average cosine similarity score of the image with each textual attribute:

$$a_{kl} = \frac{1}{M} \sum_{j=1}^{M} \frac{\boldsymbol{z}^{(k)\intercal} \cdot \boldsymbol{s}_j^{(l)}}{||\boldsymbol{z}^{(k)}|| \cdot ||\boldsymbol{s}_j^{(l)}||}, \tag{4}$$

where $\boldsymbol{z}^{(k)}$ and $\boldsymbol{s}_j^{(l)}$ are the latent representations of the kth image and the jth word of the lth text, respectively, and M is the number of words. (Other similarity functions than cosine similarity could be used here, e.g., unimodal attention mechanisms which compute the similarity between a query vector (here the disentangled representation) and keys (here the textual attributes) are also suitable. We experimented with the stacked attention mechanism of [29] but this did not improve the results.) A triplet loss is used to enforce that the alignment score of an image with its corresponding text a_{dd} should be higher than of the image with the hardest negative text a_{dl}, and vice versa for each text:

$$\mathcal{L}_{IT} = \sum_{d=1}^{D} \max(0, \Delta - a_{dd} + a_{dl}) + \max(0, \Delta - a_{dd} + a_{kd}), \tag{5}$$

where the alignment scores $a_{..}$ are computed with Equation (4), l and k are the indices of respectively the hardest negative text and the hardest negative image for the dth image-text pair, i.e., $l = \text{argmax}_{l \neq d} a_{dl}$ and $k = \text{argmax}_{k \neq d} a_{kd}$, and Δ is the margin.

Through the coarse-grained alignment of the textual attributes and fashion images, the disentangled representations of the fashion images become multimodal. Hence, we consider the disentangled representations of the fashion images $\boldsymbol{z}$ as our multimodal item representations.

3.2.3. Fine-grained Alignment of Image Regions and Textual Attributes

To facilitate the alignment of the images and textual attributes, we also align the textual attributes with regions. This ensures a more fine-grained visual contextualization of the textual attributes. We obtain the image regions representations $\hat{\boldsymbol{v}}_i$ from the last convolutional layer of the encoder and use another fully-connected layer to project them to the disentangled space:

$$\boldsymbol{v}_i = W_v \hat{\boldsymbol{v}}_i, \tag{6}$$

with $W_v \in \mathbb{R}^{d_z \times d_i}$. Then, we use the bidirectional focal attention mechanism of [30] to find the latent alignment of the regions $\boldsymbol{v}_i$ and textual attributes $\boldsymbol{s}_j$ in the disentangled space. Bidirectional focal attention is designed to identify which fragments (i.e., which regions or words) are irrelevant in finding these alignments and cancel them out. Irrelevant fragments are fragments that refer to complementary attributes or that are just noise. The relevance of fragments is determined by computing their relative importance to other fragments. More precisely, it is based on the intuition that compared to relevant fragments in a modality, irrelevant fragments in that modality obtain low attention scores with the fragments in the other modality. Cancelling such irrelevant fragments out is useful when aligning visual and textual e-commerce data which are usually quite complementary. Bidirectional focal attention works in two directions, i.e., it applies text-to-image and image-to-text attention. Here, we describe the text-to-image direction, but the image-to-text direction is completely analogous. In text-to-image attention, we try to find relevant image regions for each word in three steps. First, we compute pre-assigned attention scores α_{ij} as the normalized cosine similarity of each region $\boldsymbol{v}_i$ and word $\boldsymbol{s}_j$:

$$\alpha_{ij} = \text{softmax}(\eta \frac{\boldsymbol{v}_i^\intercal \cdot \boldsymbol{s}_j}{||\boldsymbol{v}_i|| \cdot ||\boldsymbol{s}_j||}), \tag{7}$$

with η a hyperparameter to further increase the gap between relevant and irrelevant regions. Second, we distinguish relevant from irrelevant regions by scoring the ith region based on its preassigned attention score α_{ij} compared with the scores α_{rj} of other regions:

$$F(\alpha_{ij}) = \sum_{r=1}^{R} f(\alpha_{ij}, \alpha_{rj}) g(\alpha_{rj}) \tag{8}$$

$$\text{with } f(\alpha_{ij}, \alpha_{rj}) = \alpha_{ij} - \alpha_{rj} \tag{9}$$

$$\text{and } g(\alpha_{rj}) = \sqrt{\alpha_{rj}}, \tag{10}$$

where R is the number of regions, and functions f and g measure the difference in preassigned attention scores and the importance of a region, respectively. Hence, $F(\alpha_{ij})$ measures the relevance of the ith region for the jth word compared to other regions. The ith region will be considered irrelevant for the jth word if it obtains a low preassigned attention score with the jth word compared to the relevant regions, i.e., if it receives a score $F(\alpha_{ij}) \leq 0$. Third, reassigned attention scores α'_{ij} are computed based on scoring function F:

$$\alpha'_{ij} = \frac{\alpha_{ij} H(\alpha_{ij})}{\sum_{i=1}^{R} \alpha_{ij} H(\alpha_{ij})} \tag{11}$$

$$\text{with } H(\alpha_{ij}) = \mathbb{I}\big(F(\alpha_{ij}) > 0\big) \in \{0, 1\}, \tag{12}$$

where H is an indicator function. The reassigned attention scores are used to compute a visual context vector for each word:

$$\boldsymbol{c}_j^v = \sum_{i=1}^{R} \alpha'_{ij} \boldsymbol{v}_i. \tag{13}$$

Finally, the alignment score of the kth image and the lth text is calculated based on the context vectors from both attention directions:

$$S_{kl} = \frac{\frac{1}{M}\sum_{j=1}^{M}\left(\frac{\boldsymbol{s}_j^{(l)\intercal} \cdot \boldsymbol{c}_j^{v(k)}}{||\boldsymbol{s}_j^{(l)}|| \cdot ||\boldsymbol{c}_j^{v(k)}||}\right) + \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\boldsymbol{v}_i^{(k)\intercal} \cdot \boldsymbol{c}_i^{t(l)}}{||\boldsymbol{v}_i^{(k)}|| \cdot ||\boldsymbol{c}_i^{t(l)}||}\right)}{2}. \tag{14}$$

Each image should have a higher alignment score with its corresponding text than with the hardest negative text, and vice versa for each text:

$$\mathcal{L}_{RT} = \sum_{d=1}^{D} \max(0, \Delta - S_{dd} + S_{dl}) + \max(0, \Delta - S_{dd} + S_{kd}), \tag{15}$$

where Δ is the margin, and l and k are the indices of respectively the hardest negative text and the hardest negative image for the dth image-text pair, i.e., $l = \text{argmax}_{l \neq d} S_{dl}$ and $k = \text{argmax}_{k \neq d} S_{kd}$.

3.3. Complete Loss Function

Since our E-VAE jointly learns to disentangle and align, the complete loss function is:

$$\mathcal{L}_{E-VAE} = \mathcal{L}_{ELBO} + \lambda_1 \cdot \mathcal{L}_{IT} + \lambda_2 \cdot \mathcal{L}_{RT}, \tag{16}$$

with λ_1 and λ_2 hyperparameters.

4. Experimental Setup

This section describes our experimental setup. The datasets used for training our models and our evaluation procedure are explained in Sections 4.1 and 4.2, respectively. The baseline methods we compare with are listed in Section 4.3. Finally, we provide the training and implementation details in Section 4.4.

4.1. Datasets

In our experiments, we use two datasets consisting of real e-commerce data from the web with images of products on a white background and phrase-like product descriptions: the Polyvore Outfits dataset [31] and the Amazon Dresses dataset [32].

4.1.1. Polyvore Outfits

For outfit recommendation we evaluate on the two versions of the Polyvore Outfits dataset [31]. The non-disjoint version consists of 251,008 items and 68,306 outfits and the disjoint version contains 152,785 items and 35,140 outfits. Items belong to one of 11 different product categories, i.e., tops, bottoms, all body, outerwear, shoes, jewellery, bags, hats, scarves, sunglasses and other accessories. As such, the visual variation is very extensive. Examples of outfits of the non-disjoint version are shown in Appendix A Figure A1. The non-disjoint version contains 53,306 outfits for training, 10,000 for testing, and 5000 for validation. The disjoint version contains 16,995 outfits for training, 15,145 for testing, and 3000 for validation. These train-test splits are the same as in [31]. We construct a fashion glossary of the 1000 most frequent fashion-related phrases for the alignment task based on the textual data of the items in the non-disjoint training set (see Appendix A).

4.1.2. Amazon Dresses

For cross-modal search we use the Amazon Dresses dataset [32] which consists of 53,689 image-text pairs describing dresses of different styles for a variety of occasions. Examples of image-text pairs are shown in Appendix A Figure A2. We take the same train-test split as in [32] which uses 48,689 image-text pairs for training, 4000 for validation and 1000 for testing. We also use their fashion glossary for the alignment task (see Appendix A).

4.2. Evaluation

We first perform an intrinsic evaluation of the explainability of the disentangled representations obtained by E-VAE. Next, we compare the performance of E-VAE with that of other baseline methods on outfit recommendation and cross-modal search. For the textual attribute representations produced by E-VAE and E-BFAN (a variant of our E-VAE described in Section 4.3) we also report the sparsity percentage which is the average amount of zero coordinates in the representation of an attribute.

4.2.1. Outfit Recommendation

For outfit recommendation, the disentangled multimodal item representations are projected to multiple type-specific compatibility spaces as proposed by [31]. More precisely, given triplets (z_u, z_v^+, z_v^-) where z_u is the disentangled representation of an item of type u, z_v^+ represents a compatible item of type v and z_v^- is a randomly sampled incompatible item of the same type v, we enforce the following compatibility loss in the type-specific space for pairs of types (u, v):

$$\mathcal{L}_C = \sum_{d=1}^{D} \max(0, \Delta - C_{dd^+} + C_{dd^-}) \tag{17}$$

$$\text{with } C_{dd^+} = \frac{(W_c^{(u,v)} z_u^{(d)})^\intercal \cdot W_c^{(u,v)} z_v^{+(d)}}{||W_c^{(u,v)} z_u^{(d)}|| \cdot ||W_c^{(u,v)} z_v^{+(d)}||}, \tag{18}$$

where Δ is the margin, $W_c^{(u,v)} \in \mathbb{R}^{d_c \times d_z}$ is the projection matrix associated with the type-specific space for the pair (u, v), C_{dd^+} is the compatibility score of a positive item pair, and C_{dd^-} is computed analogous to C_{dd^+}. We add the compatibility loss $\mathcal{L}_C$ to the complete loss (Equation (16)) with a factor λ_3.

