Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback

Hui Wu[*1,2] Yupeng Gao[*2] Xiaoxiao Guo[*1,2] Ziad Al-Halah[3]
Steven Rennie[4] Kristen Grauman[3] Rogerio Feris[1,2]

[1] MIT-IBM Watson AI Lab [2] IBM Research [3] UT Austin [4] Pryon

Abstract

Conversational interfaces for the detail-oriented retail fashion domain are more natural, expressive, and user friendly than classical keyword-based search interfaces. In this paper, we introduce the Fashion IQ dataset to support and advance research on interactive fashion image retrieval. Fashion IQ is the first fashion dataset to provide human-generated captions that distinguish similar pairs of garment images together with side-information consisting of real-world product descriptions and derived visual attribute labels for these images. We provide a detailed analysis of the characteristics of the Fashion IQ data, and present a transformer-based user simulator and interactive image retriever that can seamlessly integrate visual attributes with image features, user feedback, and dialog history, leading to improved performance over the state of the art in dialog-based image retrieval. We believe that our dataset will encourage further work on developing more natural and real-world applicable conversational shopping assistants.[1]

1. Introduction

Fashion is a multi-billion-dollar industry, with direct social, cultural, and economic implications in the world. Recently, computer vision has demonstrated remarkable success in many applications in this domain, including trend forecasting [2], modeling influence relations [1], creation of capsule wardrobes [23], interactive product retrieval [18, 68], recommendation [41], and fashion design [46]. In this work, we address the problem of interactive image retrieval for fashion product search. High fidelity interactive image retrieval, despite decades of research and many great strides, remains a research challenge. At the crux of the challenge are two entangled elements: empowering the user with ways to express what they want, and empowering the retrieval machine with the information, capacity, and learning objective to realize high performance.

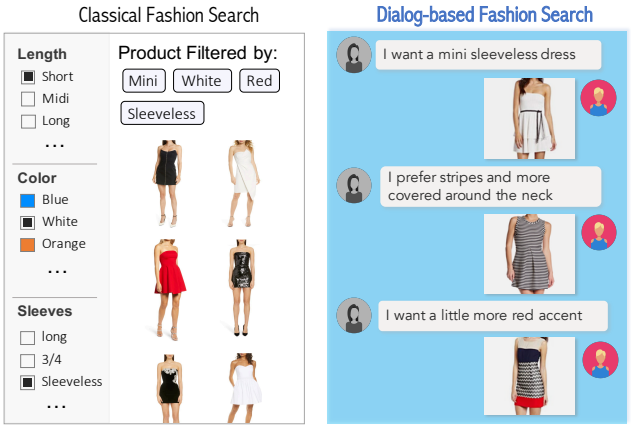

Figure 1: A classical fashion search interface relies on the user selecting filters based on a pre-defined fashion ontology. This process can be cumbersome and the search results still need manual refinement. The Fashion IQ dataset supports building dialog-based fashion search systems, which are more natural to use and allow the user to precisely describe what they want to search for.

To tackle these challenges, traditional systems have relied on relevance feedback [47, 68], allowing users to indicate which images are "similar" or "dissimilar" to the desired image. Relative attribute feedback (e.g., "more formal than these", "shinier than these") [33, 32] allows the comparison of the desired image with candidate images based on a fixed set of attributes. While effective, this specific form of user feedback constrains what the user can convey. More recent work utilizes natural language to address this problem [65, 18, 55], with relative captions describing the differences between a reference image and what the user has in mind, and dialog-based interactive retrieval as a principled and general methodology for interactively engaging the user in a multimodal *conversation* to resolve their intent

[*] Equal contribution.
[1] Fashion IQ is available at: https://github.com/XiaoxiaoGuo/fashion-iq

