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

Xiaoxiao Guo*
IBM Research AI
xiaoxiao.guo@ibm.com

Hui Wu*
IBM Research AI
wuhu@us.ibm.com

Yupeng Gao
IBM Research AI
yupeng.gao@ibm.com

Steven Rennie
Fusemachines Inc.
srennie@gmail.com

Rogerio Feris
IBM Research AI
rsferis@us.ibm.com

Abstract

We contribute a new dataset for natural language based fashion image retrieval. In contrast with existing fashion datasets, we provide both image attributes derived from real-world product descriptions, and human-generated relative captions that distinguish similar image pairs, to facilitate research on interactive image retrieval systems that jointly leverage such information. We empirically demonstrate that combining natural language feedback with visual attribute representations results in superior user feedback modeling, single-shot retrieval, and interactive retrieval performance, even when no visual attributes are available at test time. We believe that our dataset will encourage further work on developing more natural and real-world applicable conversational shopping assistants.

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 [1], creation of capsule wardrobes [18], interactive product retrieval [13], recommendation [31], and fashion design [35].

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: 1) Empowering the user with ways to express what they want, and 2) Empowering the retrieval machine with the information, capacity, and learning objective to realize high performance.

To tackle these challenges, traditional systems have relied on relevance feedback [36], 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") [26, 25] 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 the information that a user can convey. Very recent work on image retrieval has demonstrated the power of utilizing natural language to address this problem [50, 13, 43], 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 [13]. 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.

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. Nevertheless, attributes and representations extracted from this data can form a strong basis for generating stronger image captions [55, 51, 54] , and more effective image retrieval [19, 4, 40, 27]. What has been previously been unavailable is a dataset for exploring how such information interacts with and enhances the state of the art systems based on relative natural language feedback.

In this paper, we introduce a new dataset and explore methods for jointly leveraging natural language feedback and side information to realize more effective image retrieval systems (see Figure 1). The dataset, which we call Fashion Interactive Queries (*Fashion IQ*) is situated in the detail-critical fashion domain, and we investigate incorpo-

Figure 1. The Fashion IQ dataset includes attribute labels as well as relative image captions, which enables building natural language feedback based interactive image retrieval systems.

rating learned, interpretable representations based on attributes, trained on metadata that is only assumed to be available during training, into state-of-the-art relative captioning and interactive image retrieval systems.

The main contributions of this paper are as follows:

- We introduce the *first* dataset of real-world product images that is annotated with both human-generated relative captions and interpretable attribute labels extracted from real-world product descriptions, to facilitate research on leveraging side information and natural language for more effective user feedback modeling and interactive image retrieval (see Figure 3).
- We show that injecting highly interpretable, "privileged" information about image attributes during training can *substantially* improve the state-of-the-art in interactive dialog-based image retrieval at test time, even when no image attributes are available (see Figure 8).
- We empirically demonstrate that incorporating side information also leads to substantially more effective natural language based user feedback modeling and single-shot image retrieval, and benchmark new and existing architectures for both tasks (see Tables 3, 4).

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 [39, 21], fashion image generation [35], and product search [19, 56]. Both Dual Attribute-Aware Ranking Networks (DARN) [19] and Where to Buy It (WTBI) [14] datasets were created to solve the problem of retrieving images from professional fashion image catalogs, using consumer photos as queries. The ModaNet [58] and Clothing Co-Parsing (CCP) [53] datasets provide pixel-wise annotations for fashion apparel segmentation. DeepFashion [29, 12] is a large-scale fashion dataset containing consumer-commercial image pairs, and labels such as clothing attributes, landmarks, and segmentation masks. UT Zappos 50k [56] is a dataset of shoes created to model fine-grained visual differences. Amazon has several datasets [31, 48] with product images and other metadata such as consumer reviews and co-purchase information. Unlike existing fashion datasets used for image retrieval, which focus on content-based or attribute-based product search, our proposed dataset is focused on *conversational* fashion image retrieval, where user feedback is provided in natural language. In a similar vein, the Multi-Modal Domain-Aware Conversations dataset [37] uses synthetic data for user feedback, while our work makes available a unique set of *human-written* relative descriptions for a large set of product images.