We compute performance on two outfit recommendation tasks: outfit compatibility (OC) and fill-in-the-blank (FITB). In the OC task, the goal is to predict how compatible the items in the outfit are in order to distinguish compatible outfits (the positive class) from incompatible outfits (the negative class). Some examples are shown in Figure 1. OC is computed as the average compatibility score across all item pairs in the outfit and evaluated using the area under the ROC curve (AUC). The ROC curve summarizes confusion matrices by plotting the true positive rate on the y-axis and the false positive rate on the x-axis for different classification thresholds of the positive and negative class. The AUC thus provides an aggregate measure of performance across all possible classification thresholds. In the FITB task, given an incomplete outfit and four candidate items, the goal is to predict which candidate item is most compatible with the incomplete outfit. Some examples are shown in Figure 2. The most compatible candidate item is the one which has the highest total compatibility score with the items in the incomplete outfit. Performance on the FITB task is evaluated with accuracy. We use the same set of OC and FITB questions as in [31], that is, 20,000 OC and 10,000 FITB test questions for the non-disjoint version and 30,290 OC and 15,145 FITB test questions for the disjoint version.

4.2.2. Cross-modal Search

In image-to-text retrieval, given an image k we retrieve the top N texts l that best describe the image, that is, the texts l with the highest total similarity score $\mathcal{R}_{kl}$ with image k in the disentangled space. Figure 3 shows some examples for $N = 1$. In text-to-image retrieval, we retrieve the top N images k that display the textual attributes expressed in a given text l, that is, the images k with the highest total similarity score $\mathcal{R}_{kl}$ with text l. Figure 4 shows some examples for $N = 5$. We experiment with two different retrieval models to compute the total similarity score of image k and text l. First, we define the total similarity $\mathcal{R}_{kl,(1)}$ as the similarity of image k and text l at the coarse-grained level:

$$\mathcal{R}_{kl,(1)} = a_{kl}, \tag{19}$$

with a_{kl} as in Equation (4). Second, we define the total similarity $\mathcal{R}_{kl,(2)}$ as the average of the similarity of image k and text l at the coarse-grained and fine-grained level:

$$\mathcal{R}_{kl,(2)} = \frac{a_{kl} + S_{kl}}{2}, \tag{20}$$

with S_{kl} as in Equation (14). We use these retrieval models during testing to evaluate retrieval performance. We compute recall@N for $N = 1, 5, 10$.

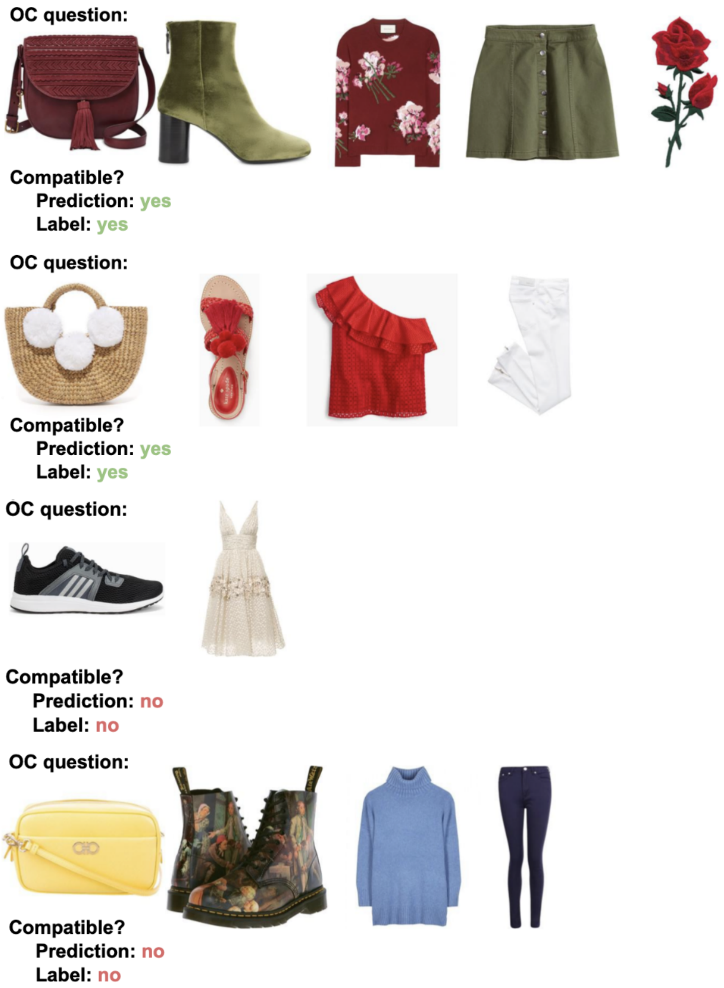


Figure 1. Outfit compatibility results on the Polyvore Outfits dataset obtained with a classification threshold of 50% (best viewed in color).

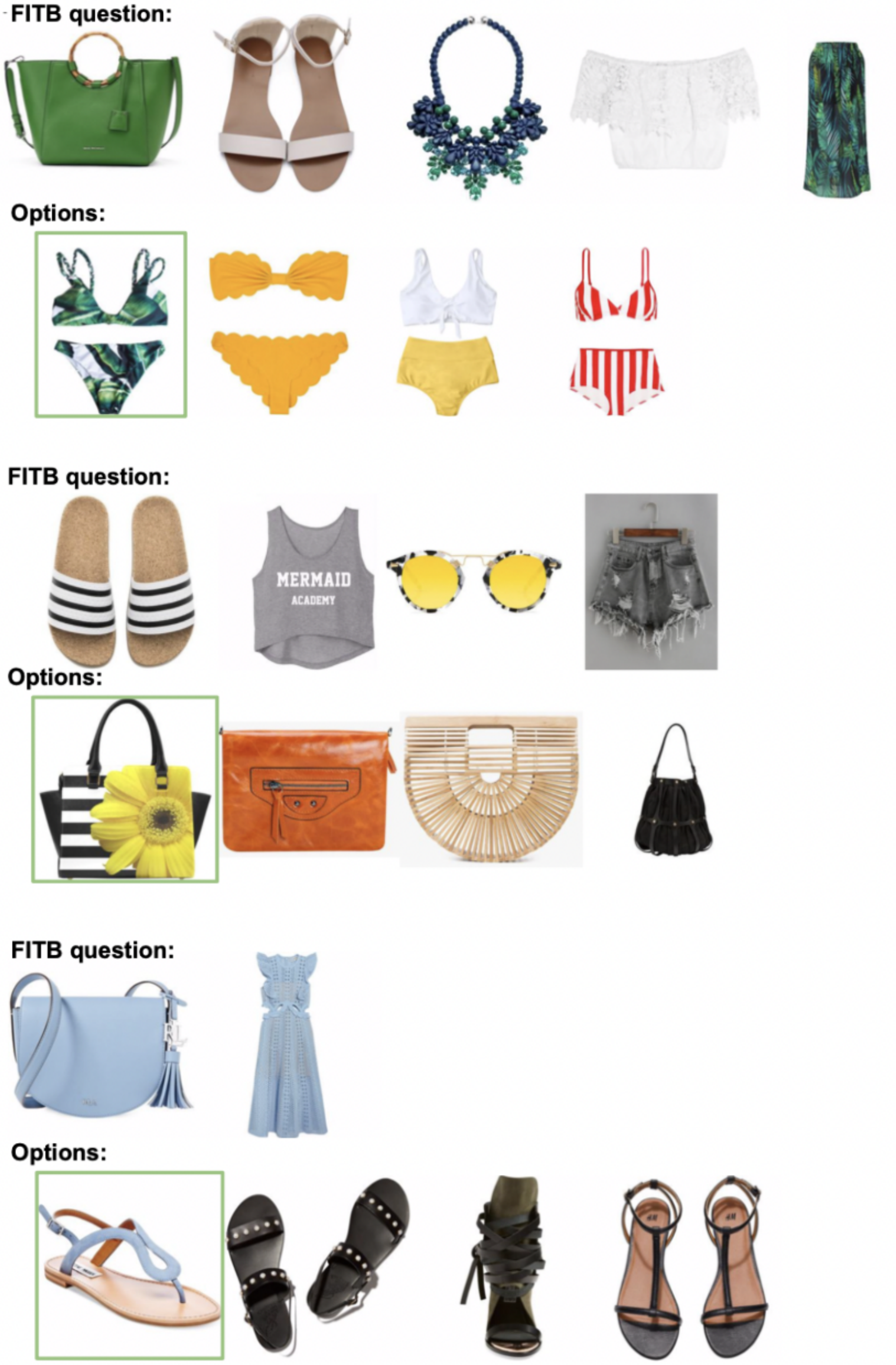


Figure 2. Outfit completion results on the Polyvore Outfits dataset (best viewed in color).