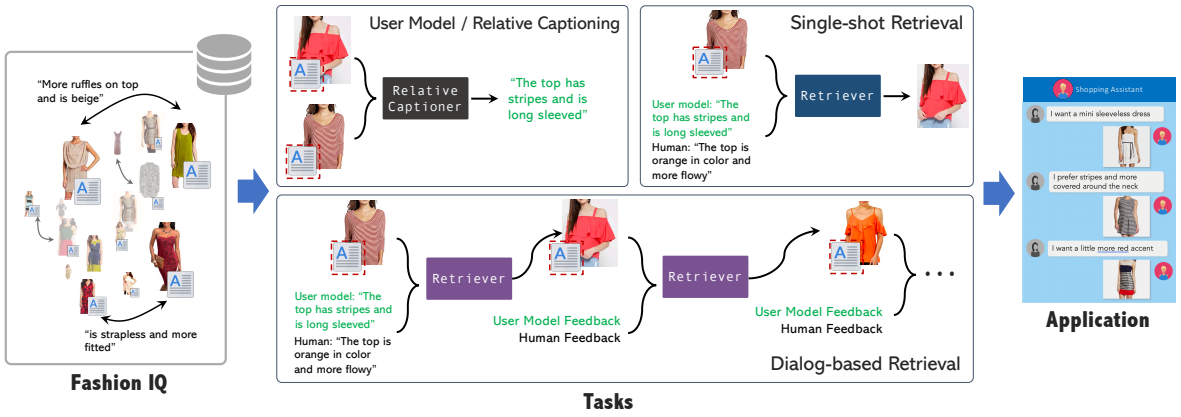

Figure 2: Fashion IQ can be used in three scenarios: user modeling based on relative captioning, and single-shot as well as dialog-based retrieval. Fashion IQ uniquely provides both annotated user feedback (black font) and visual attributes derived from real-world product data (dashed boxes) for system training.

[18]. When empowered with natural language feedback, the user is not bound to a pre-defined set of attributes, and can communicate compound and more specific details during each query, which leads to more effective retrieval. For example, with the common attribute-based interface (Figure 1 left) the user can only define what kind of attributes the garment has (e.g., white, sleeveless, mini), however with interactive and relative natural language feedback (Figure 1 right) the user can use comparative forms (e.g., more covered, brighter) and fine-grained compound attribute descriptions (e.g., red accent at the bottom, narrower at the hips).

While this recent work represents great progress, several important questions remain. In real-world fashion product catalogs, images are often associated with *side information*, which in the wild varies greatly in format and information content, and can often be acquired at large scale with low cost. Descriptive representations such as attributes can often be extracted from this data, and form a strong basis for generating stronger image captions [71, 66, 70] and more effective image retrieval [25, 5, 51, 34]. How such side information interacts with natural language user inputs to improve the state of the art dialog-based image retrieval systems are important open research questions. Furthermore, a challenge with implementing dialog-interface image search systems is that currently conversational systems typically require cumbersome hand-engineering and/or large-scale dialog data [35, 6]. In this paper, we investigate the extent to which side information can alleviate these issues, and explore ways to incorporate side information in the form of visual attributes into model training to improve interactive image retrieval. This represents an important step toward the ultimate goal of constructing commercial-grade conversational interfaces with much less data and effort, and much wider real-world applicability.

Toward this end, we contribute a new dataset, Fashion Interactive Queries (*Fashion IQ*). Fashion IQ is distinct from existing fashion image datasets (see Figure 4) in that it uniquely enables joint modeling of natural language feedback and side information to realize effective and practical image retrieval systems. As we illustrate in Figure 2, there are two main settings to utilize Fashion IQ to drive progress on developing more effective interfaces for image retrieval: *single-shot retrieval* and *dialog-based retrieval*. In both settings, the user can communicate their fine-grained search intent via natural language relative feedback. The difference of the two settings is that dialog-based retrieval can progressively improve the retrieval results over the interaction rounds. Fashion IQ also enables *relative captioning*, which we leverage as a user model to efficiently generate a large amount of low-cost training data, to further improve training interactive fashion retrieval systems.[2]

To summarize, our main contributions are as follows:

- We introduce a novel dataset, Fashion IQ, a publicly available resource for advancing research on conversational fashion retrieval. Fashion IQ is the first fashion dataset that includes both human-written relative captions that have been annotated for similar pairs of images, and the associated real-world product descriptions and attribute labels as side information.
- We present a transformer-based user simulator and interactive image retriever that can seamlessly leverage multimodal inputs (images, natural language feedback, and attributes) during training, and leads to significantly improved performance. Through the use of self-attention, these models consolidate the traditional components of user modeling and interactive retrieval, are highly extensible, and outperform existing methods for the relative captioning and interactive image retrieval of fashion images on Fashion IQ.

[2]Relative captioning is also a standalone vision task [26, 57, 43, 15], which Fashion IQ serves as a new training and benchmarking dataset.

- To the best of our knowledge, this is the first study to investigate the benefit of combining natural language user feedback and attributes for dialog-based image retrieval, and it provides empirical evidence that incorporating attributes results in superior performance for both user modeling and dialog-based image retrieval.

2. Related Work

Fashion Datasets. Many fashion datasets have been proposed over the past few years, covering different applications such as fashionability and style prediction [50, 28, 22, 51], fashion image generation [46], product search and recommendation [25, 72, 19, 41, 63], fashion apparel pixelwise segmentation [27, 74, 69], and body-diverse clothing recommendation [24]. DeepFashion [38, 16] is a large-scale fashion dataset containing consumer-commercial image pairs and labels such as clothing attributes, landmarks, and segmentation masks. iMaterialist [17] is a large-scale dataset with fine-grained clothing attribute annotations, while Fashionpedia [27] has both attribute labels and corresponding pixelwise segmented regions.

