11-2023

Multimodal fashion knowledge extraction as captioning

Yifei YUAN
Chinese University of Hong Kong

Wenxuan ZHANG
Alibaba DAMO Academy

Yang DENG
Singapore Management University, ydeng@smu.edu.sg

Wai LAM
Chinese University of Hong Kong

Citation

YUAN, Yifei; ZHANG, Wenxuan; DENG, Yang; and LAM, Wai. Multimodal fashion knowledge extraction as captioning. (2023). *SIGIR-AP '23: Proceedings of the 11th International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Beijing, November 23-28*. 52-62.

Available at: https://ink.library.smu.edu.sg/sis_research/9170

Social Media Fashion Knowledge Extraction as Captioning

Yifei Yuan
The Chinese University of Hong Kong
Hong Kong SAR
yfyuan@se.cuhk.edu.hk

Wenxuan Zhang
Alibaba DAMO Academy
Singapore
isakzhang@gmail.com

Yang Deng
National University of Singapore
Singapore
dengyang17dydy@gmail.com

Wai Lam
The Chinese University of Hong Kong
Hong Kong SAR
wlam@se.cuhk.edu.hk

ABSTRACT

Social media plays a significant role in boosting the fashion industry, where a massive amount of fashion-related posts are generated every day. In order to obtain the rich fashion information from the posts, we study the task of social media fashion knowledge extraction. Fashion knowledge, which typically consists of the occasion, person attributes, and fashion item information, can be effectively represented as a set of tuples. Most previous studies on fashion knowledge extraction are based on the fashion product images without considering the rich text information in social media posts. Existing work on fashion knowledge extraction in social media is classification-based and requires to manually determine a set of fashion knowledge categories in advance. In our work, we propose to cast the task as a captioning problem to capture the interplay of the multimodal post information. Specifically, we transform the fashion knowledge tuples into a natural language caption with a sentence transformation method. Our framework then aims to generate the sentence-based fashion knowledge directly from the social media post. Inspired by the big success of pre-trained models, we build our model based on a multimodal pre-trained generative model and design several auxiliary tasks for enhancing the knowledge extraction. Since there is no existing dataset which can be directly borrowed to our task, we introduce a dataset consisting of social media posts with manual fashion knowledge annotation. Extensive experiments are conducted to demonstrate the effectiveness of our model.

KEYWORDS

fashion knowledge extraction, social media analysis, multimodal data mining

ACM Reference Format:
Yifei Yuan, Wenxuan Zhang, Yang Deng, and Wai Lam. 2023. Social Media Fashion Knowledge Extraction as Captioning. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '23), November 26–28, 2023, Beijing, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3624918.3625315

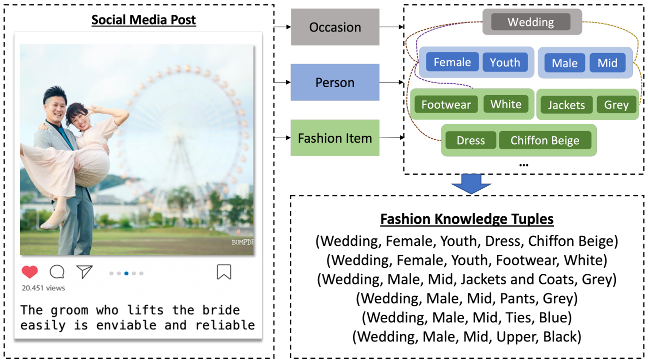

Social Media Post
20,451 views
The groom who lifts the bride easily is enviable and reliable
Occasion
Person
Fashion Item
Wedding
Female
Youth
Male
Mid
Footwear
White
Jackets
Grey
Dress
Chiffon Beige
...
Fashion Knowledge Tuples
(Wedding, Female, Youth, Dress, Chiffon Beige)
(Wedding, Female, Youth, Footwear, White)
(Wedding, Male, Mid, Jackets and Coats, Grey)
(Wedding, Male, Mid, Pants, Grey)
(Wedding, Male, Mid, Ties, Blue)
(Wedding, Male, Mid, Upper, Black)

Figure 1: One example for the fashion knowledge extraction task. The fashion knowledge tuple consists of the occasion, person attributes including gender and age group, and the type and appearance of the fashion items in the post.

1 INTRODUCTION

Fashion, as one of the most important aspects of modern life, has been flourishing and evolving over the past decades. As a new style of sharing information, social media has become an important platform of constantly updating fashion information. Given the rich fashion-related posts, extracting fashion knowledge from them can assist many downstream applications such as personalized recommendation [7, 8, 11, 16], fashion image retrieval [20, 39, 45] and so on, therefore attracting increasing attention in recent years [20, 39, 45].

As shown in Figure 1, a fashion post on the left side often consists of an image and the post text content written by the user for sharing his/her feelings. The Fashion Knowledge Extraction (FKE) task thus aims to elicit key fashion information from the post, including the occasion, person attributes, and the detailed fashion item information. Following the previous study [25, 50], the extracted fashion knowledge is usually denoted as a set of tuples containing essential fashion information. As listed on the right bottom part, the organized knowledge tuples represent the fashion-related information from the post in a more structured manner.

Solving the FKE task on social media posts is an interesting yet challenging problem. Firstly, most existing FKE works are directly

SIGIR-AP '23, November 26–28, 2023, Beijing, China

ACM ISBN 979-8-4007-0408-6/23/11…$15.00
https://doi.org/10.1145/3624918.3625315

conducted on the fashion product images [4, 12, 14, 21] where a single image taken in the professional studio is provided. However, social media posts often contain information of different modalities, including both the image and text. As shown in Figure 1, apart from the post image, the corresponding post text also indicates essential information such as where the image is taken, who in the post is, and what the person is wearing, and thus attaching great importance to extracting the fashion knowledge from the post. Therefore, how to make full use of the multimodal post information for the FKE task is underexplored.

To handle the multimodal information for harvesting the fashion knowledge, some initial attempts have been made. Ma et al. [25] propose a pipeline-based model which first extracts person and clothing boxes from the image, then classifies the detected regions into different attribute categories with the text as an additional input. However, the model merely incorporates the image and text features by simple concatenation which fails to capture the deep interplay between different modalities. Moreover, similar to text-based structure prediction problems [48, 49], formulating the knowledge extraction task as a classification problem needs to manually determine a set of fashion knowledge categories in advance. However, the format of fashion knowledge aspects is typically quite varied. For example, "muslin white" and "chiffon beige" can both be used to describe the appearance of the dress the woman wears in Figure 1. Besides, strong dependencies are often observed between different aspects in the fashion knowledge data [25]. For instance, the person clothes can be affected by the occasion of the post. Traditional classification-based models tend to determine the category of each fashion knowledge aspect separately, thus failing to capture such relationship. Furthermore, their method is pipeline-based and can give rise to the problem of error propagation. Some potential errors in the preceding steps such as the inaccurate prediction of person boxes could lead to a negative influence on not only the extraction of person attributes information but also all the fashion item knowledge corresponding to the person.