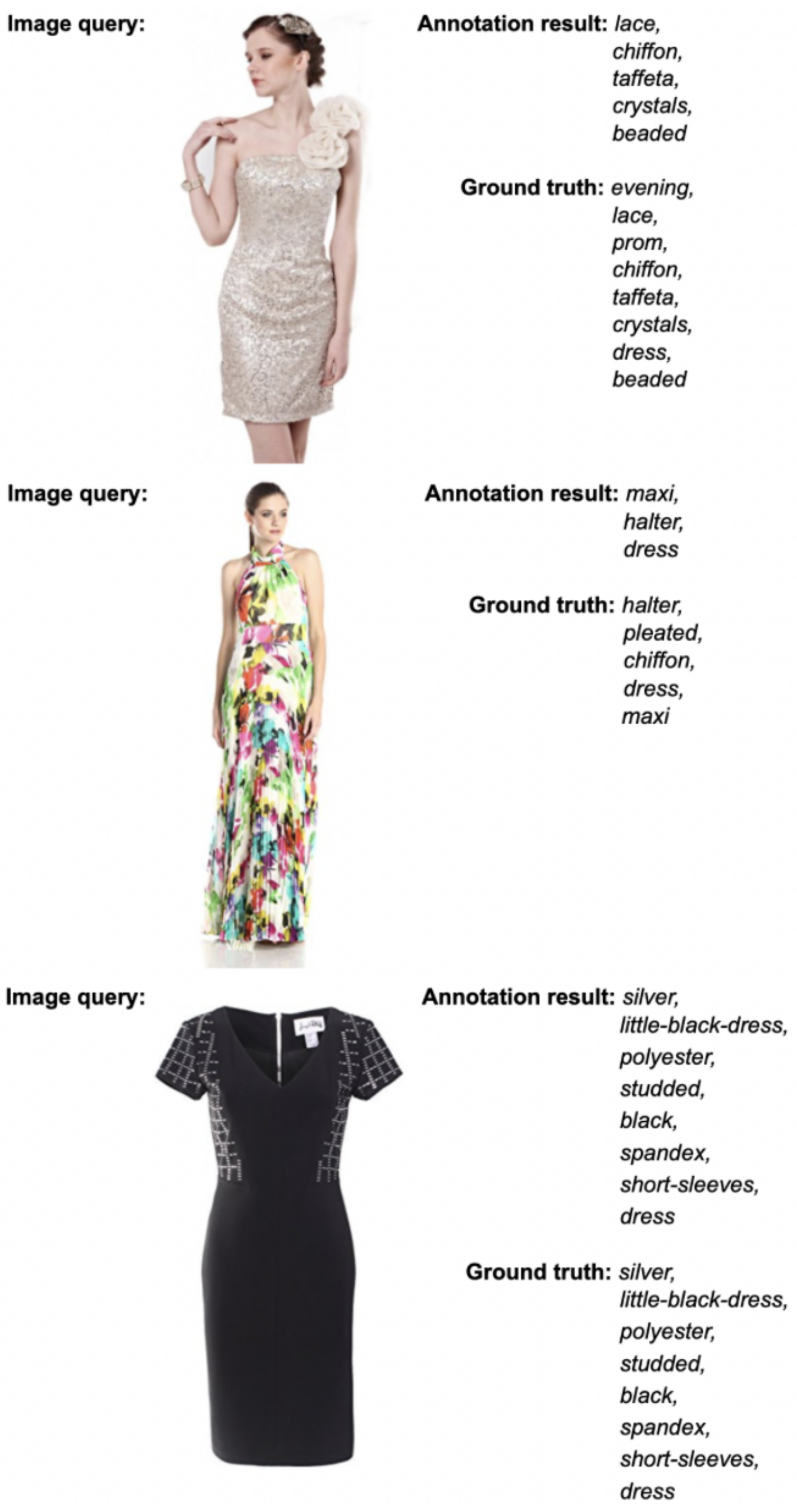


Image query:
Annotation result: lace, chiffon, taffeta, crystals, beaded
Ground truth: evening, lace, prom, chiffon, taffeta, crystals, dress, beaded
Image query:
Annotation result: maxi, halter, dress
Ground truth: halter, pleated, chiffon, dress, maxi
Image query:
Annotation result: silver, little-black-dress, polyester, studded, black, spandex, short-sleeves, dress
Ground truth: silver, little-black-dress, polyester, studded, black, spandex, short-sleeves, dress

Figure 3. Image-to-text retrieval results on the Amazon Dresses dataset (best viewed in color).

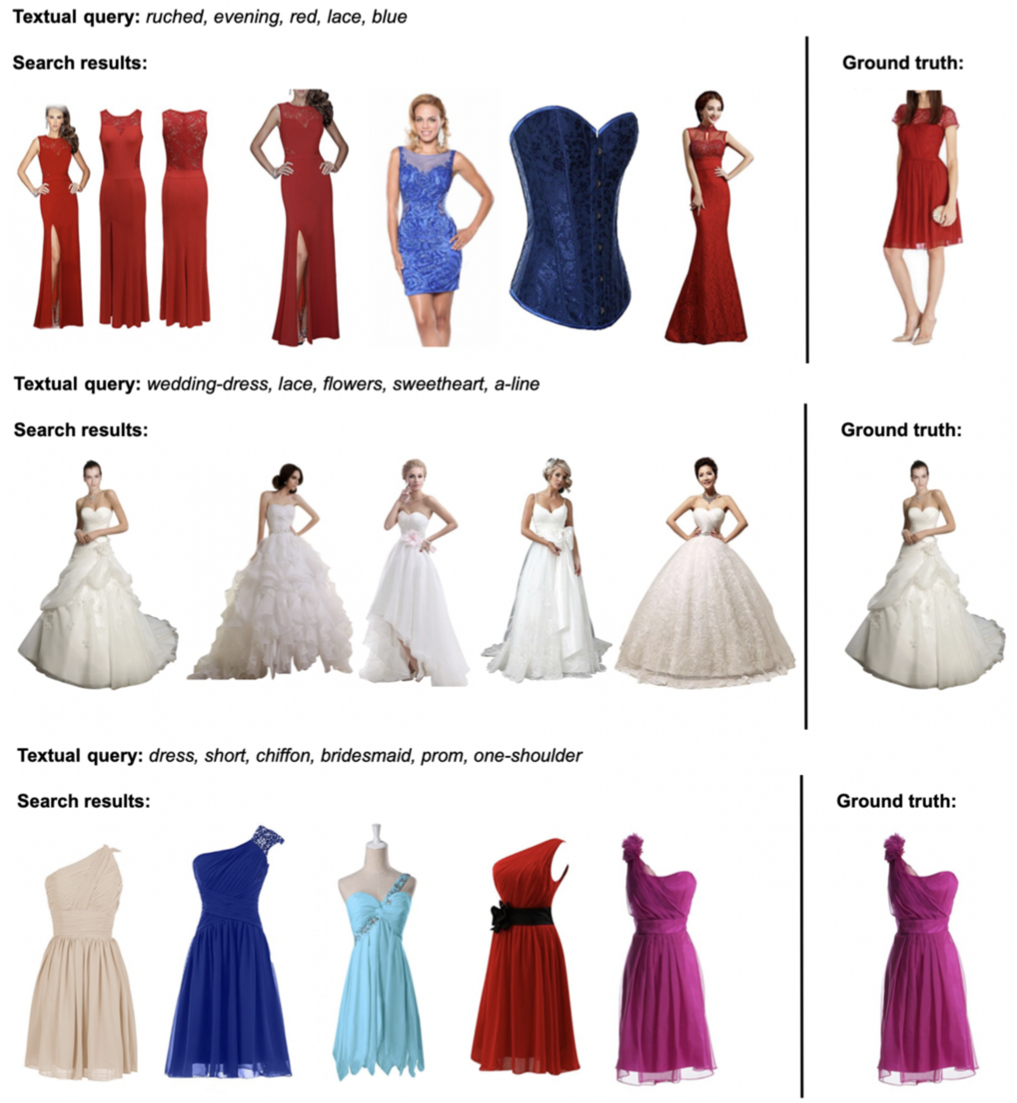

Textual query: ruched, evening, red, lace, blue
Search results:
Ground truth:
Textual query: wedding-dress, lace, flowers, sweetheart, a-line
Search results:
Ground truth:
Textual query: dress, short, chiffon, bridesmaid, prom, one-shoulder
Search results:
Ground truth:

Figure 4. Text-to-image retrieval results on the Amazon Dresses dataset (best viewed in color).

It is important to note that recall computed at the cut-off of N items regards a very strict evaluation because it relies on incomplete product descriptions and an incomplete ground truth reference collection. This means we might retrieve an image for a text that satisfies the text but which is different from the text's ground truth image, or we might retrieve a text for an image which is not (part of) the original image description but which still accurately describes it. Therefore, the actual evaluation results might be higher than the reported results.

4.3. Baseline Methods

The baselines we compare with are listed below. For fair comparison, we selected state-of-the-art models that use the exact same ground truth data, that is both visual and textual item data, and no additional knowledge such as attribute groups or user behaviour data.

4.3.1. Outfit Recommendation

- **TypeAware [31]** This is a state-of-the-art multimodal outfit recommender system which infers a multimodal embedding space for item understanding where semantically related images and texts are aligned as well as images/texts of the same type. Jointly, the system learns about item compatibility in multiple type-specific compatibility spaces.

- ***β*-VAE [5]** Here we first use β-VAE to learn to disentangle the fashion images and afterwards learn the two-level alignment in the disentangled space and compatibility in the type-specific compatibility spaces while keeping the layers of β-VAE frozen.
- **DMVAE [20]** This is a state-of-the-art VAE for multimodal disentangled representation learning which infers three spaces: a single shared multimodal disentangled space and a private unimodal disentangled space both for vision and language. We use a triplet matching loss (similar to Equations (5) and (15)) to learn cross-modal relations. Then, the representations of these three spaces are concatenated to represent an item and projected to the type-specific compatibility spaces for compatibility learning.
- **E-BFAN** This is a variant of E-VAE which infers the two-level alignment in a multimodal shared space but does not aim for disentanglement. The multimodal shared space is enforced with the same loss function as E-VAE (Equation (16)) but with the $\mathcal{L}_{ELBO}$ term removed. Next, the resulting explainable item representations are projected to multiple type-specific compatibility spaces.

4.3.2. Cross-Modal Search

- **3A [27]** This is a state-of-the-art neural architecture for cross-modal and multimodal search of fashion items which learns intermodal representations of image regions and textual attributes using three alignment losses, i.e., a global alignment, local alignment and image cluster consistency loss. In contrast with our work, these alignment losses do not result in explainable representations.
- **SCAN [33]** This is a state-of-the-art model for image-text matching to find the latent alignment of image regions and words referring to objects in general, everyday scenes.
- **BFAN [30]** This is an extension of SCAN [33] that eliminates irrelevant regions and words when inferring their latent alignments while SCAN integrates all of them, which can lead to semantic misalignment.
- ***β*-VAE [5]** We first use β-VAE to learn to disentangle the fashion images and afterwards learn the two-level alignment in the disentangled space while keeping the layers of β-VAE frozen.
- **VarAlign [19]** This is a state-of-the-art neural architecture for image-to-text retrieval consisting of three VAEs. A mapper VAE performs cross-modal variational alignment of the latent distributions of two other VAEs, that is, one VAE that learns to reconstruct the text based on the image and a second VAE that reconstructs the text based on the text itself. We use the disentangled image representations obtained after the mapper and disentangled text representations produced by the second VAE for image-to-text retrieval.
- **DMVAE [20]** This is a state-of-the-art VAE for multimodal disentangled representation learning which infers a single shared multimodal disentangled space and a private unimodal disentangled space per modality. For fair comparison, we use a triplet matching loss (similar to Equations (5) and (15)) to learn the cross-modal relations and use the image shared features and text shared features for cross-modal search.
- **E-BFAN** This is a variant of E-VAE which creates explainable item representations but does not aim for disentanglement and therefore discards $\mathcal{L}_{ELBO}$ from the loss function (Equation (16)).