Attributes for Interactive Fashion Search. Visual attributes, including color, shape, and texture, have been successfully used to model clothing images [19, 17, 18, 1, 57, 5, 30]. Relative attributes [32, 41] have been exploited as a richer form of feedback for interactive fashion image retrieval [25, 26, 23, 24]. In [57], 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. All these methods rely on a fixed, pre-defined set of attributes, whereas our work explores the use of feedback as relative queries in *natural language*, allowing more flexible and more precise descriptions of the items to be searched.

Image Retrieval with Natural Language Queries. Methods that lie in the intersection of computer vision and natural language processing, including image captioning [34, 49, 52] and visual question-answering [2, 6, 45], have received much attention from the research community. Recently, several techniques have been proposed for image or video retrieval based on natural language queries [28, 3, 46]. In [50], both image and text are used as queries for retrieval, where the text specifies a desired modification to the image. [43] decomposes a complex scene into local regions and utilized local region descriptions as iterative queries. In another line of work, visually-grounded dialog systems [7, 42, 9, 8] 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 [13], we consider the setting of goal-driven dialog, where the user provides feedback in natural language, and the agent outputs retrieved images. Unlike [13], we provide a large dataset of relative captions anchored in a dataset with real-world contextual information, which is available to the community. In addition, we show that the use of side information can improve the per-

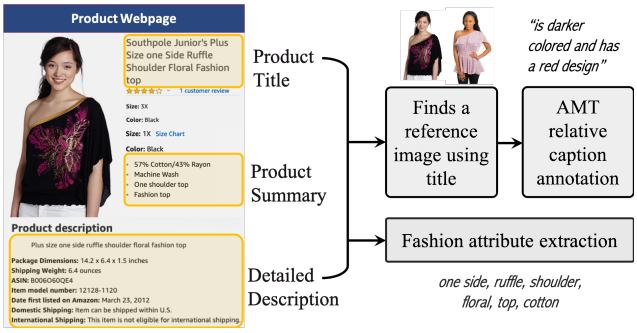

Figure 2. Overview of the Fashion IQ dataset collection process.

formance of image retrieval using natural language by improving: (1) single-turn retrieval, (2) user modeling, and (3) dialog-based 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 [47], with many applications in computer vision [38, 19]. In the context of fashion, [19] 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 [4], learn weakly supervised clothing representations [40], and improve search based on noisy and incomplete product descriptions [27]. In our work, for the first time, we explore the use of side information for image retrieval with a natural language feedback interface.

3. Fashion IQ Dataset

3.1. Dataset Collection

The images of fashion products that comprise our Fashion IQ dataset were originally sourced from *Amazon.com*. Similar to [1], we selected three categories of product items from the original Amazon Review data [31, 16], specifically: Dresses, Tops&Tees, and Shirts. For each image, we crawled *Amazon.com* and extracted corresponding product information, when available. To facilitate research on the benefits of using natural language for interactive image retrieval, we additionally collected natural language based user feedback, describing the differences between each target product image and a single reference product image. Note that these human-written relative descriptions are associated with real-world context, including side information derived from product descriptions and customer reviews. This unique feature of the Fashion IQ dataset allows researchers to investigate the advantages of natural language feedback in conjunction with such contextual information, which is often available in practice. The overall data collection procedure is illustrated in Figure 2. Basic statistics of the resulting Fashion IQ dataset are summarized in Table 1.

In the following subsections, we further describe how we collected the fashion attribute labels and relative captions.