To tackle the research challenges discussed above, we propose to cast the social media based FKE task as a captioning problem. Inspired by the classic image captioning task [38, 42] that generates a natural language description for a given image, we transform the FKE task to a captioning problem for better modeling the interplay of the image and text information and alleviating the issues of the classification-based models. Specifically, given the multimodal social media post including an image and the corresponding text, we aim to generate a natural language caption for the post, which contains the key fashion information. The fashion knowledge tuples can then be easily extracted from the generated caption. During the training stage, we first transform the original fashion knowledge tuples into a pseudo caption with a sentence transformation method. Then the multimodal post and the pseudo caption can be paired as training instances to learn a multimodal generation model. With such caption generation formulation, we can tackle the FKE task in an end-to-end manner, alleviating the potential error propagation issue in pipeline-based solutions. Moreover, compared with existing classification-based models, our model incorporates the multimodal information from both the image and text as input and utilizes the natural language caption as the output, which can better capture the interactions between different modalities. In addition, the dependencies between different fashion knowledge aspects can also be fully exploited by learning to generate them in an autoregressive manner.

Motivated by the big success of pre-trained language models for various vision-language tasks such as image-text retrieval [23], we build our model based on a multimodal pre-trained generation model named VL-Bart [5] to utilize its rich knowledge of processing information from different modalities. We further design several auxiliary tasks including visual question answering (VQA), sentence reconstruction, and image-text matching to warm-up the model. These tasks are designed to equip the model with fashion-related knowledge via different formats but under the same model architecture. After training with multiple relevant tasks, the model can obtain some prior task-specific knowledge, which helps tackle the main concerned FKE task.

Since existing datasets used in previous studies are either single-modal with only fashion item information [21] or not publicly available [25], there is no dataset that can be directly adopted for the concerned task. Therefore, we introduce a large-scale fashion knowledge dataset based on user-generated social media fashion-related posts. For each post including an image and text, we manually annotate its corresponding occasion, person attributes, as well as the type and appearance of the fashion items they wear to construct the fashion knowledge tuples. We provide detailed statistics on this newly introduced dataset and conduct extensive experiments on it[1].

To sum up, the main contributions of our paper are as follows:

- We propose to tackle fashion knowledge extraction from multimodal social media posts as a captioning task, which effectively captures the interplay of different modalities via generating a natural language caption for extracting the fashion knowledge tuples in an end-to-end manner.
- To equip the model with fashion-related knowledge, we design several auxiliary tasks including sentence reconstruction, image-text matching, and visual question answering, which helps tackle the main concerned FKE task.
- We contribute a benchmark dataset and conduct extensive experiments to demonstrate the effectiveness of our model. We show that our method outperforms various state-of-the-art methods, especially under the difficult multi-person multi-fashion-item situation.

2 RELATED WORK

2.1 Fashion Knowledge Extraction

Fashion knowledge plays a vital role in fashion-related tasks such as clothing recognition [4, 15], fashion trend forecasting [24, 26, 27], fashion sentiment analysis [46], and fashion-related information retrieval [41, 45]. Therefore, there has been an increasing interest on knowledge extraction tasks in the fashion domain recently [12, 14]. Early studies mostly rely on handcrafted features and mainly focus on extracting simple clothing-related knowledge using techniques such as conditional random field [4, 21]. Huang et al. [12] propose a Dual Attribute-aware Ranking Network (DARN) consisting of two sub-networks for retrieval feature learning. DeepFashion,

[1] The dataset and code are available in https://github.com/yfyuan01/FKE.

which is first proposed by Liu et al. [21], is annotated with clothing items with rich fashion knowledge information. They propose a dataset where each picture is annotated with some fashion item attributes. Jia et al. [14] propose a data-driven approach for recognizing fashion attribute where a modified version of Faster R-CNN model is trained. Furthermore, Wang et al. [39] solve the problem of fashion landmark localization and clothing category classification via a knowledge-guided fashion network. Yan et al. [43] address unconstrained fashion landmark detection, where clothing bounding boxes are not provided in both training and testing phases. To the best of our knowledge, Ma et al. [25] are the first to focus on social media based fashion knowledge extraction, which aims to conduct automatic fashion knowledge extraction from social media posts by unifying the occasion, person attributes and clothing prediction in a contextualized module. Although the model incorporates multimodal information from the social media posts, it is pipeline-based and requires to extract all person boxes in advance. Moreover, they do not publish their dataset for safety reasons.

2.2 Multimodal Pre-training

Following the success of large pre-trained models in natural language understanding (NLU) [6, 9, 44] and generation (NLG) tasks [3, 17, 32], some multimodal pre-trained models have shown their superiority over traditional non-pretrained methods in many tasks recently. Some of them mainly focus on video-text pretraining such as VideoBERT [36], HERO [18], MIL-NCE [28, 29] and so on, while others focus on the image-text domain. Among these image-text pretrained methods, ViLBERT [23], LXMERT [44], and VL-BERT [35] are the extensions of the popular BERT model [9] and are used for learning task-agnostic joint representations of the image content and natural language. Following this line, unified models are proposed to deal with both understanding and generation tasks. For example, Oscar [19] leverages object tags detected in images as anchor points to significantly ease the learning of image-text alignments. CLIP [31] connects image and text representations by learning visual concepts from natural language supervision. Huo et al. [13] propose a two-tower pre-trained model named WenLan within the cross-modal contrastive learning framework. CogView [10] and DALLE [33] are powerful generative models that focus on text-to-image generation. Among them, VL-Bart [5] is the state-of-the-art model designed for vision text generation and shows good generalization ability on different tasks. Therefore, we adopt it as the backbone of our model (to leverage its knowledge in processing information from different modalities) in this work.

3 OUR METHOD

3.1 Problem Definition

We aim to automatically extract fashion knowledge from a social media post, which is composed of an image and the post text content. Following the definition given in previous study [25], the fashion knowledge is denoted as a set of tuples, each tuple k is defined as the combination of the occasion, person attributes, and fashion item information: $k = (o, p, f)$. Here o denotes the occasion category, which belongs to a set of occasions such as wedding, school, sports, etc. $p = (age, gender)$ denotes the gender and age information of a specific person in the post, where $gender \in \{Male, Female\}$ and $age \in \{Kid, Youth, Mid, Old\}$. The fashion item information $f = (type, app)$ contains the fashion item type $type$ such as "pants" and the appearance app of the fashion item, where the appearance is usually a short text such as "lace white" describing both its pattern, color, and style. Therefore, the fashion knowledge tuple k can also be unfolded and represented as $k = (occ, age, gender, type, app)$.