4.4. Training Details

All models are trained for 50 epochs using the Adam [34] optimizer and a learning rate of 5×10^{-5}. We apply early stopping if there is no improvement for five consecutive epochs. We use batch sizes of 8 and 16 for the Polyvore Outfits and Amazon Dresses dataset, respectively (only for DMVAE [20] we had to use a batch size of 4 for the Polyvore Outfits dataset). For VarAlign [19] we use the setup above for each of the training phases. Images are resized to $W \times H = 256 \times 256$ and represented with the *conv5_block3_out*-layer of size $8 \times 8 \times 2048$. Hence, dimension d_i equals 2048 and d_f equals 131,072. Texts are cleaned by only retaining words from the fashion glossary, removing duplicates if any, to obtain a set of fashion tags for each item. The fashion tags are represented with Skipgram word embeddings [28] of dimension $d_t = 300$ (see Appendix B). The latent space dimension d_z affects how many factors of variation we can discover. Here, we find $d_z = 128$ a reasonable choice for the amount of relevant factors of variation. A sensitivity analysis for d_z on the Amazon Dresses dataset can be found in Appendix D. Dimension d_c also equals 128.

For VAE-based models, weights are initialized according to truncated normal distributions with mean 0 and standard deviation 0.001. For other models, weights are initialized using the Xavier uniform initializer. For hyperparameter β we perform a grid search over interval $\{2, 4, 6, 8, 10, 20, 25, 30, 40, 50\}$ and find the optimal values to be $\beta = 25$ and $\beta = 20$ for the Polyvore outfits and the Amazon Dresses dataset, respectively. For λ_1, λ_2 and λ_3 we perform grid search over interval $\{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7\}$. A $\lambda_1 = 10^5$ and $\lambda_2 = 10^5$ work best for both datasets. For λ_3 we found the optimal value to be 10 for E-BFAN and 10^6 for E-VAE. Furthermore, a Δ of 0.2 and η of 20 were found to work well based on the validation set. For DMVAE [20] we obtained the best results when dividing the latent space dimension d_z into a private space for each modality of dimension 16 and a shared space across all modalities of dimension 112. For outfit recommendation, we project a concatenation of the item representation in the private visual space, private textual space and shared space to the type-specific compatibility spaces. Furthermore, a grid search showed that we achieve the best results when we weight the text reconstruction loss with a factor 10^5 and the matching loss with a factor 10^5. For VarAlign [19], we set the KL loss weighting factor w_1^{KL} of the first VAE equal to 0.9 and the KL loss weighting factor w_2^{KL} of the second VAE to 0.999 as reported in their paper. They do not report how they set their remaining hyperparameters though. Therefore, we set the KL loss weighting factor w_m^{KL} of the mapper to 0.9 and the weighting factors of the text reconstruction losses w_1^r and w_2^r of both VAEs as well as the weighting factor w^W of the Wasserstein distance loss of the mapper to 10^5 inspired by the hyperparameter settings for the text reconstruction and matching loss of DMVAE [20] and find these settings result in a good performance.

5. Results and Discussion

5.1. Explainability

First, we look at the explanations that we can generate for the content of the disentangled item representations through the alignment with the textual attributes. We use the textual explanations to provide insight in how the item representations are disentangled and to evaluate whether this disentanglement is meaningful.

If the item representations are disentangled, this means that each coordinate or subset of coordinates of the representation contains information about only one factor of variation. Recall that we assume that the attributes in our vocabulary can be clustered in attribute groups that correspond with the factors of variation. If the representations are disentangled, attributes sharing non-zero components should belong to the same factor of variation. For example, since color is a factor of variation different colors are expected to share non-zero components. Since for E-BFAN the attribute representations are less sparse as can be seen from Tables 1 and 2, we only examine this for E-VAE.

Table 1. Outfit recommendation results on the Polyvore Outfits dataset.

	Polyvore Outfits					
	Disjoint			Non-Disjoint		
Model	**OC**	**FITB**	**Sparsity**	**OC**	**FITB**	**Sparsity**
	auc	**acc**	**%**	**auc**	**acc**	**%**
TypeAware [31]	81.29	53.28	-	81.94	54.35	-
β-VAE [5]	63.70	37.29	-	66.74	39.13	-
DMVAE [20]	80.20	52.97	-	82.36	55.76	-
E-BFAN	**82.65**	**54.90**	73.28	**89.11**	**60.34**	80.73
E-VAE	81.90	52.24	**82.60**	88.41	58.87	**83.93**

Table 2. Cross-modal search results on the Amazon Dresses dataset. *R@N* denotes recall@*N*.

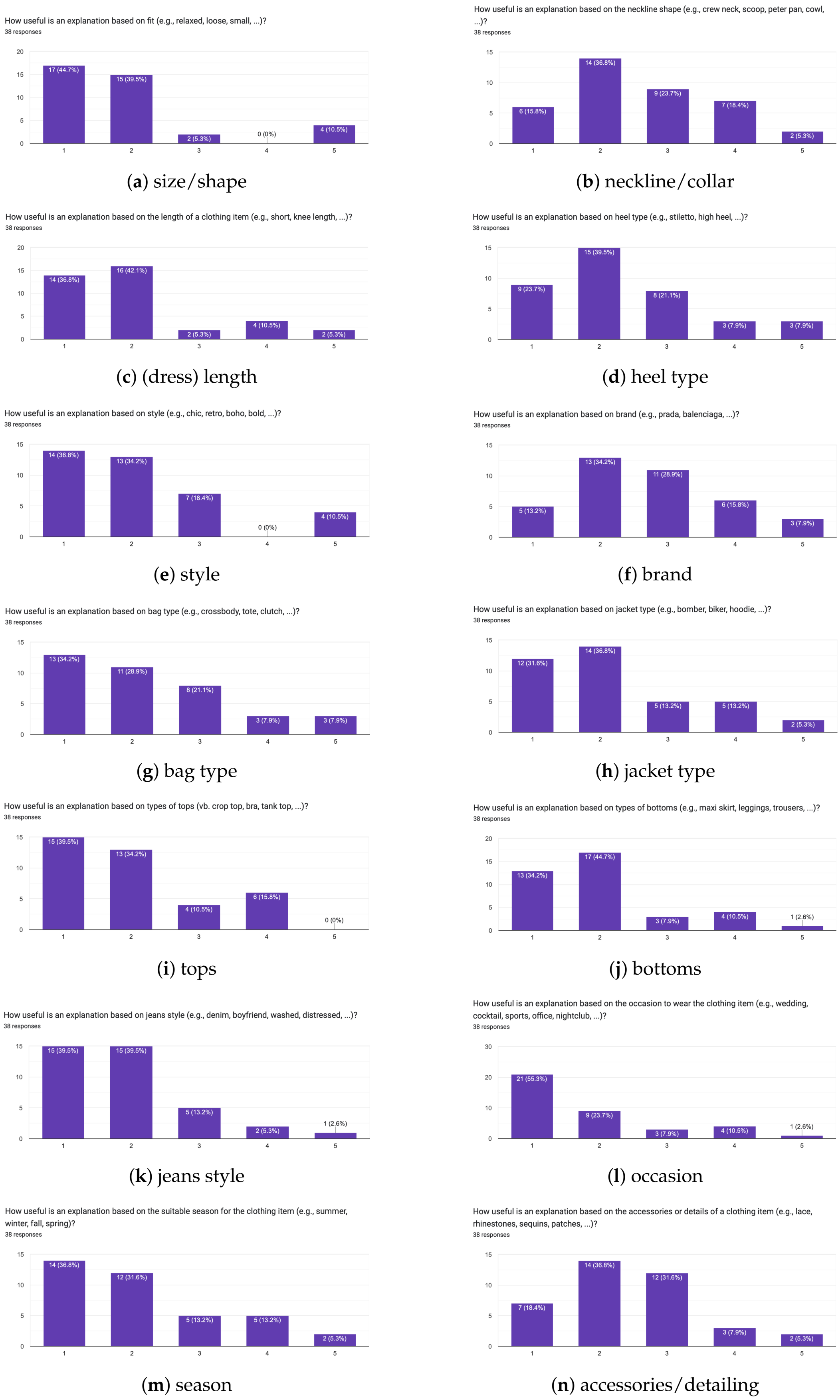
	Amazon Dresses						
Model	**Image to Text**			**Text to Image**			**Sparsity**
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**%**
3A [27]	6.80	19.30	30.00	8.80	21.80	32.10	-
SCAN [33]	0.00	0.20	0.80	0.30	0.40	0.80	-
BFAN [30]	9.20	28.80	41.90	16.90	38.60	52.00	-
β-VAE [5]	0.00	1.20	2.20	0.40	1.20	2.40	-
VarAlign [19]	5.30	15.10	21.40	-	-	-	-
DMVAE [20]	**9.80**	28.70	42.40	11.10	28.60	41.00	-
E-BFAN ($\mathcal{R}_{kl,(1)}$)	6.20	26.10	39.60	15.40	36.40	50.10	69.62
E-BFAN ($\mathcal{R}_{kl,(2)}$)	6.90	27.30	41.30	17.50	39.60	52.40	69.67
E-VAE ($\mathcal{R}_{kl,(1)}$) $- \mathcal{L}_{RT}$	6.80	27.00	39.70	15.20	37.10	47.30	**90.59**
E-VAE ($\mathcal{R}_{kl,(1)}$)	6.70	26.40	42.20	15.60	37.20	49.10	88.89
E-VAE ($\mathcal{R}_{kl,(2)}$)	9.20	**32.10**	**47.20**	**18.20**	**42.10**	**53.40**	88.87

Table 3 shows some examples of attributes and the five attributes they share the most non-zero components with. More precisely, these five attributes are found by ranking the attributes given by the fashion glossary by the number of overlapping non-zero dimensions with the given attribute. In most cases, we can see that attributes that share multiple non-zero components indeed are instances of the same factor of variation. In cases where they are not, we may assume other reasons why they share non-zero components. For instance, *short* (referring to dress length) is visually similar with *short sleeves*. Furthermore, *brown* is the main color of a *leopard-print* and colors and prints could be considered very similar factors of variation. Lastly, *ankle-strap* has a connection with the shoe type *leather-sandals*. Overall, these results show that our E-VAE indeed seems to find the factors of variation considered by humans. For the Amazon Dresses dataset especially colors, textures/prints and fabrics are very well disentangled. For the Polyvore Outfits dataset we see nice examples for different kinds of factors of variation.