Unlike most existing fashion datasets used for image retrieval, which focus on content-based or attribute-based product search, our proposed dataset facilitates research on *conversational* fashion image retrieval. In addition, we enlist real users to collect the high-quality, natural language annotations, rather than using fully or partially automated approaches to acquire large amounts of weak attribute labels [41, 38, 46] or synthetic conversational data [48]. Such high-quality annotations are more costly, but of great benefit in building and evaluating conversational systems for image retrieval. We make the data publicly available so that the community can explore the value of combining high-quality human-written relative captions and the more common, web-mined weak annotations.

Visual Attributes for Interactive Fashion Search. Visual attributes, including color, shape, and texture, have been successfully used to model clothing images [25, 22, 23, 2, 73, 7, 40]. More relevant to our work, in [73], a system for interactive fashion search with attribute manipulation was presented, where the user can choose to modify a query by changing the value of a specific attribute. While visual attributes model the presence of certain visual properties in images, they do not measure the relative strength of them. To address this issue, relative attributes [42, 52] were proposed, and have been exploited as a richer form of feedback for interactive fashion image retrieval [32, 33, 30, 31]. However, in general, attribute based retrieval interfaces require careful curation and engineering of the attribute vocabulary. Also, when attributes are used as the sole interface for user queries, they can lead to inferior performance relative to both relevance feedback [44] and natural language feedback [18]. In contrast with attribute based systems, our work explores the use of relative feedback in *natural language*, which is more flexible and expressive, and is complementary to attribute based interfaces.

Image Retrieval with Natural Language Queries. Methods that lie in the intersection of computer vision and natural language processing, including image captioning [45, 64, 67] and visual question-answering [3, 10, 59], have received much attention from the research community. Recently, several techniques have been proposed for image or video retrieval based on natural language queries [36, 4, 60, 65, 55]. In another line of work, visually-grounded dialog systems [11, 53, 13, 12] have been developed to hold a meaningful dialog with humans in natural, conversational language about visual content. Most current systems, however, are based on purely text-based questions and answers regarding a single image. Similar to [18], we consider the setting of goal-driven dialog, where the user provides feedback in natural language, and the agent outputs retrieved images. Unlike [18], we provide a large dataset of relative captions anchored with real-world contextual information, which is made available to the community. In addition, we follow a very different methodology based on a unified transformer model, instead of fragmented components to model the state and flow of the conversation, and show that the joint modeling of visual attributes and relative feedback via natural language can improve the performance of interactive image retrieval.

Learning with Side Information. Learning with privileged information that is available at training time but not at test time is a popular machine learning paradigm [61], with many applications in computer vision [49, 25]. In the context of fashion, [25] showed that visual attributes mined from online shopping stores serve as useful privileged information for cross-domain image retrieval. Text surrounding fashion images has also been used as side information to discover attributes [5, 20], learn weakly supervised clothing representations [51], and improve search based on noisy and incomplete product descriptions [34]. In our work, for the first time, we explore the use of side information in the form of visual attributes for image retrieval with a natural language feedback interface.

3. Fashion IQ Dataset

One of our main objectives in this work is to provide researchers with a strong resource for developing interactive dialog-based fashion retrieval models. To that end, we introduce a novel public benchmark, Fashion IQ. The dataset contains diverse fashion images (dresses, shirts, and tops&tees), side information in form of textual descriptions and product meta-data, attribute labels, and most importantly, large-scale annotations of high quality relative captions collected from human annotators. Next we describe the data collection process and provide an in-depth analy-

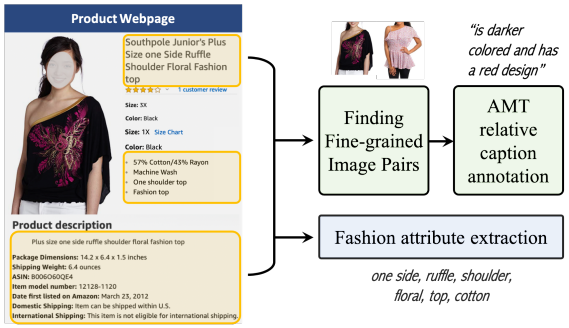

Figure 3: Overview of the dataset collection process.

	#Image	# With Attr	# Relative Cap.
Dresses	19,087	12,955	20,052
Shirts	31,728	20,071	20,130
Tops&Tees	26,869	16,438	20,090

Table 1: Dataset statistics. We use 6:2:2 splits for each category for training, validation and testing, respectively.

sis of Fashion IQ. The overall data collection procedure is illustrated in Figure 3.