Collecting attribute labels While the Amazon Review data contains product metadata information on titles and categories, this information tend to be short, generic and incomprehensive (c.f. Figure 5), and does not correlate well with the visual appearance of fashion images. Instead, we leveraged the rich textual information contained in the product website, and 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 [29], 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 [29] 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 6, we provide examples of the original side information provided in Amazon Review dataset and the corresponding attribute labels that were extracted.

Collecting Relative Captions The goal in supporting relative captions is to allow users to use natural language expressions to more flexibly describe how a reference image (e.g. a current search result) differs from an image of what they are searching for, to realize more interactive and effective image retrieval. To amass relative captions for the Fashion IQ data, we adopted a data collection interface similar to the one presented in [13], and collected data using Amazon Mechanical Turk. Briefly, the users were situated in a 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.[1] To ensure that the relative captions described 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 data split subset) with the maximum sum of the TF-IDF weights on each overlapping word. We randomly selected ∼10,000 target images for each of the three fashion categories, and collected two sets of captions for each pair. Inconsistent

[1]For brevity, we refer the readers to Appendix A of [13] for further details of the data collection interface.

	Dresses		Tops&Tees		Shirts	
	train / val / test	total	train / val / test	total	train / val / test	total
# Images	11452 / 3817 / 3818	19087	16121 / 5374 / 5374	26869	19036 / 6346 / 6346	31728
# Images with side info	7741 / 2561 / 2653	12955	9925 / 3303 / 3210	16438	12062 / 4014 / 3995	20071
# Relative Captions	11970 / 4034 / 4048	20052	12054 / 3924 / 4112	20090	11976 / 4076 / 4078	20130

Table 1. Dataset statistics on Fashion IQ.

Semantics	Quantity	Examples
Direct reference of target image	49%	is solid white and buttons up with front pockets
Comparative reference	32%	**has longer sleeves** and is **lighter in color**
Mixed use of direct and comparative references	19%	has a geometric print with **longer sleeves**
Single-attribute phrase	30.5%	**is more bold**
Compositional attribute phrases	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 between the target and the reference images.

captions were filtered. Figure 6 shows examples of image pairs presented to the user, and the resulting relative image captions that were collected. To ensure the quality of the collected captions, we only included workers from three dominant English-speaker countries, with master level of expertise and with an acceptance rate above 95%.

4. Fashion IQ Analysis and Applications

4.1. Dataset Analysis

Figure 4 depicts empirical distributions of relative caption length and number of attributes per image for all subsets of Fashion IQ, and shows that all three datasets are similarly distributed. Figure 6 depicts examples of collected relative captions in the Fashion IQ dataset, and Figure 7 displays word-frequency clouds of the relative captions in each fashion category. The natural language based data annotation process results in relatively rich fashion vocabularies for each subtask, with prominent visual differences often being implicitly agreed upon by both annotators, and resulting in semantically related descriptions. In most cases, the attribute labels and relative captions contain complimentary information, and thus jointly form a stronger basis for ascertaining the relationships between images.

To further obtain insight on the linguistic properties of the relative captions, we conducted a semantic analysis on a subset of 200 randomly chosen relative captions. The results of the analysis are summarized in Table 2, and suggest that there is an approximately even chance that the user will ignore the reference image entirely and directly describe the target image, rather than utilizing the reference image for comparative purposes. Further, the majority (69.5%) of the captions contain rich information about the target image and consist of composite attribute phrases. The diversity in the structure and content of the captions provide a fertile resource for modeling user feedback and for learning natural language feedback based image retrieval models.

4.2. Using Fashion IQ

The relative captions and the attribute labels in the Fashion IQ dataset can be used in a multitude of ways to drive progress on developing more effective interfaces for image retrieval (as shown in Figure 3). These tasks can be developed as standalone components of the final retrieval application, or can be investigated in conjunction. For example, a trained user model can serve as a substitute for real users and bootstrap or augment training of the interactive retrieval model. Next, we briefly introduce the component tasks associated with developing interactive image retrieval applications, and discuss how Fashion IQ can be utilized to realize and enhance these components.