Given a post x consisting of an image v and text t denoted as $x = \{v, t\}$, the problem is to develop a framework which outputs N fashion knowledge tuples of the post, represented as $K = \{k\}_{i=1}^{N}$, where the number of tuples N varies from post to post.

3.2 Framework Overview

Figure 2 presents an overview of our proposed method. In general, we formulate the concerned FKE task as a captioning problem. We tackle it via an encoder-decoder structure based on a pre-trained generative model named VL-Bart [5], as shown in the left part. By treating it as a multimodal generation problem, the interactions between different modalities can be effectively captured. Then the structured fashion knowledge tuples are recovered from the generated caption. Besides, as shown in the right part, before captioning, we leverage several fashion-related auxiliary tasks to warm-up the pre-trained models and equip it with task-specific knowledge.

In detail, for the captioning phase in the left part, we fine-tune the model to generate the fashion knowledge captions. Instead of generating the tuple-like fashion knowledge directly, the model generates captions in a natural language manner. To facilitate such training, given the original training instance with the format of post-tuple pair (x, K), we transform the fashion knowledge tuples K to a pseudo caption y containing all the desired fashion knowledge elements of the post via a caption construction method. The transformed training instance can thus be represented as (x, y) for learning a multimodal generative model. To add fashion information to the pre-trained model, as shown in the right part of Figure 2, we design various auxiliary tasks processes including sentence reconstruction (SRC), visual question answering (VQA), and image text matching (ITM) before fine-tuning. These tasks are designed to focus on one or several fashion knowledge aspects and equip the model with task-specific fashion knowledge.

3.3 Image and Text Encoding

3.3.1 *Text Encoding.*

As shown in the bottom part of Figure 2, the input text t of our model consists of three parts: task prefix, task text, and post text. Post text is the original text content written by the user in the social media post. To auxiliary task training, we also include a task prefix which indicates which task the model should perform, followed by the task text used as an additional textual input for a specific task (e.g. it can be a question in the visual question answering task). The three textual inputs are concatenated with a special token [SEP] and fed to the embedding layer to obtain the text embedding of the model. The positional embeddings for denoting the absolute token positions are added to the token embeddings and learned during the training. Then for each training instance, the text input t is encoded to a vector represented as e^t.

3.3.2 *Image Encoding.*

To extract image features, we first detect several object regions from the image, denoted as Region of Interest (ROI). By utilizing ROI instead of the raw image pixels, we can align

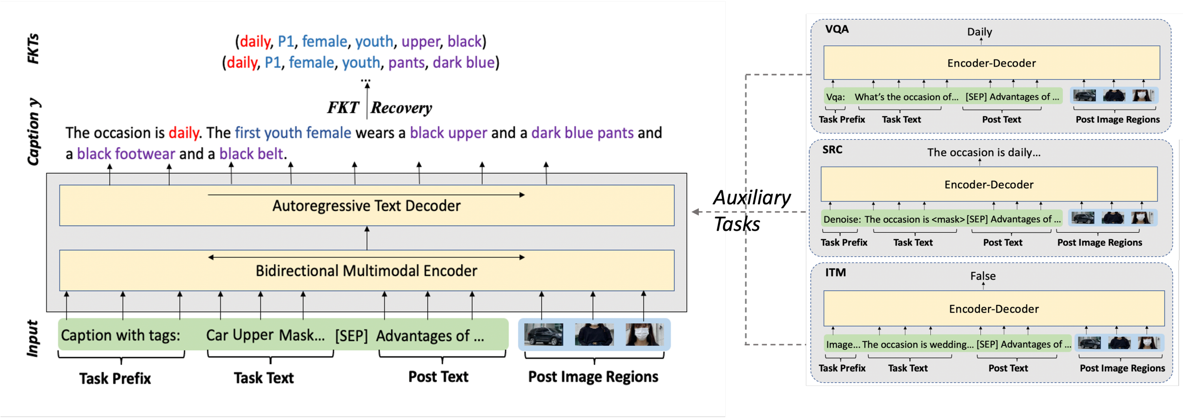


Figure 2: The overall architecture of our framework. The left part depicts transforming the FKE task as a captioning problem, where FKT denotes the fashion knowledge tuples. The right part shows the detail of the three auxiliary tasks.

the multimodal information between the image and text [19, 23]. To obtain ROI features, following previous studies [19, 23], we generate r image object regions with Faster-RCNN [34]. For each region, we also detect the object tag in the format of text such as "Upper", "Woman", etc. The final embedding is the sum of four types of features: ROI object features, ROI bounding box coordinates, image ids, and region ids. The ROI object feature is the encoding result from Faster-RCNN. The bounding box coordinate is the position vector of the ROI. Image id is set to be 1 in our task, and region id $\in \{1, ..., r\}$. The visual embedding of image v is represented as e^v.

3.4 FKE as Captioning

We cast the original FKE task as a captioning problem. We aim to train a generation model for learning the mapping function given the natural language caption y transformed from the fashion knowledge tuples K and the social media post x.

3.4.1 Caption Construction. To facilitate the training process of the generative model, we propose a strategy to construct the pseudo caption y from the N fashion knowledge tuples of a post represented as $K = \{k\}_{i=1}^{N}$. For the caption construction, we wish to incorporate the major fashion knowledge elements into the caption while neglecting the unnecessary information. The rule of transforming the fashion knowledge tuples into the natural language caption is designed as follows.

As shown in Algorithm 1, since the occasion of all the fashion knowledge tuples corresponding to a certain post is the same, we first transform the occasion information into a sentence at the beginning of the target sequence with the template "`The occasion is [occ]`". In the example shown in Figure 2, the sentence saying "The occasion is daily" is constructed to incorporate the occasion category. We then group and gather all the fashion knowledge tuples by different persons. For each person, we write a sentence containing his/her gender and age information. With the same example, we write "The first youth female" at the beginning of the second sentence. We then list all the fashion items the person wears including their type and appearance and incorporate them into a fashion item description sentence. For different fashion items of the same person, we concatenate them with the word "`and`" to mimic the writing method the users often use. Therefore, for the girl in the Figure 2, we add the fashion item information by saying that she wears a black upper and a dark blue pants, etc. After the sentence transformation process that transforms the original tuple-like data into a natural language caption, the input-to-target generation can be modeled with a classic encoder-decoder architecture.