3.1. Image And Attribute Collection

The images of fashion products that comprise our Fashion IQ dataset were originally sourced from a product review dataset [21]. Similar to [2], we selected three categories of product items, specifically: Dresses, Tops&Tees, and Shirts. For each image, we followed the link to the product website available in the dataset, in order to extract corresponding product information.

Leveraging the rich textual information contained in the product website, we extracted fashion attribute labels from them. More specifically, product attributes were extracted from the product title, the product summary, and detailed product description. To define the set of product attributes, we adopted the fashion attribute vocabulary curated in DeepFashion [38], which is currently the most widely adopted benchmark for fashion attribute prediction. In total, this resulted in 1000 attribute labels, which were further grouped into five attribute types: texture, fabric, shape, part, and style. We followed a similar procedure as in [38] to extract the attribute labels: an attribute label for an image is considered as present if its associated attribute word appears at least once in the metadata. In Figure 4, we provide examples of the original side information provided in the product review dataset and the corresponding attribute labels that were extracted. To complete and denoise attributes, we use an attribute prediction model pretrained on DeepFashion (details in Appendix A).

Semantics	Quantity	Examples
Direct reference	49%	is solid white and buttons up with front pockets
Comparison	32%	**has longer sleeves** and is **lighter in color**
Direct & compar.	19%	has a geometric print with **longer sleeves**
Single attribute	30.5%	**is more bold**
Composite attr.	69.5%	black **with** red cherry pattern **and** a deep V neck line
Negation	3.5%	is white colored with a graphic and **no lace design**

Table 2: Analysis on the relative captions. Bold font highlights comparative phrases.

3.2. Relative Captions Collection

The Fashion IQ dataset is constructed with the goal of advancing conversational image search. Imagine a typical visual search process (illustrated in Figure 1): a user might start the search by describing general keywords which can weed out totally irrelevant search instances, then the user can construct natural language phrases which are powerful in specifying the subtle differences between the search target and the current search result. In other words, relative captions are more effective to narrow down fine-grained cases than using keywords or attribute label filtering.

To ensure that the relative captions can describe the fine-grained visual differences between the reference and target image, we leveraged product title information to select similar images for annotation with relative captions. Specifically, we first computed the TF-IDF score of all words appearing in each product title, and then for each target image, we paired it with a reference image by finding the image in the database (within the same fashion category) with the maximum sum of the TF-IDF weights on each overlapping word. We randomly selected $\sim$10,000 target images for each of the three fashion categories, and collected two sets of captions for each pair. Inconsistent captions were filtered (please consult the suppl. material for details).

To amass relative captions for the Fashion IQ data, we collected data using crowdsourcing. Briefly, the users were situated in the context of an online shopping chat window, and assigned the goal of providing a natural language expression to communicate to the shopping assistant the visual features of the search target as compared to the provided search candidate. Figure 4 shows examples of image pairs presented to the user, and the resulting relative image captions that were collected. We only included workers from three predominantly English-speaking countries, with master level of expertise and with an acceptance rate above 95%. This criterion makes it more costly to obtain the captions, but ensures that the human-written captions in Fashion IQ are indeed of high quality. To further improve

(a) Examples of the textual descriptions and attribute labels

(b) Examples of relative captions, i.e., natural-language relative feedback.

Figure 4: Fashion IQ uniquely provides natural-language relative feedback, textural descriptions and fashion attributes.

the quality of the annotations and speed up the annotation process, the prefix of the relative feedback "Unlike the provided image, the one I want" is provided with the prompt, and the user only needs to provide a phrase that focuses on the visual differences of the given image pairs.

3.3. Dataset Summary and Analysis

Basic statistics and examples of the resulting Fashion IQ dataset are in Table 1 and Figure 4, with additional details presented in Appendix A, including dataset statistics by each data split, the word-frequency clouds of the relative captions, and the distributions of relative caption length and number of attributes per image. As depicted in Figure 2, Fashion IQ can be applied to three different tasks, *single-shot retrieval*, *dialog-based retrieval* and *relative captioning*. These tasks can be developed independently or jointly to drive the progress on developing more effective interfaces for image retrieval. We provide more explanations on how Fashion IQ can be applied to each task in Appendix B. In the main paper, we focus on the multi-turn retrieval setting, which includes the *dialog-based retrieval* task and the *relative captioning* task. Appendix C includes auxiliary study on the *single-shot retrieval* task.