Modeling user feedback User models are a common way to bootstrap and augment the training of dialog systems [10]. When appropriately constructed, user models can mimic the behavior of a real user, and provide a large amount of low-cost training data for dialog models. Simulated user feedback has also been recently utilized in previous work on image retrieval [50, 43], and for directly modeling and generating text descriptions describing the differences between pairs of images [13]. More broadly, several relative captioning systems for describing visual differences have been proposed for several tasks [20, 44, 33, 11]. The work in [13] has established the efficacy of relative captioning and the utility of relative captioning simulators in developing dialog-based interactive image retrieval systems. Nevertheless, high fidelity user simulation is inherently very difficult and unsolved, and remains a major focus of cur-

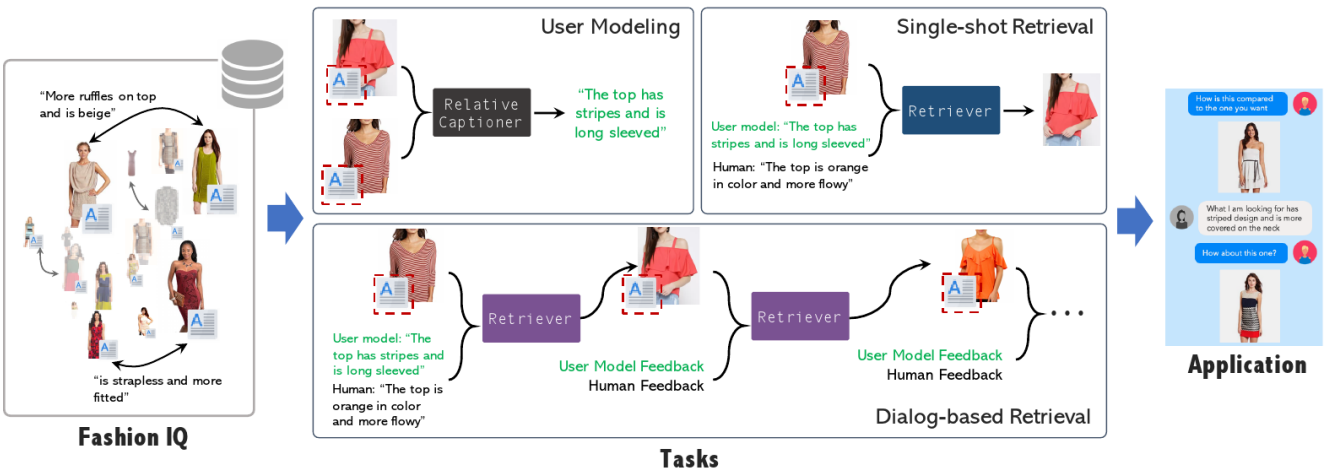

Figure 3. Fashion IQ can be used in different scenarios to enhance the development of an interactive fashion retrieval system with natural language interaction. We provide three example scenarios: user modeling and two types of retrieval tasks. Fashion IQ uniquely provides both annotated user feedback (black font) and visual attributes derived from real-world product data (dashed boxes) for system training.

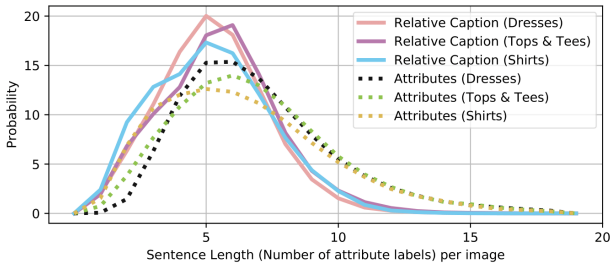

Figure 4. Distribution of sentence lengths and number of attribute labels per image.

rent research efforts. Fashion IQ introduces the opportunity to utilize both attribute labels and human-annotated relative captions to realize stronger user simulators, and correspondingly stronger interactive image retrieval systems.