Algorithm 1 Caption Construction

```
Input:
    $N$ fashion knowledge tuples of a post $\{k\}_{i=1}^{N}$
    Each tuple $k = (occ, gender, age, type, app)$
    Number of persons in the post $n_p$
Output:
    Natural Language Sequence $y$
$o \leftarrow$ "The occasion is "$+occ$
$s \leftarrow o$
for $m = 1$ to $n_p$ do
    $y \leftarrow y+$ "The " $+m+gender+age+$"person wears "
    for $n = 1$ to $N$ do
        if $n! = N$ then
            $y \leftarrow y+$ "a" $+ app + type +$ " and "
        else
            $y \leftarrow y$ "a" $+ app + type +$ "."
        end if
    end for
end for
```

3.4.2 Encoder-Decoder Structure. We use transformer [37] encoder-decoder to incorporate image and text features and generate the fashion knowledge caption. The encoder is composed of m transformer blocks, each of which consists of a self-attention layer and a fully-connected layer with residual connections. The decoder is also a stack of m transformers with an additional cross-attention layer in each block. Given the multimodal post input x, the image v and text t are first fed into the bidirectional encoder and incorporated together into a contextualized sequence. Given the sequence, the decoder models the conditional probability distribution of the target sentence to generate caption y. At each time step, the decoder iteratively predicts the probability of current caption tokens based on previously generated tokens and the encoder output.

3.4.3 Training. Given a pretrained model with the encoder-decoder structure, we fine-tune our model parameters θ on the input-target pair. We utilize standard sentence generation loss as our loss function. At each time step j, the decoder output y_j is determined based on the generated caption by previous time steps $y_{<j}$, the image and text embedding e^v and e^t. We minimize the negative log-likelihood of generating the target caption y given the input text embedding e^t and image embedding e^v:

$$\min \; -logP_{\theta}(y|e^t, e^v) = -\sum_{j=1}^{|y|} logP_{\theta}(y_j|y_{<j}, e^t, e^v) \qquad (1)$$

where P_θ is the likelihood of generating the target caption y given the image text input, and $|y|$ is the length of the target caption.

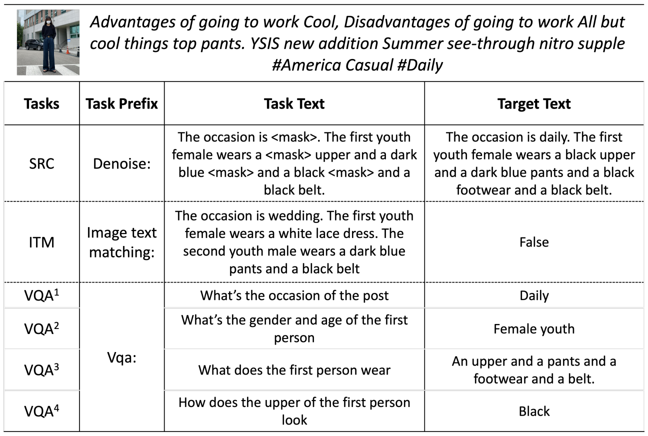
3.4.4 Inference and Tuple Recovery. During inference, we generate the target caption sequence y' in an autoregressive manner given the post image and text pair. Same as mentioned in the training phase, the input text also consists of the task and the post text separated by the separation token [SEP].

At each time step, we choose the token with the highest probability over the vocabulary set to obtain the natural language caption. When recovering the fashion knowledge tuples from the caption, we first split the output sequence into several sentences. As shown in the top left part of Figure 2, the occasion information can then be extracted from the first sentence having the format of “`The occasion is `”. With respect to the remaining sentences, we extract the person attributes in the sentence, and pair them with all the fashion item information including the type and appearance in that sentence. According to the figure, for the sentence “`The first youth female wears a black upper`”, we can obtain the fashion knowledge tuple `(youth,female,upper,black)` from it following the rules. After extracting the fashion knowledge elements from the sequence, we compare them with the ground-truth label for evaluation. Notably, if the decoding fails, say the generated sequence violates the format, we treat the prediction as null.

3.5 Auxiliary Task Training

To obtain task-specific knowledge, we further design several auxiliary tasks including sentence reconstruction, visual question answering, and image-text matching. These tasks are designed to focus on one or several fashion knowledge aspects which can warm up the pre-trained model before training on the main captioning task. In order to fit to different auxiliary tasks, we assign different task prefixes to each task and add them before the original task text. The examples of the task prefix, task text, and target output text of each task are listed in Figure 3.

3.5.1 Sentence Reconstruction (SRC). Based on the assumption that different aspects in the fashion knowledge data are not strictly independent but strongly related (e.g. the type and appearance of the fashion item can be affected by the occasion), the goal of the SRC task is to predict some masked tokens based on their surrounding tokens and the image feature. Therefore, we randomly mask out 30% of the input tokens and ask the model to predict and reconstruct the original sentence. The task text is the masked fashion knowledge sequence and the output is the original full text. For each masked token, we replace it with the special mask token <mask>.



Advantages of going to work Cool, Disadvantages of going to work All but cool things top pants. YSIS new addition Summer see-through nitro supple #America Casual #Daily

Tasks	Task Prefix	Task Text	Target Text
SRC	Denoise:	The occasion is <mask>. The first youth female wears a <mask> upper and a dark blue <mask> and a black <mask> and a black belt.	The occasion is daily. The first youth female wears a black upper and a dark blue pants and a black footwear and a black belt.
ITM	Image text matching:	The occasion is wedding. The first youth female wears a white lace dress. The second youth male wears a dark blue pants and a black belt	False
VQA^1	Vqa:	What’s the occasion of the post	Daily
VQA^2		What’s the gender and age of the first person	Female youth
VQA^3		What does the first person wear	An upper and a pants and a footwear and a belt.
VQA^4		How does the upper of the first person look	Black

Figure 3: The task prefix and Input-Output formats of our three auxiliary tasks.

3.5.2 Image-Text Matching (ITM). This task takes a pair of image and natural language text as input. The model needs to determine if the text corresponds to the image or not. In our setting, we aim to determine if the given fashion knowledge caption corresponds to the post or not. We transform the original binary classification task into a generation problem following the rule that if the text is the corresponding caption of the post, the model generates “true”, while if not, the model generates “false”. We consider the ground-truth post-caption pair as positive samples. To construct negative samples, with the probability of 50%, we randomly sample the pseudo caption from another post in the training dataset.