Relative captions vs attributes. The length of relative captions and the number of attributes per image of Fashion IQ have similar distributions across all three categories (c.f. Figure 8 in the Appendix). In most cases, the attribute labels and relative captions contain complementary information, and thus jointly form a stronger basis for ascertaining the relationships between images. To further obtain insight on the unique properties of the relative captions in comparison with classical attribute labels, we conducted a semantic analysis on a subset of 200 randomly chosen relative captions. The results of the analysis are summarized in Table 2. Almost 70% of all text queries in Fashion IQ consist of compositional attribute phrases. Many of the captions are simpler adjective-noun pairs (e.g. "red cherry pattern"). Nevertheless, this structure is more complex than a simple "bag of attributes" representation, which can quickly become cumbersome to build, necessitating a large vocabulary and compound attributes, or multi-step composition. Furthermore, in excess of 10% of the data involves more complicated compositions that often include direct or relative spatial references for constituent objects (e.g. "pink stripes on side and bottom"). The analysis suggests that relative captions are a more expressive and flexible form of annotation than attribute labels. The diversity in the structure and content of the relative captions provide a fertile resource for modeling user feedback and for learning natural language feedback based image retrieval models, as we will demonstrate below.

4. Multimodal Transformers for Interactive Image Retrieval

To advance research on the Fashion IQ applications, we introduce a strong baseline for dialog-based fashion re-

trieval based on the modern transformer architecture [62]. Multimodal transformers have recently received significant attention, achieving state-of-the-art results in vision and language tasks such as image captioning and visual-question answering [75, 56, 37, 54, 39]. To the best of our knowledge, multimodal transformers have not been studied in the context of goal-driven dialog-based image retrieval. Specifically, we adapt the transformer architecture in (1) a relative captioner transformer, which is then used as a user simulator to train our interactive retrieval system, and (2) a multimodal retrieval framework, which incorporates image features, fashion attributes, and a user's textual feedback in a unified fashion. This unified retrieval architecture allows for more flexibility in terms of included modalities compared to the RNN-based approaches (e.g., [18]) which may require a systemic revision whenever a new modality is included. For example, integrating visual attributes into traditional goal-driven dialog architectures would require specialization of each individual component to model the user response, track the dialog history, and generate responses.

4.1. Relative Captioning Transformer

In the relative captioning task, the model is given a reference image I_r and a target image I_t and it is tasked with describing the differences of I_r relative to I_t in natural language. Our transformer model leverages two modalities: image visual feature and inferred attributes (Figure 5). While the visual features capture the fine-grained differences between I_r and I_t, the attributes help in highlighting the prominent differences between the two garments. Specifically, we encode each image with a CNN encoder $f_I(\cdot)$, and to obtain the prominent set of fashion attributes from each image, we use an attribute prediction model $f_A(\cdot)$ and select the top $N = 8$ predicted attributes from the reference $\{a_i\}^r$ and the target $\{a_i\}^t$ images based on confidence scores from $f_A(I_r)$ and $f_A(I_t)$, respectively. Then, each attribute is embedded into a feature vector based on the word encoder $f_W(\cdot)$. Finally, our transformer model attends to the difference in image features of I_r and I_t and their attributes to produce the relative caption $\{w_i\} = f_R(I_r, I_t) = (f_I(I_r) - f_I(I_t), f_W(\{a_i\}^r), f_W(\{a_i\}^t))$, where $\{w_i\}$ is the word sequence generated for the caption.

4.2. Dialog-based Image Retrieval Transformer

In this interactive fashion retrieval task, to initiate the interaction, the system can either select a random image (which assumes no prior knowledge on the user's search intent), or retrieve an image based on the keywords-based query from the user. Then at each turn, the user provides textual feedback based on the currently retrieved image to guide the system towards a target image, and the system responds with a new retrieved image, based on all of the user feedback received so far. Here we adopt a transformer

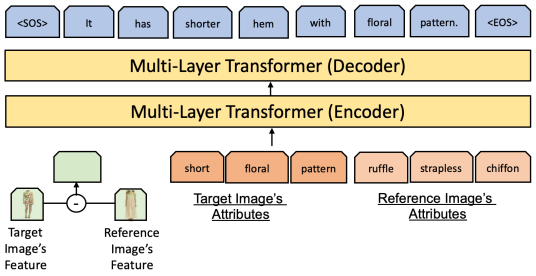

Figure 5: Our multimodal transformer model for relative captioning, which is used as a user simulator for training our interactive image retrieval system.