Composing text and image data for image retrieval Single-turn image retrieval systems have now evolved to support multimodal queries that include both images and text. Recent work, for example, has attempted to use attribute labels to manipulate retrieval queries [57], and to construct synthetic natural phrases from image labels [50]. Fashion IQ enables new research on multimodal retrieval systems by virtue of human-annotated relative feedback sentences, as well as interpretable attribute labels associated with image metadata.

Dialog-based interactive image retrieval Recently, relative captions for a shoe dataset were used in [13] as a basis for training the first dialog-based interactive image retrieval systems. The additional attribute labels in Fashion IQ, as well as a more diverse set of fashion categories, and more data (∼6x larger), allows for more comprehensive research on interactive product retrieval systems. In the next section, we provide baselines that demonstrate how attribute labels can be incorporated into this general framework, and show that this significantly advances the state of the art in dialog-based interactive image retrieval.

5. Baseline Experiments

In this section, we provide experiments on applying Fashion IQ to build different sub-systems for interactive image retrieval. We conducted experiments on using images and relative captions as the lone source of training data, and for the first time, investigated the value of using both attribute labels and relative captions in the context of interactive image retrieval with a natural language interface. Specifically, in the rest of this section, we provide empirical studies on *user modeling* (Section 5.1), *composing text and images for image retrieval* (Section 5.2) and *dialog-based interactive image retrieval* (Section 5.3).

Experiment setup All experiments were performed on the three categories (Dresses, Shirts and Tops&Tees), with the same data split shown in Table 1. In all models, we study different combinations of training data modalities, including images, relative captions, and attribute information. To reduce noise and avoid any dependence on attribute labels at test time, we extracted attribute features from the penultimate layer of a trained *attribute prediction network* (AttrNet)[2], and used them in all models which involve attribute

[2]The architecture and the performance of AttrNet are included in Appendix A.

Figure 5. Examples of the original product titles and the derived attributes (more examples available in the appendix, Figure 13).

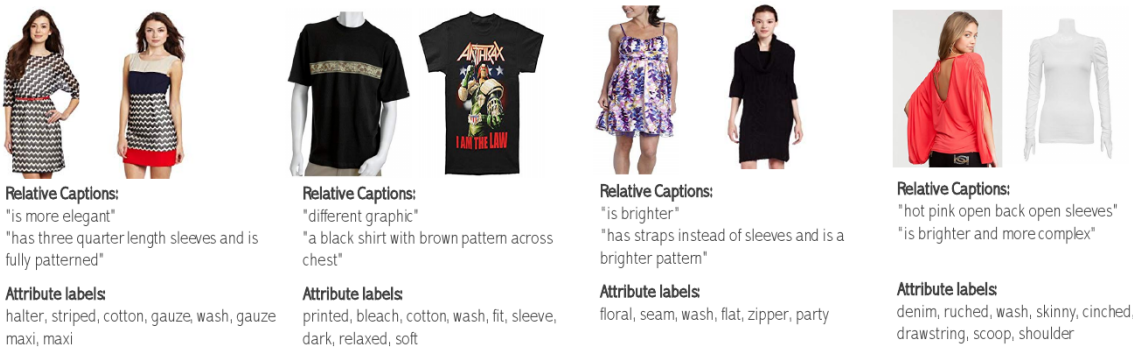

Figure 6. Examples of relative captions and image attributes in the dataset. Attributes are associated with the left image in each pair.

Figure 7. Vocabulary of relative captions scaled by frequency.

information. The parameters for each experiment were selected based on retrieval performance on the validation sets. The Adam optimizer was used to train all networks. For a detailed explanation of all network configurations and parameter settings, please consult Appendix C.