3.5.3 Visual Question Answering (VQA). In the general visual question answering problem [1], the model aims to generate the answer of the input question for a given image. In our setting, we design four types of question-answering pairs which correspond to the occasion, person attributes, and fashion item aspects of each post. Each data sample in the training set has one-quarter probability of transforming into one of the four QA pairs during training. As listed in Figure 3, the first type of question has a fixed form, which says “what’s the occasion of the post”, and the answer is constructed based on the occasion category in the ground-truth training dataset. The second type of question focuses on the person information of the post, which asks the gender and age group of a random person in the post. For example, “what’s the gender and age of the first person”. The third and fourth types of questions are related to the type and appearance of a random fashion item in the post. With respect to the third type of question, we first randomly select a person in the post, then asks what he/she is wearing. Considering the appearance of the fashion items, with the random selection of one fashion item in the post, we set the format of the questions as “how does the [Fashion Item Type] of the [Person ID] person look”.

3.5.4 Multitask Training. For multitask training, we train the model on different auxiliary tasks with the pretrained parameter weights.

Occ.	**Wedding**	**Daily**	**School**	**Graduation**	**Sports**	**Vacation**
Post	1507	2543	931	2444	1769	1078
Person	2422	2877	1292	2050	2227	1337
Item	5684	7467	3580	5987	5960	3761

Person Attr.	**Male**				**Female**			
	Kid	Youth	Mid	Old	Kid	Youth	Mid	Old
Post	329	3618	610	84	358	6601	565	40
Person	885	9932	1529	206	901	17372	1504	110

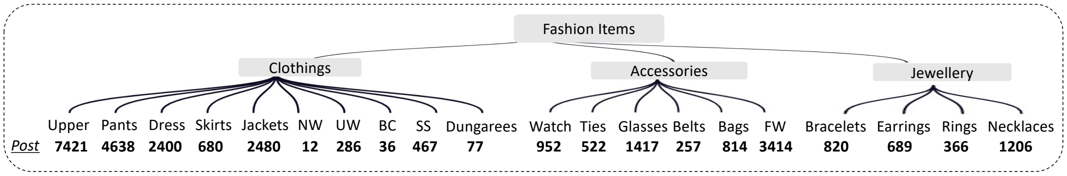


Figure 4: Detailed information of our dataset w.r.t different fashion knowledge aspects, where NW, UW, BC, SS, FW are the abbreviations of nightwear, underwear, babyclothes, swimsuits, and footwear respectively. From bottom to top, we report the number of fashion items, persons, and posts.

For each training step, we randomly sample a mini-batch from one of the three tasks. We differentiate the tasks by using different task prefixes. It is worth noting that the four subtasks of VQA are considered as the same task and share the same task prefix. Since we only change the input-output format without changing the pretrained model structure, we use the same loss function as in Section 3.4.3. We then set a bunch of weights according to the partial loss of each task to form the final loss.

$$L_{all} = \sum_{i=1}^{|T|} w_i L_i \tag{2}$$

Where $|T| = 3$ is the number of tasks, L_i is the partial loss according to each task, w_i is the hyperparameter representing the corresponding weight. After using with multiple relevant tasks to warm-up the model, the model is equipped with fashion-related knowledge, which can help tackle the main concerned FKE task.

4 EXPERIMENT

4.1 Dataset

Since there is no existing dataset that can be directly adopted to our setting, we collect and contribute a large-scale annotated dataset for the FKE task. Our dataset contains 9,272 posts with 32,439 fashion knowledge tuples in total, with an average of 3.5 fashion knowledge tuples per post and 2.7 fashion knowledge tuples per person. The detailed statistics of our dataset concerning different fashion knowledge aspects are reported in Figure 4.

Post Collection and Preprocessing Our dataset is collected from Instagram[2], which is a popular social media platform where large amount of posts are generated by users every day. To obtain the fashion-related posts, we first define six occasions, including school, graduation, sports, wedding, daily wear, and vacation. Under each occasion, we then choose some typical hashtags and crawl the related posts given the hashtag. After that, we filter out posts without any texts or containing only emojis. Since the raw text in social media is often noisy, we employ several text cleaning methods to deal with the crawled texts. We first preprocess the texts by removing all the unnecessary tokens including emojis, URL, whitespace, HTML characters, punctuation marks, and mentions. We then detect and translate all the text into English using the Google translate API [3].

Fashion Knowledge Annotation For the filtered fashion-related posts, we hire 10 fashion experts to manually annotate the fashion knowledge information for each post. The annotators first need to determine the occasion of the posts. Since sometimes similar images may result in different occasion results, before making the choice, the annotators are asked to read both texts and images to make sure that the occasion type is determined by both of them. After that, the annotators annotate the person attributes including the gender and age group in the images, as well as the type and appearance of the fashion items they wear. Considering the unfixed format of the appearance, when annotating it, the annotators are asked to use two or three words to describe how the fashion item looks, including its color, pattern, and texture. After all the annotations are finished, we ask two annotators to check the completeness and correctness of the results, making sure that all the fashion knowledge is correctly annotated in each post.

4.2 Comparison Methods

To validate the model effectiveness, we compare with both existing classification-based and generation-based methods. The first four are classification-based methods.

- **DARN** [12] is a Dual Attribute-aware Ranking Network originally used for retrieval feature learning. Same as in [25], we also only keep one stream for our task.
- **FashionNet** [21] is a pipeline-based model which simultaneously predicts landmarks and attributes. It consists of a global appearance branch, a local appearance branch and a pose branch.
- **HDF** [25] extracts the fashion knowledge from social media posts. It unifies three tasks of occasion, person and clothing discovery from multiple modalities of images, texts and metadata.
- **ViLBERT** [23] directly takes image text features as inputs and treats the task as a classification task. For occasion prediction, the input is the post image and text, and the output is the occasion category. While for fashion item information extraction, the

[2] https://www.instagram.com/

[3] https://pypi.org/project/googletrans/

Table 1: The experimental results of our model compared with the baseline methods, as well as the ablated results where text and image tags are removed from our model.