The ability to explain disentangled representations makes it possible to inspect which features are encoded in which components of the item representations. In future applications, this allows control over which features are used to produce search and recommendation results. In addition, it could increase the fairness, transparency, trust and effectiveness of search and recommender systems by generating explanations for why certain products are retrieved. We perform a user study to verify whether our textual explanations would be useful to real users of fashion e-commerce websites. Our survey was filled out by 38 women between the ages of 18 and 65 of whom 50% visit a webshop multiple times a week and 89.50% visit a webshop multiple times a month. Of the 38 respondents, 92.10% stated they like to receive recommendations that are personalized to their preferences and 78.90% mentioned that they would like to receive explanations that reveal what information these personalized recommendations are based on. Next, we evaluate for different groups of product attributes how useful it would be to use them in textual explanations for personalized recommendations. More precisely, for each of the factors of variation in Table 3, we asked the respondents how useful it would be to base an explanation on that product attribute group. We used a five-point Likert scale where 1 corresponds with *very useful* and 5 with *very useless*. We found that the majority of the product attribute groups were considered useful to include in explanations. The product attribute groups that were considered most useful can be classified into four classes: fit (size/shape, neckline/collar, (dress) length, heel type), style (style, brand, bag type, jacket type, type of top, type of bottom, jeans style), occasion (occasion, season) and small details (accessories/detailing). The results for these product attribute groups are shown in Figure 5. In the absence of online fitting rooms customers can assess the product completely only after it arrives at their home. This makes fit-related issues one of the main reasons for product returns. Furthermore, another prominent reason for product returns is that the product does not align with the customer expectations. Including fit-related explanations can then help customers gauge the suitability of a product. We assume these are the reasons why users value explanations with fit-related product attributes (Figure 5a–d). Apart from body type, a recommendation engine should also make suggestions based on style preferences. Therefore, style-related explanations are useful for users since they are an indicator whether the system as correctly captured these style preferences (Figure 5e–k). Furthermore, it seems users do not only like to receive recommendations on what to wear but also on when to wear it. Occasion- and season-related explanations were also rated as very useful in our survey (Figure 5l–m). In addition, such explanations reveal whether the user intent while browsing was correctly estimated. Finally, we found that explanations related to small details such as accessories or detailing in the clothing are considered useful as well (Figure 5n). We argue this is because these are difficult to search for in a webshop as they are often not mentioned in the product description nor available as a product attribute group to automatically filter on. However, these small details often give a clothing item something extra and can make or break the fact that the user likes the item. For the remaining factors of variation, that is, color, texture/print, fabric and ankle strap type, only a small majority of respondents found explanations based on these product attribute groups useful. The reason might be the difference in importance of such product attributes, e.g., some users might prefer wool sweaters while for others the fabric does not matter. Alternatively, some users might find that certain explanations are straightforward, e.g., if you have clicked on several black dresses or checkered coats it is logical that you will receive recommendations of black dresses or checkered coats making an explanation of where these recommendations come from less useful. The results can be found in Figure A3 of Appendix C. Overall, we conclude that the textual explanations that could be produced with our explainable disentangled representations would be useful for online customers.

Table 3. Attributes that are instances of a factor of variation and the five attributes they share the most non-zero components with when using E-VAE (attributes that belong to another factor of variation are crossed out.).

	Amazon Dresses	
Attribute	**Five Attributes Sharing Non-Zero Components**	**Factor of Variation**
purple	{gray, navy, brown, blue, ~~zebra-print~~}	color
pink	{beige, coral, brown, yellow, ~~summer~~)}	color
reptile	{plaid, aztec, chevron, ombre, horizontal-stripes }	texture/print
zebra-print	{aztec, chevron, snake-print, ombre, animal-print}	texture/print
floral-print	{paisley, gingham, aztec, snake-print, plaid}	texture/print
short	{tea-length, knee-length, ~~wear-to-work~~, ~~shift~~, ~~short-sleeves~~}	dress length
viscose	{lyocell, cashmere, ramie, wool, acetate}	fabric
acetate	{ramie, lyocell, cashmere, wool, viscose}	fabric
rhinestones	{ruffles, ~~homecoming~~, ~~acrylic~~, chains, sequins}	accessories/detailing
scalloped	{crochet, metallic, fringe, ~~mandarin~~, embroidered}	accessories/detailing
cocktail	{nightclub, homecoming, sports, ~~khaki~~, career}	style/occassion
wear-to-work	{career, office, ~~turtleneck~~, ~~boatneck~~, athletic}	style/occassion
crew	{turtleneck, boatneck, peter-pan, ~~athletic~~, ~~sweater~~}	neckline/collar
scoop	{~~modal~~, square-neck, cowl, turtleneck, boatneck}	neckline/collar
spring	{summer, fall, winter, ~~khaki~~, ~~neutral~~}	season
	Polyvore Outfits	
Attribute	**Five Attributes Sharing Non-Zero Components**	**Factor of Variation**
brown	{khaki, camel, ~~leopard-print~~, tan, burgundy}	color
gingham	{check, plaid, polka, polka-dot, ~~ruffled~~	texture/print
crepe	{silk, satin, twill, acetate, silk-blend}	fabric
beaded	{braided, crochet, feather, leaf, flower}	accessories/detailing
chic	{retro, stylish, boho, bold, inspired}	style
high-neck	{neck, scoop, halter, hollow, neckline}	neckline/collar
Alexander-McQueen	{dsquared, Prada, Balenciaga, Balmain, Dolce-Gabbana}	brand
ankle-strap	{~~leather-sandals~~, slingback, buckle-fastening, strappy, straps}	ankle strap type
bag	{canvas, crossbody, handbag, tote, clutch}	bag type
covered-heel	{high-heel, heel-measures, slingback, stiletto, stiletto-heel}	heel type
bomber	{jacket, hoodie, biker, moto, letter}	jacket type
distressed	{ripped, boyfriend-jeans, denim, topshop-moto, washed}	jeans style
crop-top	{bra, cami, crop, tank, top}	tops
pants	{leggings, maxi-skirt, pant, trousers, ~~twill~~}	bottoms
relaxed-fit	{relaxed, ~~roll~~, loose-fit, rounded, size-small }	size/shape



(**a**) size/shape

(**b**) neckline/collar

(**c**) (dress) length

(**d**) heel type

(**e**) style

(**f**) brand

(**g**) bag type

(**h**) jacket type

(**i**) tops

(**j**) bottoms

(**k**) jeans style

(**l**) occasion

(**m**) season

(**n**) accessories/detailing

Figure 5. Results from a user study on the usefulness of certain product attribute groups in textual explanations. We used a five-point Likert scale where 1 corresponds with *very useful* and 5 with *very useless*. All the product attribute groups shown here are considered useful.

5.2. Outfit Recommendation

The outfit recommendation results on the Polyvore Outfits dataset are reported in Table 1.

We find that E-VAE largely outperforms the β-VAE [5] baseline. This clearly shows the importance of the weak supervision of our alignment to steer the disentanglement process towards finding the relevant factors of variation. Without it, the factors of variation that are found by β-VAE [5] do not correspond with those in the fashion glossary and are less relevant for the OC and FITB task.

Furthermore, our E-VAE surpasses TypeAware [31] with a large margin on the non-disjoint dataset split. For the disjoint dataset split, E-VAE only outperforms TypeAware [31] on the OC task. We argue that this performance difference between the two dataset splits is due to their size. The non-disjoint version (approximately 250,000 items of 11 different product categories) is easier to disentangle by E-VAE than the disjoint version (approximately 150,000 items of 11 different product categories) as disentangling requires enough training examples to discover the relevant factors of variation.

The results obtained with DMVAE [20] also indicate that disentangling the disjoint dataset split is harder. On the disjoint dataset split, E-VAE performs better on the OC task but not on the FITB task. On the larger non-disjoint dataset split, E-VAE outperforms DMVAE [20] with a large margin. We argue that this is due to our two-level alignment in the disentangled shared space, which is more fine-grained than the one inferred by DMVAE [20] and can therefore better model fine-grained product attributes involved in making fashionable outfit combinations.

While the representations created with E-VAE are sparser, the best results on both tasks and dataset splits are obtained with the E-BFAN model which creates explainable but more entangled representations. However, note that the Polyvore Outfits dataset is extremely challenging to disentangle. It contains extensive visual variation due to the presence of 11 different product categories and contains fewer examples per product category compared with the Amazon Dresses dataset. Furthermore, we consider an extensive attribute glossary of 1000 fine-grained attributes as instances of factors of variation to disentangle.

Some examples of OC and FITB questions answered by E-VAE are shown in Figures 1 and 2.

5.3. Cross-Modal Search

We report our cross-modal search results on the Amazon Dresses dataset in Table 2. We make several observations.

First, when comparing E-BFAN and E-VAE, we see that both achieve the best results with retrieval model $\mathcal{R}_{kl,(2)}$. We find that E-BFAN ($\mathcal{R}_{kl,(2)}$) is surpassed by E-VAE ($\mathcal{R}_{kl,(2)}$) on all metrics. In addition, the textual attribute representations created by E-VAE are more sparse than those of E-BFAN, which seems to indicate that the item representations produced by E-BFAN are indeed more entangled. E-VAE ($\mathcal{R}_{kl,(1)}$) $- \mathcal{L}_{RT}$ shows the effect of the additional alignment at the fine-grained level. Without $\mathcal{L}_{RT}$ (Equation (5)) we can only use retrieval model $\mathcal{R}_{kl,(1)}$ since we then no longer learn how to compute S_{kl}. We observe that the alignment of the textual attributes and image regions improves the results when it is both enforced during training ($\mathcal{L}_{RT}$) and incorporated in the retrieval model (S_{kl}).