5.1. User Modeling using Relative Captioning

The role of a user model is to act as a surrogate for real human users, and to provide text-based feedback describing the difference between the target and candidate images. Our baseline user model is a relative image captioning model that employs the show-and-tell [49] architecture, with the input being the difference between two image features (#1, Figure 9). Since item attributes are an elemental part of many of the phrases people use to search for items, they naturally share similar semantics with and can enhance the quality of the relative feedback simulator. A simple way to incorporate attribute information is to augment the image representation (#2, Figure 9). Specifically, we incorporate attribute features into the relative captioner by first linearly projecting each set of predicted attribute features to match the dimension of the hidden state of the decoder RNN, and then concatenating them with the image features. The difference between the resulting target and candidate features is then input into the initial state of the RNN. We also investigated an attention-enabled model variation (#3, Figure 9): hidden state vectors are used as query vectors to generate attention weights over the combined image-attribute features, and the new feature after applying additive attention [52] is then concatenated with the embedding of the previous word and input into the decoder RNN at the next timestep.

Results The performance of each method is summarized in Table 4. The attribute-aware method outperform the image-only baseline across all metrics, suggesting that attribute information is complementary to the raw visual signals and improves relative captioning performance. The attention-enabled attribute-aware captioner, moreover, scores significantly higher than the concatenation-based model, suggesting that the attention mechanism is better able to utilize the attribute prediction information. To assess the relative importance of the image and attribute components, we investigated removing each of the components from the inputs of attribute-aware attention model. The performance degradation from removing the image component

		R@10 (R@50)		
		Dresses	Shirts	Tops&Tees
Side information features				
A	Full model: side information, gating on text features.	11.24 (**32.39**)	**13.73** (**37.03**)	**13.52** (**34.73**)
B	A without side information features.	**11.49** (29.99)	13.68 (35.61)	11.36 (30.67)
C	Image and text concatenation, linear projection [13].	10.52 (28.98)	13.44 (34.60)	11.36 (30.42)
Variants of gating connection				
D	A without gating connection.	10.42 (27.99)	12.33 (33.94)	11.48 (30.35)
E	Gating on image features (with image feature embedding).	9.73 (25.64)	11.62 (30.75)	10.09 (27.21)
F	TIRG [50] (E without image feature embedding).	8.10 (23.27)	11.06 (28.08)	7.71 (23.44)
Single-modality retrieval				
G	Relative feedback only.	6.94 (23.00)	9.24 (27.54)	10.02 (26.46)
H	Image feature only.	4.20 (13.29)	4.51 (14.47)	4.13 (14.30)
I	Side information feature only.	2.57 (11.02)	4.66 (14.96)	4.77 (13.76)

Table 3. Results on composing text and image features for image retrieval.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	Rouge-L	CIDEr	SPICE
#3 (D)	**61.3**	**44.1**	**29.0**	**19.7**	**26.2**	**55.5**	**59.4**	**34.7**
#3 - Attribute(D)	60.3	43.5	28.8	19.0	25.9	54.7	58.5	33.8
#3 - Image (D)	56.6	39.9	24.0	14.5	22.8	51.2	32.8	29.4
#2 (D)	58.5	42.0	26.7	17.5	24.0	53.2	42.7	30.8
#1 (D)	58.1	41.0	26.3	17.4	24.8	53.6	48.9	32.1
#3 (S)	**57.7**	**46.3**	**32.9**	**22.3**	**27.9**	**57.1**	**78.8**	**36.6**
#3 - Attribute(S)	56.0	45.4	31.6	19.8	25.9	54.7	58.5	33.8
#3 - Image(S)	48.2	39.1	24.8	16.1	23.8	51.7	50.3	31.7
#2 (S)	54.5	42.6	29.1	19.4	25.8	53.5	47.1	31.8
#1 (S)	53.2	41.9	29.0	19.6	25.9	53.8	52.6	32.0
#3 (T)	**58.4**	**44.1**	**29.6**	**20.3**	**26.5**	**54.1**	**63.3**	**35.3**
#3 - Attribute(T)	57.3	43.2	28.9	19.7	26.2	53.6	62.4	34.7
#3 - Image(T)	44.1	35.5	21.4	13.1	21.9	50.1	33.7	29.6
#2 (T)	55.9	41.0	26.0	17.0	25.4	51.5	40.7	31.1
#1 (T)	54.0	39.4	24.6	15.7	24.3	50.5	41.1	30.6