Model	Occasion				Category				Appearance				Overall			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
DARN [12]	44.1	40.2	42.6	42.4	25.1	73.2	47.1	57.4	10.3	64.2	40.8	50.0	9.1	23.5	14.4	17.9
FashionNet [21]	43.2	40.5	43.6	42.0	26.3	72.8	46.3	56.6	10.6	62.9	41.2	49.8	8.7	22.4	14.5	17.6
HDF [25]	50.3	47.4	43.7	45.5	29.4	77.1	52.7	62.8	14.3	68.8	44.0	53.7	12.1	27.9	17.6	21.6
ViLBERT [23]	59.6	50.3	58.4	54.1	32.5	80.1	53.6	64.2	15.2	71.3	52.9	60.7	12.5	28.7	20.4	23.5
Oscar [19]	75.6	75.2	76.0	75.5	7.7	21.1	33.7	26.0	7.1	20.1	23.2	21.5	5.5	15.2	17.4	16.2
VL-Bart [5]	75.2	69.1	74.9	71.4	30.8	80.9	48.6	60.7	17.8	52.9	31.6	39.6	15.4	35.6	21.9	27.1
Ours w/o text	69.2	64.6	70.1	68.8	32.7	78.5	64.2	70.6	20.4	71.8	57.7	64.0	15.4	33.6	28.2	30.7
Ours w/o img tags	72.8	68.4	73.0	70.6	30.8	75.6	64.2	69.4	18.1	70.2	57.0	62.9	16.0	35.1	28.4	31.4
Ours	**74.7**	**69.2**	**75.4**	**71.1**	**36.4**	**81.8**	**67.9**	**74.2**	**22.2**	**73.9**	**60.7**	**66.5**	**20.2**	**39.1**	**32.3**	**35.4**

image input is the fashion item box, the output is the type and appearance classes of the fashion item.

To further evaluate the effectiveness of our proposed captioning method, we also adopt the following generation-based baselines:

- **Oscar** [19]. We utilize Oscar to generate the fashion knowledge tuples. Oscar is BERT-like and does not have an encoder-decoder structure. The model is pre-trained on several classification tasks and one generation task (COCOCaption). During fine-tuning, the words in the tuples serve as input and are masked randomly at the rate of 15%. During inference, the generation process terminates when the model outputs the [STOP] token.
- **VL-Bart** [5]. We also construct a baseline which uses the same pre-trained VL-Bart model as ours but without the proposed captioning method. Specifically, we directly employ the fashion knowledge tuples in the natural language form as the target sequence, instead of transforming them into a natural language caption with the sentence transformation strategy.

4.3 Experimental Setting

In our experiment, we randomly split the dataset into the training, testing, and validation set with the percentage of 80%, 10%, 10%. We conduct 5 runs for our experiment, each with a different random seed and report the average score. When comparing our method with existing models, since most of the existing classification-based methods take the fashion item boxes as an input, we use an existing tool [47] to extract and predict all the person attributes in the post, following [25]. For each person box, we then use the same Faster-RCNN network mentioned in Section 3.3.2 to extract all the fashion items. Our code is based on PyTorch and Huggingface Transformers [40]. We use AdamW [22] with (β_1,β_2) = (0.9,0.999) and the learning rate 1e-4 with 5% linear warmup schedule. By default, each training process is run for 40 epochs. We report the results from the top-20 fashion item boxes with the confidence score greater than 0.5 from the original extraction results.

We use several evaluation metrics in our experiment. At the fashion item level, we report the precision, recall, F1 rate of each fashion item tuple. A tuple prediction is counted as correct only when all the elements are the same as the ground-truth label. We also report the post-wise accuracy score, which is the probability of post predictions our model got right. Except for that, to measure the semantic similarity between the generated caption and the transformed gold standard, we also employ some caption evaluation metrics including BLEU [30] and METEOR [2].

4.4 Main Experiment Results

Table 1 shows the main experiment results of our model and the baseline models. Except for the overall fashion tuple prediction performance ("Overall"), we also report the performance of the occasion, fashion item category and appearance prediction for a more comprehensive comparison. We have the following observations:

First of all, error propagation is the main problem in the pipeline-based methods. The inaccurate prediction of the person boxes can lead to the fashion item information prediction errors, affecting both the category, appearance and overall performance. Secondly, for pipeline-based models, there remains a gap between the precision and recall rate in most tasks. For example, the precision rate of HDF is 77.1 while the recall rate is 52.7 in the category prediction subtask. The gap mainly comes from the inaccurate fashion item box extraction, where many fashion items are not extracted or misextracted, thus leading to the low recall rate in the model result. In addition, some models take the dependencies between different fashion knowledge items into account (e.g. HDF), thus achieving better performance than those not (e.g. FashionNet). Moreover, it can also be observed that pre-trained models (e.g. ViLBERT) have a better performance than the non-pretrained models (e.g. DARN), showing the effectiveness of large pre-trained models in our task.

Our model achieves the best overall performance among all the methods. Although Oscar outperforms our method in the occasion prediction subtask, where each post contains only one occasion label belonging to one of the six categories, which does not require the model to have a strong generation ability. Oscar has difficulty in generating the more complex fashion item information, getting only 5.5 of the overall accuracy score. In addition, our model gets further improved by transforming the original tuple-like fashion knowledge into natural language sentences. Compared with directly generating the fashion knowledge tuples (i.e. VL-Bart), the overall F1 score improves from 27.1 to 35.4. The result proves that generating the natural language caption helps the generation model capture the dependencies between different fashion knowledge aspects, thus resulting in a better prediction.

To further study the role of different modalities in this task, we remove the post text and the image object tags respectively. Without post texts, the overall F1 score drops from 35.4 to 30.7. This

Table 2: Performance comparison regarding different auxiliary tasks, where base denotes directly fine-tuning on our dataset without post-training.

Setting	BLEU$_1$	BLEU$_2$	METEOR	Acc	F1
Base	69.77	64.51	38.17	13.76	30.43
+SRC	71.20	65.81	38.69	15.08	31.79
+ITM	71.90	66.05	38.44	15.31	32.04
+VQA	71.79	65.03	38.81	15.10	31.87
+ITM+VQA	72.49	66.68	38.90	18.23	34.01
+SRC+VQA	72.81	67.04	38.79	17.97	32.56
+SRC+ITM	73.46	67.61	38.86	18.54	33.87
Ours	75.40	69.29	39.80	20.22	35.43

result verifies that the post text contains rich fashion knowledge information with respect to where the post is located, who is in the post, and what the person is wearing. Besides, the image tags also play a vital role in our task, improving the overall performance from 31.4 to 35.4. The reason is that some tags (e.g. woman, dress) can be aligned to the corresponding image regions and provide hints for the fashion knowledge such as the person gender and fashion item categories. The results verify that essential information is contained in different modalities of the post, which can be effectively captured by our model.

4.5 Ablation Study

To evaluate the effect of different auxiliary tasks, we report the performance of our model with several variants. As shown in Table 2, we remove one or two auxiliary tasks at each time and report the corresponding accuracy and F1 scores. To further analyze the effects of those tasks on the generation results at the semantic level, BLEU and METEOR scores are also presented.