Second, our E-VAE surpasses the β-VAE [5] baseline with a great margin. This clearly demonstrates the effectiveness and necessity of the two-level alignment of the vision and language data in the disentangled space. Hence, our proposed extensions to β-VAE [5] to disentangle multimodal e-commerce data create item representations that better capture relevant fine-grained product attributes. Furthermore, they have the additional benefit that they result in textual explainability.

Also SCAN [33] is largely outperformed by the other models. We expect this is because the other models that find the latent alignments of image regions and words such as 3A [27], BFAN [30], E-BFAN and E-VAE all include some method to eliminate irrelevant regions and words, while SCAN integrates all of them. This is often not that big of a problem for multimodal datasets with images of objects in everyday scenes where the annotations are generated and curated to describe the image content. However for multimodal e-commerce data, where the images and descriptions are often complementary and noisy, failing to eliminate such attributes will lead to semantic misalignment. Therefore, image-text matching models created for general multimodal data are often not suitable to align multimodal e-commerce data.

Compared to the original BFAN [30] which is not explainable, E-VAE ($\mathcal{R}_{kl,(2)}$) performs better on all metrics and E-BFAN ($\mathcal{R}_{kl,(2)}$) performs slightly better on text-to-image retrieval but not on image-to-text retrieval.

VarAlign [19] is also surpassed by E-VAE. This indicates that for image-to-text retrieval a two-level cross-modal alignment of the visual and textual item data is more suited than a cross-modal alignment of the latent distributions of two unimodal disentangled spaces.

DMVAE [20], which jointly learns to disentangle and align an image and its description in a shared multimodal space, outperforms E-VAE ($\mathcal{R}_{kl,(2)}$) on recall@1 for image-to-text retrieval but not on other metrics. DMVAE performs very well on image-to-text retrieval but less so on text-to-image retrieval. We argue that this is because their disentangled full-text representations make less expressive search queries than our sparse textual attribute representations.

Overall, E-VAE ($\mathcal{R}_{kl,(2)}$) achieves the highest cross-modal search results on the Amazon Dresses dataset, except on recall@1 for image-to-text retrieval. This shows that through the alignment of the disentangled representations with textual attributes we obtain explainable representations that are separated in factors of variation corresponding with relevant product attribute groups and that provide a useful mechanism to identify whether a particular item has a particular product attribute. The most sparse attribute representations are created with E-VAE ($\mathcal{R}_{kl,(1)} - \mathcal{L}_{RT}$). There is a positive correlation between the sparsity of the attribute representations and the quality of disentanglement of the item representations. We provide some qualitative examples obtained by E-VAE ($\mathcal{R}_{kl,(2)}$) on image-to-text and text-to-image retrieval in Figures 3 and 4. These figures demonstrate that, given a description, E-VAE ($\mathcal{R}_{kl,(2)}$) can retrieve images that exhibit the requested attributes, and given an image, E-VAE ($\mathcal{R}_{kl,(2)}$) can retrieve suitable descriptions.

5.4. Summary and Future Work

The experimental results (Tables 1 and 2) demonstrate that E-VAE achieves state-of-the-art performance on outfit recommendation and surpasses the state of the art on cross-modal search. Our E-VAE largely outperforms β-VAE [5] on both tasks which clearly shows the technical contribution of our model for disentangling multimodal e-commerce data.

In addition, we illustrate that in contrast with state-of-the-art methods our E-VAE has the additional benefit that it learns disentangled representations that are interpretable and explainable. We justify the disentanglement and explainability by proving two points. First, the textual attribute representations are very sparse (Tables 1 and 2) meaning that information related to these product attributes is encoded in a small subset of dimensions of the disentangled space. Second, attributes that are encoded in similar subsets of dimensions indeed belong to the same factor of variation (Table 3). Future research might focus on further improving the alignment of the disentangled representations with textual attributes. In addition, in future work our method to create interpretable and explainable disentangled item representations could be used to build transparent search and recommender systems that generate explanations to why certain products are retrieved. There it could be explored how to generate multimodal explanations instead of only textual explanations in search and recommendation. Finally, it would be interesting to combine the strengths of our proposed method with those of the transformer architecture.

We found that especially the Polyvore Outfits dataset was challenging to disentangle. With 11 different product categories and fewer examples per product category than in the Amazon Dresses dataset, the visual variation was very extensive and the noise abundant. Furthermore, our Polyvore attribute glossary considered 1000 fine-grained attributes for disentanglement compared to 205 for the Amazon Dresses dataset. Overall, our E-VAE succeeded to find the factors of variation considered by humans. Especially colors, prints and fabrics are very well disentangled for the Amazon Dresses dataset. For the Polyvore Outfits dataset we showed nice examples of discovered factors of variation that are general, e.g., color, style, size and brand, as well as specific to a particular clothing category, e.g., bag type, heel type and jeans type. An interesting research direction to explore here is the design of self-supervised pretraining tasks for e-commerce images to obtain better representations of the fine-grained discriminative regions in these images that could make the disentanglement easier. Moreover, future work could explore ways to change the saliency definition of a VAE through the use of another reconstruction loss more suitable for e-commerce images.

6. Conclusions

We proposed E-VAE which learns explainable disentangled representations from multimodal e-commerce data. We achieved this by jointly learning to disentangle the product images and learning to align the product tags with the product images and their regions in the disentangled space. Through the weak supervision of the alignment we could steer the disentanglement process towards discovering factors of variation of interest which was necessary given the huge search space of visual variations. In addition, our proposed model is the first to generate textual explanations for the content of disentangled representations and gave us insight in whether the disentanglement is meaningful, which groups of product attributes are well separated and which are not. Finally, we obtained state-of-the-art outfit recommendation results on the Polyvore Outfits dataset and new state-of-the-art cross-modal search results on the Amazon Dresses dataset.

Author Contributions: Conceptualization, K.L.; methodology, K.L.; software, K.L.; validation, K.L.; formal analysis, K.L.; investigation, K.L.; resources, M.-F.M.; data curation, K.L.; writing—original draft preparation, K.L.; writing—review and editing, K.L. and M.-F.M.; visualization, K.L.; supervision, M.-F.M.; project administration, K.L.; funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: The first author was funded by the KU Leuven Postdoctoral Mandate grant number 3E210691. The second author was funded by the ERC Advanced Grant CALCULUS H2020-ERC-2017-ADG 788506.

Data Availability Statement: The Polyvore Outfits dataset is available at https://github.com/mvasil/fashion-compatibility (accessed on 17 October 2022) and the Amazon Dresses dataset can be accessed through https://liir.cs.kuleuven.be/software_pages/fashion_wsdm_ijcee.php (accessed on 17 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC curve
β-VAE	Beta-Variational AutoEncoder
BFAN	Bidirectional Focal Attention Network
CNN	Convolutional Neural Network
DMVAE	Disentangled Multimodal Variational AutoEncoder
FITB	Fill-In-The-Blank
E-BFAN	Explainable Bidirectional Focal Attention Network
E-VAE	Explainable Variational AutoEncoder
OC	Outfit Compatibility
POS	Part-Of-Speech
ROC	Receiver Operating Characteristic
SCAN	Stacked Cross-Attention
VAE	Variational AutoEncoder
XAI	eXplainable Artificial Intelligence

Appendix A. Datasets

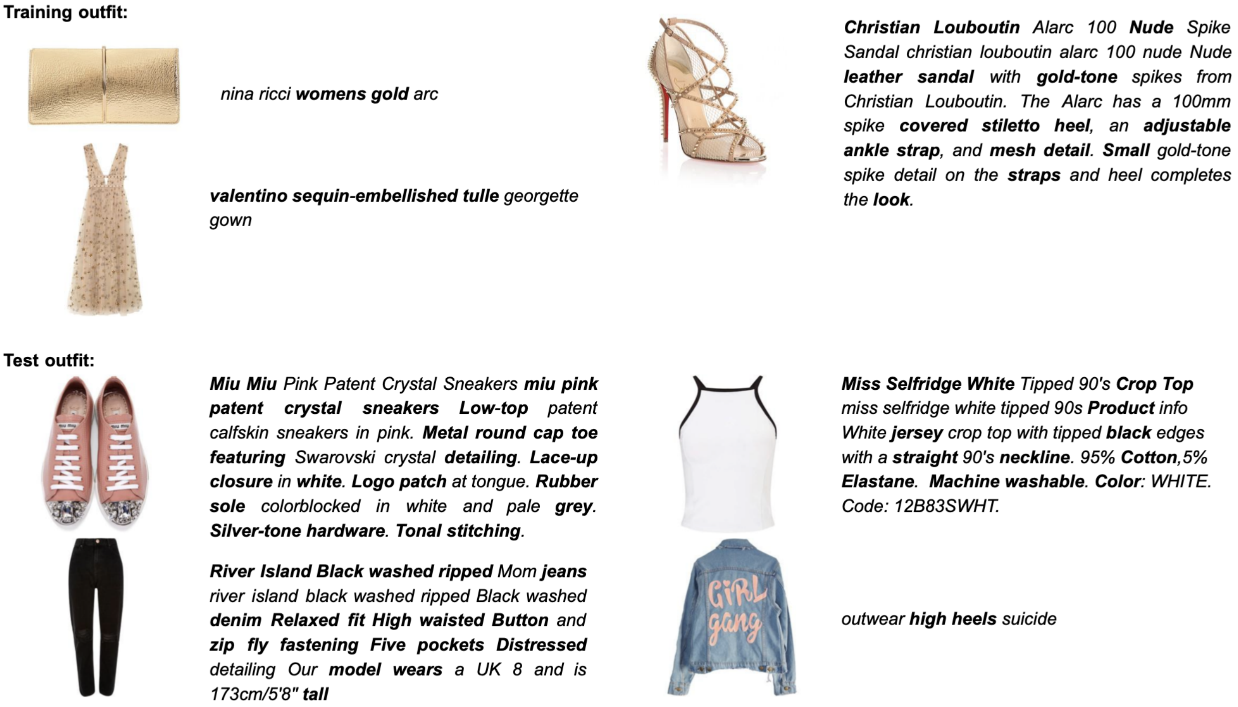
For our experiments, we chose datasets with real-world e-commerce data from the web consisting of images of products on a white background and phrase-like descriptions. Such images of products on a white background are common on e-commerce websites and already pose a great challenge in discovering the generative factors given the extensive visual variations. Furthermore, the phrase-like descriptions allow to train word embeddings on a fashion vocabulary where embeddings of attributes belonging to the same attribute group are often already embedded close together as they appear in similar contexts. This benefits the disentanglement process where attributes need to be clustered in attribute groups that make up the factors of variation. Two datasets that meet these requirement are the Polyvore Outfits [31] and the Amazon Dresses [32] datasets.