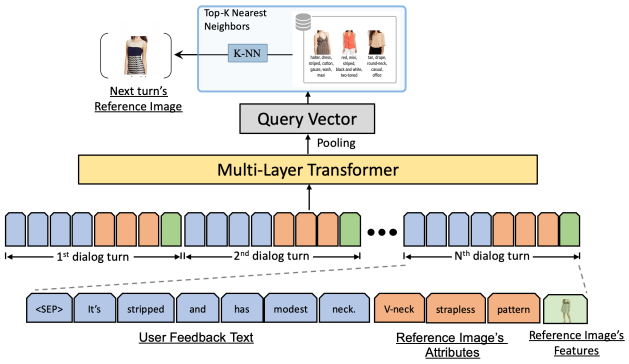

Figure 6: Our multimodal transformer model for image retrieval, which integrates, through self-attention, visual attributes with image features, user feedback, and the entire dialog history during each turn, in order to retrieve the next candidate image.

architecture that enables our model to attend to the entire multimodal history of the dialog during each dialog turn. This is in contrast with RNN-based models (e.g., [18]), which must systemically incorporate features from different modalities, and consolidate historical information into a low-dimensional feature vector.

During training, our dialog-based retrieval model leverages the previously introduced relative captioning model to simulate the user's input at the start of each cycle of the interaction. More specifically, the user model is used to generate relative captions for image pairs that occur during each interaction (which are generally not present in the training data of the captioning model), and enables efficient training of the interactive retriever without a human in the loop as was done in [18]. For commercial applications, this learning procedure would serve as pre-training to bootstrap and then boost system performance, as it is fine-tuned on real multi-turn interaction data that becomes available. The relative captioning model provides the dialog-based retriever at each iteration j with a relative description of the differences between the retrieved image I_j and the target image I_t. Note that only the user model f_R has access to I_t, and f_R

communicates to the dialog model f_D only through natural language. Furthermore, to prevent f_R and f_D from developing a coded language among themselves, we pre-train f_R separately on relative captions, and freeze the model parameters when training f_D.

To that end, at the J-th iteration of the dialog, f_D receives the user model's relative feedback $\{w_i\}^J = f_R(I_J, I_t)$, the top N attributes from I_J, and image features of I_J (see Figure 6). The model attends to these features and features from previous interactions with a multi-layer transformer to produce a query vector $q_J = f_D(\{\{w_i\}^j, f_W(\{a_i\}^j), f_I(I_j)\}_{j=1}^J)$. We follow the standard multi-head self-attention formulation [62]: $\text{head}_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V)$, and the output at each layer is $\text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$. The output at the last layer is q_J, which is used to search the database for the best matching garment based on the Euclidean distance in image feature vector space. The top searched image is returned to the user for the next iteration, denoted as I_{J+1} .

5. Experiments

We evaluate our multimodal transformer models on the user simulation and interactive fashion retrieval tasks of Fashion IQ. We compare against the state-of-the-art hierarchical RNN-based approach from [18] and demonstrate the benefit of the design choices of our baselines and the newly introduced attributes in boosting performance. All models are evaluated on the three fashion categories: Dresses, Shirts and Tops&Tees, following the same data split shown in Table 1. These models establish formal performance benchmarks for the user modeling and dialog-based retrieval tasks of Fashion IQ, and outperform those of [18], even when not leveraging attributes as side information (cf. Tables 3, 4).

5.1. Experiment Setup

Input Encoders. We train an EfficientNet-b7 [58] on the attribute prediction task from the DeepFashion dataset, and we use the features from the last average pooling layer of that network to realize the image encoder f_I. For the attribute model f_A, we fine-tune the last linear layer of the previous EfficientNet-b7 using the attribute labels from our Fashion IQ dataset, and use the top-8 attribute predictions for the garment images as input. Finally, for f_W we use randomly initialized embeddings and optimize these end-to-end with other components. We use GloVe [3] to encode user feedback words in the retriever.

Transformer Details. The multimodal retrieval model is a 6-layer transformer (256 hidden units, 8 attention heads)[4].

[3]https://nlp.stanford.edu/projects/glove

[4]Our transformer implementation is based on the Harvard NLP library (https://nlp.seas.harvard.edu/2018/04/03/attention).

	Dialog Turn 1			Dialog Turn 5		
	P	R@10	R@50	**P**	R@10	R@50
Dresses						
[18]	89.45	6.25	20.26	98.56	39.12	72.21
Ours	93.14	12.45	35.21	98.39	41.35	73.63
Ours+Attr	**93.50**	**13.39**	**35.56**	**98.69**	**46.28**	**77.24**
Shirts						
[18]	89.39	3.86	13.95	98.48	32.94	62.03
Ours	92.75	**11.05**	28.99	98.28	**33.91**	63.42
Ours+Attr	**92.92**	11.03	**29.03**	**98.46**	33.69	**64.60**
Tops&Tees						
[18]	87.89	3.03	12.34	98.30	29.59	60.82
Ours	93.03	11.24	30.45	98.22	33.52	63.85
Ours+Attr	**93.25**	**11.74**	**31.52**	**98.44**	**35.94**	**66.56**

Table 3: **Dialog-based Image Retrieval.** We report the performance on ranking percentile (P) and recall at N (R@N) at the 1st and 5th dialog turns.