It can be noted that introducing auxiliary tasks improves the performance compared to directly fine-tuning the model on our dataset, which enhances the model with fashion-related knowledge. Among the three tasks, ITM is the most beneficial to the performance improvement, which improves the BLEU$_1$ and BLEU$_2$ scores by 2.13 and 1.54 percent. The reason is that the captions of different images are constructed by the same transformation method and share similar structure, recognizing the right caption from the negative image-caption pairs helps the model understand the fashion knowledge elements better. Compared with other tasks, removing SRC (corresponds to "+ITM+VQA" in the table) has the least influence on the F1 score. The reason is that when masking the caption, some less important tokens which appear with high frequency are masked with an equal probability with tokens containing rich fashion knowledge. For example, in the caption sentence "The woman wears a black upper", token "The" has the same probability of being masked as the token "upper".

4.6 Extensive Analysis

4.6.1 Performance under Different Person and Fashion Item Numbers. We analyze the performance of our model compared with the baseline models under different person and fashion item numbers, and plot the performance change in Figure 5. We can see that although the F1 score decreases for every method as the number of persons and fashion items grows, our model shows a greater advantage when it comes to the multi-person or multi-fashion-item setting. Specifically, when the number of persons and fashion items is small, both classification-based models and our model achieve a reasonable performance. However, as the case becomes more complicated, which means more persons are included in the image, traditional models often fail to extract all the fashion knowledge from the post. The performance gap between our method and a baseline HDF model reaches to the largest when there are 3 persons and 4 fashion items in the post image. Such failure on the first place, may result from the error propagation for the pipeline-based method, which means the inaccurate extraction of person boxes may give rise to the wrong prediction of all the fashion items associated with that person. On the other place, compared with our model, most traditional models fail to capture the relationship between different fashion knowledge elements. For example, the occasion "wedding" can be related to a young woman wearing a lacy white dress and a young man wearing a black suit. Our model captures such correlation by generating a caption where the occasion and person attributes are generated first, which provides some prior hints for the upcoming fashion item knowledge generation.

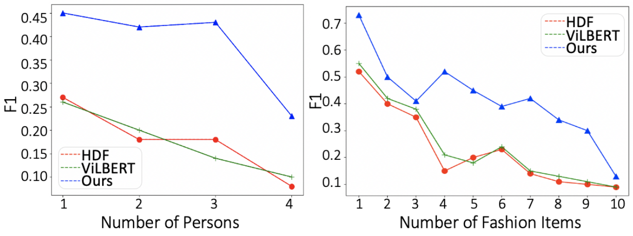
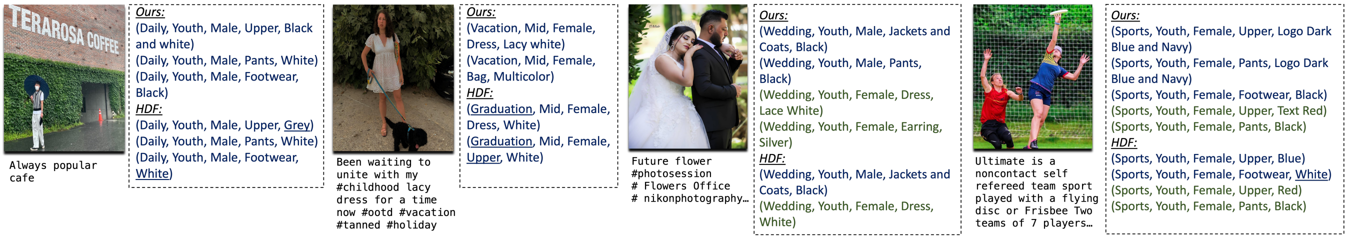


Figure 5: Performance comparison with respect to different person and fashion item numbers.

4.6.2 Generative v.s. Discriminative Methods. As can be observed from Table 1, generative methods generally achieve better performance than the classification type methods. To further investigate such a phenomenon, we break down the testing dataset into three groups, namely common, rare, and unseen set. Specifically, we define the testing fashion tuples appearing more than 5 times in the training set as the common set, those contained in the training set but appear less than 5 times as the rare set. For fashion knowledge tuples that never appear in the training set, we denote them as the unseen set. Table 3 shows the recall score of the three groups, which is the likelihood of the corresponding tuples being correctly predicted by the model.

As shown in the table, our model improves upon the discriminative baselines across all the tuple categories. This improvement is more significant on the rare data, where the recall score improves by 15.01 percent compared with ViLBERT. The result demonstrates the effectiveness of generative models in the FKE task, showing that when it comes to unfamiliar cases, generative models learn to describe the fashion items using the given knowledge compared to discriminative methods. What's more, by generating a natural language caption, the recall rate improves by 7.32 and 3.46 percent in rare and unseen cases, which proves that the interplay between



Always popular cafe

Ours:
(Daily, Youth, Male, Upper, Black and white)
(Daily, Youth, Male, Pants, White)
(Daily, Youth, Male, Footwear, Black)
HDF:
(Daily, Youth, Male, Upper, Grey)
(Daily, Youth, Male, Pants, White)
(Daily, Youth, Male, Footwear, White)

Been waiting to unite with my #childhood lacy dress for a time now #ootd #vacation #tanned #holiday

Ours:
(Vacation, Mid, Female, Dress, Lacy white)
(Vacation, Mid, Female, Bag, Multicolor)
HDF:
(Graduation, Mid, Female, Dress, White)
(Graduation, Mid, Female, Upper, White)

Future flower #photosession # Flowers Office # nikonphotography…

Ours:
(Wedding, Youth, Male, Jackets and Coats, Black)
(Wedding, Youth, Male, Pants, Black)
(Wedding, Youth, Female, Dress, Lace White)
(Wedding, Youth, Female, Earring, Silver)
HDF:
(Wedding, Youth, Male, Jackets and Coats, Black)
(Wedding, Youth, Female, Dress, White)

Ultimate is a noncontact self refereed team sport played with a flying disc or Frisbee Two teams of 7 players…

Ours:
(Sports, Youth, Female, Upper, Logo Dark Blue and Navy)
(Sports, Youth, Female, Pants, Logo Dark Blue and Navy)
(Sports, Youth, Female, Footwear, Black)
(Sports, Youth, Female, Upper, Text Red)
(Sports, Youth, Female, Pants, Black)
HDF:
(Sports, Youth, Female, Upper, Blue)
(Sports, Youth, Female, Footwear, White)
(Sports, Youth, Female, Upper, Red)
(Sports, Youth, Female, Pants, Black)

Figure 6: Real Case results of our model and HDF. Different color represents different person information. The underline in the tuple denotes the errors in the prediction.

Table 3: Recall rate of generative and discriminative methods on different test categories.

Method	Common	Rare	Unseen	Overall
Discriminative				
HDF	18.32	6.07	1.79	17.62
ViLBERT	22.31	6.21	1.97	20.41
Generative				
VL-Bart	26.21	13.90	3.51	21.85
Ours	34.63	21.22	6.97	32.28

image and text can be better captured compared with generating fashion knowledge tuples directly.