Appendix A.1. Polyvore Outfits

For the alignment task, we cannot use the same fashion glossary as for the Amazon Dresses dataset as that one is mostly focused on fashion attributes of dresses. Furthermore, fashion glossaries provided with other datasets such as the DeepFashion dataset [35] are also not suitable since as fashion changes quickly, new fashion terms are also introduced frequently and the same fashion concept can be described differently depending on the vendor. In addition, the DeepFashion dataset does not have shoes, however, this is a product category which has a quite specific vocabulary. Therefore, we use the textual data of the items in the training set of the non-disjoint version of the Polyvore Outfits dataset to obtain a glossary of fashion terms to use in the alignment task. More precisely, for each item we use both the title, description and url name. We only keep alphabetical characters and punctuation, split contractions into multiple tokens, replace punctuation by a white space, remove multiple consecutive white spaces and put everything in lower case. Next, we use a part-of-speech (POS) tagger on the tokens and only keep adjectives, adverbs, verbs and nouns. Then, we look at all unigrams, bigrams and trigrams for potential attributes, but remove those words referring to sizes (i.e., containing *mm*, *cm*, *inch* or *xxs* to *xxl*), words consisting of only two letters or less, and the verbs *has*, *have*, *is* and *are*. Finally, we compute the frequency of the remaining unigrams, bigrams and trigrams and take the 1000 most frequent ones as our fashion glossary. The resulting glossary will be made publicly available. Examples of items and their description before and after cleaning with the fashion glossary are shown in Figure A1.

Appendix A.2. Amazon Dresses

The fashion glossary used to clean the descriptions of the Amazon Dresses dataset is shown in Table A1. The terms in this fashion glossary are product attributes which are instances of different factors of variations. Examples of items and their description before and after cleaning with the fashion glossary are shown in Figure A2.



Training outfit:

nina ricci **womens gold** arc

valentino sequin-embellished tulle georgette gown

Christian Louboutin Alarc 100 **Nude** Spike Sandal christian louboutin alarc 100 nude Nude **leather sandal** with **gold-tone** spikes from Christian Louboutin. The Alarc has a 100mm spike **covered stiletto heel**, an **adjustable ankle strap**, and **mesh detail**. **Small** gold-tone spike detail on the **straps** and heel completes the **look**.

Test outfit:

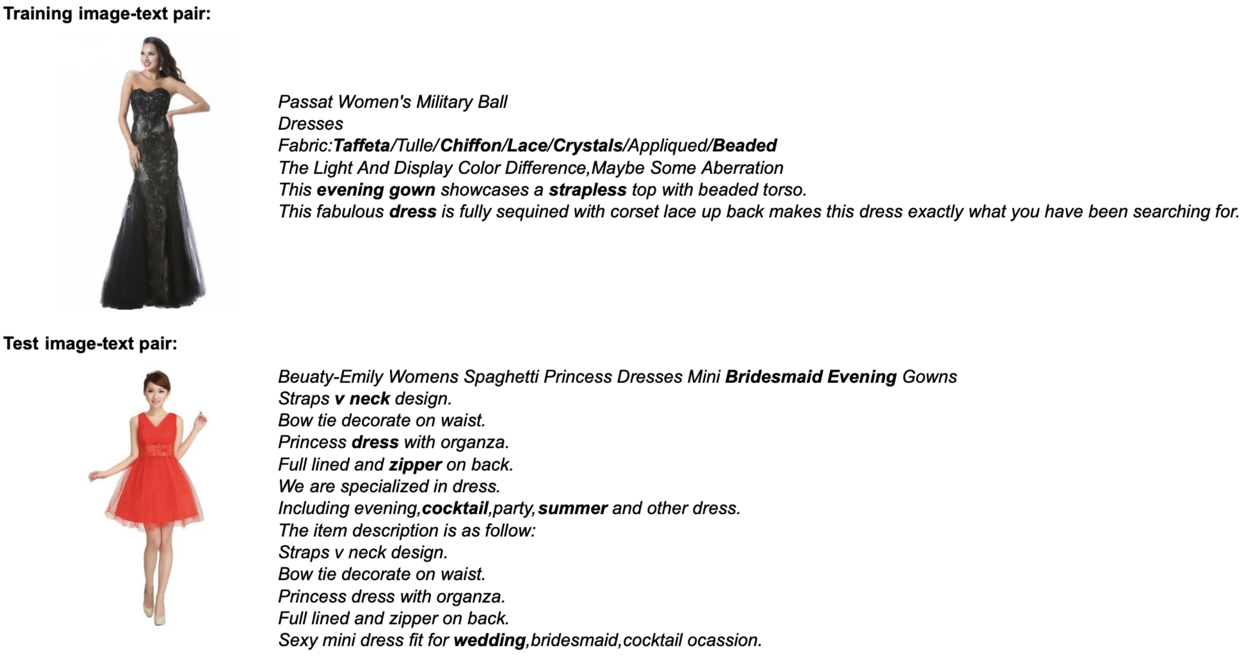
Miu Miu Pink Patent Crystal Sneakers **miu pink patent crystal sneakers Low-top** patent calfskin sneakers in pink. **Metal round cap toe featuring** Swarovski crystal **detailing**. **Lace-up closure** in **white**. **Logo patch** at tongue. **Rubber sole** colorblocked in white and pale **grey**. **Silver-tone hardware**. **Tonal stitching**.

River Island Black washed ripped Mom **jeans** river island black washed ripped Black washed **denim Relaxed fit High waisted Button** and **zip fly fastening Five pockets Distressed** detailing Our **model wears** a UK 8 and is 173cm/5'8" **tall**

Miss Selfridge White Tipped 90's **Crop Top** miss selfridge white tipped 90s **Product** info White **jersey** crop top with tipped **black** edges with a **straight** 90's **neckline**. 95% **Cotton**,5% **Elastane**. **Machine washable**. **Color**: WHITE. Code: 12B83SWHT.

outwear **high heels** suicide

Figure A1. Examples of outfits in the training and test set (best viewed in color). Bold phrases in the item descriptions are those that remain after cleaning with the fashion glossary.



Training image-text pair:

Passat Women's Military Ball
Dresses
Fabric:**Taffeta**/Tulle/**Chiffon**/**Lace**/**Crystals**/Appliqued/**Beaded**
The Light And Display Color Difference,Maybe Some Aberration
This **evening gown** showcases a **strapless** top with beaded torso.
This fabulous **dress** is fully sequined with corset lace up back makes this dress exactly what you have been searching for.

Test image-text pair:

Beuaty-Emily Womens Spaghetti Princess Dresses Mini **Bridesmaid Evening** Gowns
Straps **v neck** design.
Bow tie decorate on waist.
Princess **dress** with organza.
Full lined and **zipper** on back.
We are specialized in dress.
Including evening,**cocktail**,party,**summer** and other dress.
The item description is as follow:
Straps v neck design.
Bow tie decorate on waist.
Princess dress with organza.
Full lined and zipper on back.
Sexy mini dress fit for **wedding**,bridesmaid,cocktail ocassion.

Figure A2. Examples of image-text pairs from the training and test set (best viewed in color). Bold phrases in the item descriptions are those that remain after cleaning with the fashion glossary.

Table A1. Amazon Dresses fashion glossary. The glossary consists of 205 terms.

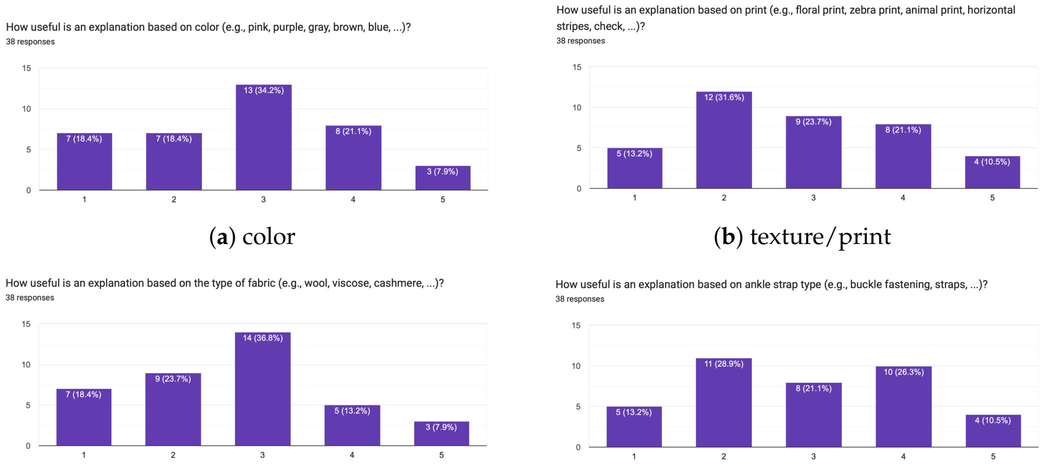
Amazon Dresses Fashion Glossary				
3-4-sleeve	a-line	abstract	acetate	acrylic
action-sports	animal-print	applique	argyle	asymmetrical
athletic	aztec	baby-doll	ballet	banded
beaded	beige	black	blue	boatneck
bone	bows	bridesmaid	brocade	bronze
brown	burgundy	bustier	buttons	camo
canvas	career	cashmere	casual	chains
checkered	cheetah-print	chevron	chiffon	coat
cocktail	contrast-stitching	coral	cotton	cover-up
cowl	crew	crochet	crystals	cut-outs
denim	dip-dyed	dress	dropped-waist	embroidered
empire	epaulette	evening	fall	faux-leather
faux-pockets	felt	floor-length	floral-print	flowers
fringe	geometric	gingham	gold	gown
gray	green	grommets	halter	hemp
high-low	high-waist	homecoming	horizontal-stripes	houndstooth
jacquard	jersey	juniors	keyhole	khaki
knee-length	lace	leather	leopard-print	linen
little-black-dress	logo	long	long-sleeves	lycra
lyocell	mahogany	mandarin	maxi	mesh
metallic	microfiber	mock-turtleneck	modal	mother-of-the-bride
multi	navy	neutral	nightclub	notch-lapel
nylon	off-the-shoulder	office	olive	ombre
one-shoulder	orange	outdoor	paisley	patchwork
peplum	peter-pan	petite	pewter	pink
piping	pique	plaid	pleated	plus-size
point	polka-dot	polyester	ponte	prom
purple	ramie	rayon	red	reptile
resort	retro	rhinestones	ribbons	rivets
ruched	ruffles	satin	scalloped	scoop
screenprint	sequins	shawl	sheath	shift
shirt	short	short-sleeves	silk	silver
sleeveless	smocked	snake-print	snap	spandex
sport	sports	spread	spring	square-neck
strapless	street	studded	summer	surf
sweater	sweetheart	synthetic	taffeta	tan
tank	taupe	tea-length	terry	tie-dye
tropical	tunic	turtleneck	tweed	twill
v-neck	velvet	vertical-stripes	viscose	wear-to-work
wedding	western	white	wide	winter
wool	wrap-dress	yellow	zebra-print	zipper