Table 4. Comparison on image-only (#1), attribute-aware (#2), and attribute-aware attentional (#3) user simulator models on common image captioning metrics. D / S / T indicate Dresses / Shirts / Tops&Tees. The highest scores per dataset are highlighted. "- Image" means the image component is removed, and "- Attribute" shows the minimum performance when removing one of the five attribute types.

(- Image) is more significant than removing any attribute component, indicating that the image component still plays the most prominent role in the relative captioning systems.

5.2. Composing Text and Image Data for Retrieval

Given a reference image and a feedback sentence, we can retrieve the target image by composing the two[3]. In this section, we provide empirical studies comparing different combinations of query modalities for retrieval, including relative feedback, image features, and learned attribute features. Specifically, the images were encoded using a pre-trained ResNet-101 network; the relative feedback sentences were encoded using Gated Recurrent Networks with one hidden layer. We used pairwise ranking loss [22] for all methods with the best margin parameters for each method selected using the retrieval score on the validation set. We included two recent methods for composing text and visual features for image retrieval as the baseline models: (1) the response encoder network in [13], which is based on concatenation of the image feature (after linear embedding) with the encoded textual features; and TIRG [50], which is a recent method based on concatenation of visual and textual features with an additional gating connection to pass the image features directly to the learned joint feature space.

Result Analysis We reported the retrieval results on the test set in Table 3. We found that the best performance was

[3]The retrieval experiments use a subset of the original dataset which have relative caption annotations. The two relative caption annotations associated with each image are considered as belong to two different queries.

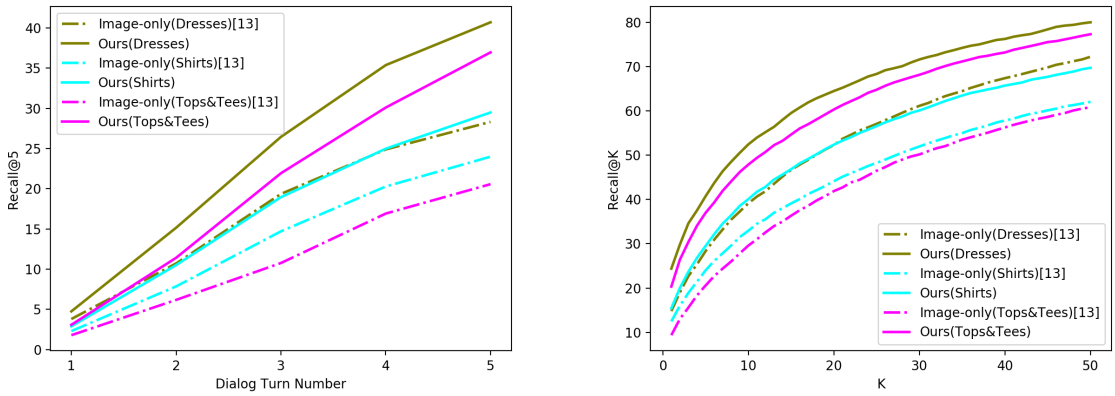

Figure 8. Recall@5 curves of different methods as a function of dialog turns (left) and recall curves at the 5-th dialog turn (right).