4.6.3 Caption Construction Analysis. Our proposed sentence construction method transforms the original fashion knowledge tuples to a natural language caption for the sequence-to-sequence mapping. To verify the effectiveness of such design, we also perform experiments based on different caption construction strategies and report the accuracy in Table 4. We use three different caption construction rules in the experiment. Some rules are designed to combine the fashion knowledge tuples in a less compact way (e.g. Rule 1 and 2). For example, we use one sentence to describe each fashion knowledge tuple respectively. We also design some rules where different fashion knowledge aspects are separated (e.g. Rule 3), where we follow the order of occasion first, person next, fashion item last when constructing the caption.

To better demonstrate the algorithms of them, we use one example to illustrate. With the input of three fashion knowledge tuples `(daily, P1, male, kid, upper, black)`, `(daily, P1, male, kid, pants, white)`, `(daily, P2, female, old, dress, blue)`, the outputs of them are as follows:

- Rule 1 *The first male kid wears a black upper in daily. The first male kid wears a white pants in daily. The second female old wears a blue dress in daily.*
- Rule 2 *The first male kid wears a black upper and a white pants in daily. The second female old wears a blue dress in daily.*
- Rule 3 *The occasion is daily. The person is a male kid and a female old. The first person wears a black upper and a white pants. The second person wears a blue dress.*
- Ours *The occasion is daily. The first male kid wears a black upper and a white pants. The second female old wears a blue dress.*

According to the results, our caption construction method achieves the best performance in all aspects. Rule 1 and 2 both put the occasion information at the end of each sentence. However, we find that it may pose a negative influence on the occasion prediction. Compared with Rule 3 where the person and fashion item information is separated, our method has a more compact form and helps to better capture the interplay between different aspects, improving the overall accuracy from 18.2 to 20.2.

Table 4: Analysis of different caption construction strategies.

	Occ.	Cat.	App.	Overall
Rule 1	72.5	36.0	21.8	18.0
Rule 2	73.6	36.1	22.0	19.1
Rule 3	74.2	35.8	22.1	18.2
Ours	74.7	36.4	22.2	20.2

4.7 Case Study

We use some real cases to compare the performance of our model with the HDF model in a more vivid way. As shown in Figure 6, there are more errors in the HDF extraction results compared with our model. For example, the appearance of the upper in the first case is misclassified as grey. In addition, our model captures the interplay between the image and text information better. For example, in the second case, the post text “*lacy dress*” corresponds to the dress in the image and the hashtag indicates that the occasion should be vacation. What’s more, our model provides more comprehensive results. For example, in the third case, the HDF model fails to extract the less obvious earring information in the image and also ignores the pants the man wears. Also concerning the appearance of the fashion items, our model outputs better description against the HDF model. As shown in the third case, our model describes the dress of the woman as “*lacy white*”, while the HDF model only classifies the dress as “*white*”.

5 CONCLUSION

We investigate social media based FKE task. For a social media post consisting of an image and text, we aim to elicit the occasion, person attributes, and fashion item information from the post. Specifically, we formulate this task as a captioning problem and transform the fashion knowledge tuples into a natural language caption. We also design several auxiliary tasks before captioning to warm-up the model with task-specific knowledge. Since no existing dataset can be directly adapted to our task, we contribute a large-scale dataset with manual annotation. Extensive experiments are conducted to demonstrate the effectiveness of our model.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[4] Huizhong Chen, Andrew Gallagher, and Bernd Girod. 2012. Describing clothing by semantic attributes. In *European conference on computer vision*. Springer, 609–623.

[5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *ICML*.

[6] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

[7] Yang Deng, Yaliang Li, Bolin Ding, and Wai Lam. 2022. Leveraging Long Short-Term User Preference in Conversational Recommendation Via Multi-Agent Reinforcement Learning. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[8] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified Conversational Recommendation Policy Learning via Graph-based Reinforcement Learning. In *SIGIR 2021*. 1431–1441.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering Text-to-Image Generation via Transformers. *arXiv preprint arXiv:2105.13290* (2021).

[11] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*. 129–138.

[12] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*. 1062–1070.

[13] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561* (2021).

[14] Menglin Jia, Yichen Zhou, Mengyun Shi, and Bharath Hariharan. 2018. A deep-learning-based fashion attributes detection model. *arXiv preprint arXiv:1810.10148* (2018).

[15] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. 105–112.

[16] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 207–216.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[18] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *EMNLP*.

[19] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *ECCV 2020* (2020).

[20] Jingyuan Liu and Hong Lu. 2018. Deep Fashion Analysis with Feature Map Upsampling and Landmark-Driven Attention. In *European Conference on Computer Vision*. Springer, 30–36.

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[24] Yunshan Ma, Yujuan Ding, Xun Yang, Lizi Liao, Wai Keung Wong, and Tat-Seng Chua. 2020. Knowledge enhanced neural fashion trend forecasting. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 82–90.

[25] Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. 2019. Who, where, and what to wear? Extracting fashion knowledge from social media. In *Proceedings of the 27th ACM International Conference on Multimedia*. 257–265.

[26] Utkarsh Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. 2019. Geostyle: Discovering fashion trends and events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 411–420.

[27] Kevin Matzen, Kavita Bala, and Noah Snavely. 2017. Streetstyle: Exploring world-wide clothing styles from millions of photos. *arXiv preprint arXiv:1706.01869* (2017).

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[31] Alec Radford, Ilya Sutskever, Gretchen Krueger Jong Wook Kim, and Sandhini Agarwal. 2021. CLIP: Connecting Text and Images.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[35] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[36] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 7463–7472.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[39] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4271–4280.

[40] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45.

[41] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11307–11317.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[43] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2017. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM international conference on Multimedia*. 172–180.

[44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[45] Yifei Yuan and Wai Lam. 2021. Conversational Fashion Image Retrieval via Multiturn Natural Language Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[46] Yifei Yuan and Wai Lam. 2021. Sentiment Analysis of Fashion Related Posts in Social Media. *Proceedings of the Fifteenth ACM International Conference on*

Web Search and Data Mining (2021). https://api.semanticscholar.org/CorpusID: 244117093

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[48] Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect Sentiment Quad Prediction as Paraphrase Generation. In *EMNLP 2021*. 9209–9219.

[49] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards Generative Aspect-Based Sentiment Analysis. In *ACL/IJCNLP 2021*. 504–510.

[50] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2023. A Comprehensive Survey on Deep Learning for Relation Extraction: Recent Advances and New Frontiers. *CoRR* abs/2306.02051 (2023).