The user's feedback text is padded to a fixed length of 8. The transformer's output representations are then pooled and linearly transformed to form the query vector. All other parameters are set to their default values. The multimodal captioning model has 6 encoding and 6 decoding transformer layers and its caption output is set to maximum word length of 8. The captioner's loss function is the cross entropy and the retrieval's is the triplet-based loss as defined in [18]. Specifically, the retrieval component minimizes the triplet-based loss over N dialog turns, $\sum_{j=1}^{N} \max(0, ||q_j - f_I(I_t)||_2 - ||q_j - f_I(I_n)||_2 + m)$, where I_t and I_n are the target image and a random image respectively. m is the constant for the margin. For further details regarding model training please consult Appendix D.

Evaluation Setup. To reduce the evaluation variance, we randomly generate a list of initial image pairs (i.e., a target and a reference image), and we evaluated all methods with the same list of the initial image pairs. We use Beam Search with beam size 5 to generate the relative captions as feedback to the retriever model. When training the retriever models, we use greedy decoding for faster training speed. The average ranking percentile is computed as $P = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{r_i}{N})$, where r_i is the ranking of the i-th target and N is the total number of candidates. Furthermore, to ablate the impact of the attribute modality, we consider two version of our approach that correspond to the exact same model with or without additional attribute features as input: **Ours** and **Ours+Attr**.

5.2. Experimental Results

Relative Captioning. Table 4 summarizes the performance of our multimodal transformer approach compared to the RNN-based approach from [18]. Our transformer method outperforms the RNN-based baseline across all metrics. Moreover, the attribute-aware transformer model improves

	BLEU-4	Rouge-L	CIDEr	SPICE
Dresses				
Guo et al. [18]	17.4	53.6	48.9	32.1
Ours	20.7	56.3	78.5	34.4
Ours+Attr	**21.1**	**57.1**	**80.6**	**36.1**
Shirts				
Guo et al. [18]	19.6	53.8	52.6	32.0
Ours	22.3	56.4	84.1	34.7
Ours+Attr	**24.2**	**57.5**	**92.1**	**35.4**
Tops&Tees				
Guo et al. [18]	15.7	50.5	41.1	30.6
Ours	20.6	54.8	79.8	**36.4**
Ours+Attr	**22.1**	**55.4**	**82.3**	35.0

Table 4: **Relative Captioning.** Our multimodal transformer captioning model outperforms the state-of-the-art RNN-based approach [18] on standard image captioning metrics across all datasets.

over the attribute-agnostic variant, suggesting that attribute information is complementary to the raw visual signals and improves relative captioning performance.

Dialog-based Image Retrieval. To test dialog-based retrieval performance, we paired each retrieval model with user models and ran the dialog interaction for five turns, starting from a random test image, to retrieve each target test image. Note that the user simulator and the retriever are trained independently, and can communicate only via generated captions and retrieved images. Image retrieval performance is quantified by the average ranking percentile of the target image on the test data split and the recall of the target image at top-N (R@N) in Table 3. Our transformer-based models outperform the previous RNN-based SOTA by a significant margin. In addition, the attribute-aware model produces better retrieval results overall, suggesting that the newly introduced attributes in our dataset are of benefit to the "downstream" dialog-based retrieval task. Figure 7 shows qualitative examples of our model (additional ablations and visualization are in Appendix D).

Transformers vs. RNNs. We proposed two Transformer-based models for the interactive image retrieval task, namely the Transformer-based user simulator and the Transformer-based retrieval model. In this ablative studies, we pair the Transformer-based models with the RNN-based counterpart [18] to assign the improvement credit. Table 10 summarizes the retrieval performance for different combinations in the Dresses dataset (see Appendix for results on all datasets). For the same retriever model, the improved user model always improves the retrieval performance for the first turn. As the interaction continues, other factors, including the retrieved image distribution and the simulated feedback diversity, jointly affect the retrieval performance. The improved user model achieved competitive or better scores on average. For the same user model,

Figure 7: Qualitative examples of our dialog-based image retrieval model.