Appendix B. Word Embeddings

We use the Skipgram model [28] to obtain word embeddings. For each dataset, we use as word embedding training data the phrase-like descriptions of the items in the training set, where we replace multiword fashion concepts from the fashion glossary with one hyphenated word such that one embedding can be learned for that concept, e.g., little black dress is replaced by little-black-dress, skinny jeans by skinny-jeans, etc. We train for 15 epochs to obtain word embeddings of dimension $d_t = 300$ using a window size of 5.

Appendix C. User Study

The results of our user study to verify the usefulness of textual explanations based on color, texture/print, fabric and ankle strap type are presented in Figure A3. For these product attribute groups there was no consensus among respondents.



(**a**) color (**b**) texture/print

(**c**) fabric (**d**) ankle strap type

Figure A3. Results from our user study on the usefulness of certain product attribute groups in textual explanations. We used a five-point Likert scale where 1 corresponds with *very useful* and 5 with *very useless*. For the product attribute groups shown here there is no general consensus on their usefulness.

Appendix D. Sensitivity Analysis

We conduct a sensitivity analysis of the dimension d_z for the best performing systems on the Amazon Dresses dataset (Table A2). The values for dimension d_z we experiment with are 32, 64, 128, 256 and 512. The results show that our E-VAE is more sensitive to the dimension d_z than the other models. This is because d_z affects how many factors of variation we expect to find and should be carefully tuned. Since the representations learned by E-VAE are much more sparse, a d_z that is too small does not leave so much room to encode relevant features. For a d_z that is too large, we hypothesize that the granularity of the factors of variation that are discovered is too small and that correspondence with the textual attributes in the product descriptions can no longer be found.

Table A2. Sensitivity analysis of dimension d_z for cross-modal retrieval conducted for the best performing models on the Amazon Dresses dataset. *R@N* denotes recall@*N*.

Model	d_z	**Amazon Dresses**					
		Image to Text			**Text to Image**		
		R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
3A [27]	32	6.60	18.10	27.40	4.30	15.40	25.60
	64	6.00	18.60	29.60	5.00	16.40	25.30
	128	6.80	19.30	30.00	8.80	21.80	32.10
	256	7.30	22.60	32.60	6.80	20.30	30.10
	512	8.90	23.80	34.20	7.40	18.60	29.50
BFAN [30]	32	7.40	26.20	38.00	12.60	34.00	46.40
	64	8.00	27.90	42.00	16.10	37.90	50.40
	128	9.20	28.80	41.90	16.90	38.60	52.00
	256	9.30	28.30	42.20	15.50	35.40	49.50
	512	8.10	29.80	41.40	15.50	37.60	50.60
E-VAE ($\mathcal{R}_{kl,(1)}$)	32	0.50	1.60	3.10	0.20	0.90	1.40
	64	6.80	23.20	36.10	12.40	32.20	45.10
	128	6.70	26.40	42.20	15.60	37.20	49.10
	256	0.30	1.30	2.80	0.30	1.20	2.30
	512	0.30	0.50	1.40	0.10	0.50	1.00

Table A2. *Cont.*

Model	d_z	Amazon Dresses					
		Image to Text			Text to Image		
		R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
E-VAE ($\mathcal{R}_{kl,(2)}$)	32	0.10	1.60	3.70	0.30	1.30	1.80
	64	6.30	22.10,	35.30	13.50	33.50	47.10
	128	9.20	32.10	47.20	18.20	42.10	53.40
	256	0.20	0.60	1.40	0.10	0.50	1.00
	512	0.10	0.80	1.30	0.10	0.50	1.00
E-BFAN ($\mathcal{R}_{kl,(2)}$)	32	5.70	25.00	38.10	12.80	35.30	46.80
	64	6.10	25.60	40.20	14.40	36.90	49.90
	128	6.90	27.30	41.30	17.50	39.60	52.40
	256	8.20	28.70	42.80	18.90	39.20	52.80
	512	9.30	30.00	44.20	18.20	42.50	53.50

References

1. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
2. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
3. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
4. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22–24 June 2014; pp. 1278–1286.
5. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
6. Kim, H.; Mnih, A. Disentangling by Factorising. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2649–2658.
7. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
8. Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; Vetter, T. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In Proceedings of the 2009 IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy, 2–4 September 2009; pp. 296–301.
9. Aubry, M.; Maturana, D.; Efros, A.A.; Russell, B.C.; Sivic, J. Seeing 3D Chairs: Exemplar Part-Based 2D-3D Alignment Using a Large Dataset of CAD Models. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3762–3769.
10. Yang, Z.; Hu, Z.; Salakhutdinov, R.; Berg-Kirkpatrick, T. Improved Variational Autoencoders for Text Modeling using Dilated Convolutions. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3881–3890.
11. Zhu, Q.; Bi, W.; Liu, X.; Ma, X.; Li, X.; Wu, D. A Batch Normalized Inference Network Keeps the KL Vanishing Away. *arXiv* **2020**, arXiv:2004.12585.
12. Ma, J.; Zhou, C.; Cui, P.; Yang, H.; Zhu, W. Learning Disentangled Representations for Recommendation. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., Red Hook, NY, USA, 2019; pp. 5711–5722.
13. Hou, Y.; Vig, E.; Donoser, M.; Bazzani, L. Learning Attribute-Driven Disentangled Representations for Interactive Fashion Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12147–12157.
14. Bouchacourt, D.; Tomioka, R.; Nowozin, S. Multi-Level Variational Autoencoder: Learning Disentangled Representations From Grouped Observations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
15. Locatello, F.; Poole, B.; Raetsch, G.; Schölkopf, B.; Bachem, O.; Tschannen, M. Weakly-Supervised Disentanglement Without Compromises. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 6348–6359.
16. Chen, H.; Chen, Y.; Wang, X.; Xie, R.; Wang, R.; Xia, F.; Zhu, W. Curriculum Disentangled Recommendation with Noisy Multi-feedback. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual Event, 6–14 December 2021.

17. Zhang, Y.; Zhu, Z.; He, Y.; Caverlee, J. Content-Collaborative Disentanglement Representation Learning for Enhanced Recommendation. In Proceedings of the Fourteenth ACM Conference on Recommender Systems, Virtual Event, 22–26 September 2020; pp. 43–52.
18. Zhu, Y.; Chen, Z. Variational Bandwidth Auto-encoder for Hybrid Recommender Systems. *IEEE Trans. Knowl. Data Eng.* **2022**, *early access*. [CrossRef]
19. Theodoridis, T.; Chatzis, T.; Solachidis, V.; Dimitropoulos, K.; Daras, P. Cross-modal Variational Alignment of Latent Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 4127–4136.
20. Lee, M.; Pavlovic, V. Private-Shared Disentangled Multimodal VAE for Learning of Latent Representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 1692–1700.
21. Chen, Z.; Wang, X.; Xie, X.; Wu, T.; Bu, G.; Wang, Y.; Chen, E. Co-Attentive Multi-Task Learning for Explainable Recommendation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Macau, China, 10–16 August 2019; pp. 2137–2143.
22. Truong, Q.T.; Lauw, H. Multimodal Review Generation for Recommender Systems. In Proceedings of the The World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 1864–1874.
23. Hou, M.; Wu, L.; Chen, E.; Li, Z.; Zheng, V.W.; Liu, Q. Explainable Fashion Recommendation: A Semantic Attribute Region Guided Approach. *arXiv* **2019**, arXiv:1905.12862.
24. Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; Zha, H. Personalized Fashion Recommendation with Visual Explanations Based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 765–774.
25. Liu, W.; Li, R.; Zheng, M.; Karanam, S.; Wu, Z.; Bhanu, B.; Radke, R.J.; Camps, O. Towards Visually Explaining Variational Autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Laenen, K.; Zoghbi, S.; Moens, M.F. Web Search of Fashion Items with Multimodal Querying. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA, 5–9 February 2018.
28. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
29. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A.J. Stacked Attention Networks for Image Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
30. Liu, C.; Mao, Z.; Liu, A.A.; Zhang, T.; Wang, B.; Zhang, Y. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 3–11.
31. Vasileva, M.I.; Plummer, B.A.; Dusad, K.; Rajpal, S.; Kumar, R.; Forsyth, D.A. Learning Type-Aware Embeddings for Fashion Compatibility. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
32. Zoghbi, S.; Heyman, G.; Gomez, J.C.; Moens, M.F. Fashion Meets Computer Vision and NLP at e-Commerce Search. *Int. J. Comput. Electr. Eng.* **2016**, *8*, 31–43. [CrossRef]
33. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. The International Conference on Learning Representations. *arXiv* **2014**, arXiv:1412.6980.
35. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.