	Turn 1	Turn 3	Turn 5	Average
Retriever (R) + User (R)	6.25	26.95	39.12	24.11
Retriever (R) + User (T)	7.00	29.07	41.57	25.88
Retriever (T) + User (R)	11.61	36.18	**42.40**	30.06
Retriever (T) + User (T)	**12.45**	**36.48**	41.35	**30.09**

Table 5: **Transformers vs. RNNs.** We report the performance on recall at 10 (R@10) at the 1st, 3rd and 5th turns on the dialog-based image retrieval task in Dresses. R / T indicate RNN-based and Transformer-based models.

the Transformer-based retriever model achieved overall better retrieval performance averaged over dialog turns, showing that Transformer-based models effectively aggregate the multimodal information for image retrieval.

6. Conclusions

We introduced Fashion IQ, a new dataset for research on natural language based image retrieval systems, which is situated in the detail-critical fashion domain. Fashion IQ is the first product-oriented dataset that makes available both high-quality, human-annotated relative captions, and image attributes derived from product descriptions. We showed that image attributes and natural language feedback are complementary to each other, and that combining them leads to significant improvements to interactive image retrieval systems. The natural language interface investigated in this paper overcomes the need to engineer brittle and cumbersome ontologies for every new application, and provides a more natural and expressive way for users to compose novel and complex queries, compared to structured interfaces. We believe that both the dataset and the frameworks explored in this paper will serve as important stepping stones toward building ever more effective interactive image retrieval systems in the future.

References

[1] Ziad Al-Halah and Kristen Grauman. From Paris to Berlin: Discovering Fashion Style Influences Around the World. In *CVPR*, 2020. 1

[2] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 1, 3, 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 3

[4] Daniel Barrett, Andrei Barbu, N Siddharth, and Jeffrey Mark Siskind. Saying what you're looking for: Linguistics meets video search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2016. 3

[5] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2, 3

[6] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *EMNLP*, 2018. 2

[7] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 2015. 3

[8] Y. Chen and L. Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *ECCV*, 2020. 1, 3

[9] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 1, 3

[10] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018. 3

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 3

[12] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017. 3

[13] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 3

[14] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 1, 3

[15] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, 2019. 2, 1

[16] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *arXiv preprint arXiv:1901.07973*, 2019. 3

[17] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *CVPR Workshop on Computer Vision for Fashion, Art and Design*, 2019. 3

[18] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018. 1, 2, 3, 6, 7, 8, 4

[19] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 3

[20] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 3

[21] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, 2016. 4

[22] Wei-Lin Hsiao and Kristen Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017. 3

[23] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *PCVPR*, 2018. 1, 3

[24] Wei-Lin Hsiao and Kristen Grauman. Vibe: Dressing for diverse body shapes. In *CVPR*, 2020. 3

[25] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 2, 3

[26] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018. 2, 1

[27] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *ECCV*, 2020. 3

[28] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 3

[29] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[30] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, 2013. 3

[31] A. Kovashka and K. Grauman. Discovering shades of attribute meaning with the crowd. In *ECCV Workshop on Parts and Attributes*, 2014. 3

[32] Adriana Kovashka and Kristen Grauman. Attributes for image retrieval. In *Visual Attributes*. Springer, 2017. 1, 3

[33] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 1, 3

[34] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. Cross-modal search for fashion attributes. In *KDD Workshop on Machine Learning Meets Fashion*, 2017. 2, 3

[35] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In *NeurIPS*, 2018. 2

[36] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 3

[37] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020. 6

[38] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 3, 4

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 6

[40] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, 2017. 3

[41] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015. 1, 3

[42] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011. 3

[43] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4624–4633, 2019. 2, 1

[44] Bryan Plummer, Hadi Kiapour, Shuai Zheng, and Robinson Piramuthu. Give me a hint! navigating image databases using human-in-the-loop feedback. In *WACV*, 2019. 3

[45] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 3

[46] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 1, 3

[47] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 8(5):644–655, 1998. 1

[48] Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*, 2018. 3

[49] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *ICCV*, 2013. 3

[50] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 3

[51] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, 2016. 2, 3

[52] Yaser Souri, Erfan Noury, and Ehsan Adeli. Deep relative attributes. In *ACCV*, 2016. 3

[53] Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*, 2017. 3

[54] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 6

[55] Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Steven Wu, Gerald Hui, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *NeurIPS*, 2019. 1, 3

[56] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 6

[57] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 2, 1

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 7

[59] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 3

[60] Stefanie Tellex and Deb Roy. Towards surveillance video search by natural language query. In *ACM International Conference on Image and Video Retrieval*, 2009. 3

[61] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 3

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 6, 7

[63] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015. 3

[64] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 3

[65] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019. 1, 3

[66] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381, 2018. 2

[67] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 3

[68] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. Visual search at ebay. In *KDD*, 2017. 1

[69] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 3

[70] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[71] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[72] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 3

[73] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017. 3

[74] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. *arXiv preprint arXiv:1807.01394*, 2018. 3

[75] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020. 6