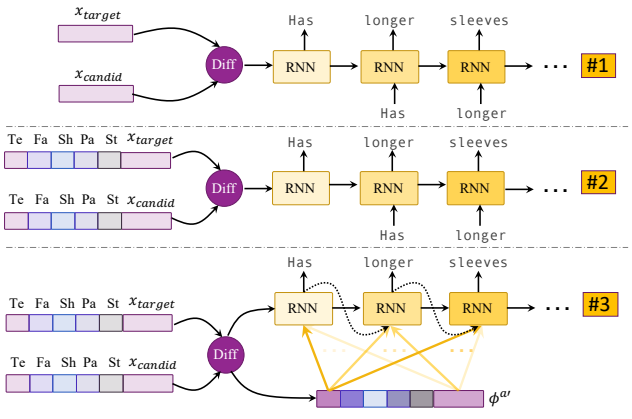

Figure 9. User models: image-only model (#1), attribute-aware model (#2), & attribute-aware attention model (#3).

achieved by using all three modalities and applying a gating connection on the encoded natural language feedback (Model A). Removing the gating connection (D), or using only image-based features and the gating connection (E and F), both produced inferior results. This confirmed the informative nature of relative feedback for image retrieval. Similar observations can be made in the cases of single-modality studies, where the relative feedback modality significantly outperformed both of the image based features. Finally, removing side information features (B and C) reduced the performance of the full model, demonstrating the benefit of incorporating attribute labels and concurring with our observation in user modeling experiments.

5.3. Interactive Image Retrieval

To test the use of Fashion IQ in the interactive retrieval setting, we consider a recent framework [13] as well as an evolved version of this framework to incorporate predicted attribute features[4]. The original framework [13] consists of three modules (namely, response encoder, state tracker, and candidate generator). The predicted attribute features are incorporated to augment image representation (similar to how they were used in user models), and we make changes in the *response encoder* module to utilize the augmented features. Specifically, we combine the image features and attribute features via additive attention mechanism [52], with GRU-encoded feedback text as the query vector. To test the retrieval performance, we paired each retrieval model with user models and ran the dialog interaction for five turns. The attribute-aware user simulator is used for all experiments.

Results Image retrieval performance is quantified by the recall of the target image at top-K (R@K) in Figure 8. Our attribute-aware method outperforms the image-only baseline, demonstrating the benefit of leveraging side information and relative feedback jointly for interactive image retrieval.[5] Figure 8 (left) shows that for recall@5, the performance gain from using side information consistently increases as function of the number of turns, as information gets aggregated. Figure 8 (right) depicts the recall@K as a function of K at the fifth dialog turn. Our results again significantly outperform the image-only baseline [13] $\forall$ K.

6. Conclusions

We introduced Fashion IQ, a new dataset for research on multimodal retrieval systems that incorporate natural language feedback, which is situated in the detail-critical fashion domain. In the paper, we discussed how Fashion IQ can be utilized to train the components of interactive image retrieval systems, and demonstrated that image attributes derived from side information can significantly improve performance of user feedback, image retrieval, and interactive dialog-based image retrieval systems, even when this information is privileged, and not available at test time. Fashion IQ is the first product-oriented dataset that makes available both human-annotated relative captions, and image attributes derived from product descriptions. We believe that

[4]Full details about the overall framework is in Appendix B.

[5]Additional retrieval results are in Appendix D.

both the dataset and the frameworks explored in this paper will serve as important stepping stones toward building ever more natural and effective interactive image retrieval systems in the future.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 1, 2, 3

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2

[3] D. Barrett, A. Barbu, N. Siddharth, and J. M. Siskind. Saying what you're looking for: Linguistics meets video search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10), 2016. 2

[4] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 1, 3

[5] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 2015. 2

[6] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. In *CVPR*, 2018. 2

[7] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *CVPR*, 2017. 2

[8] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017. 2

[9] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 2

[10] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *ACL*, 2017. 4

[11] M. Forbes, C. Kaeser-Chen, P. Sharma, and S. Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, 2019. 4

[12] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *arXiv preprint arXiv:1901.07973*, 2019. 2

[13] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018. 1, 2, 3, 4, 5, 7, 8, 11

[14] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 11

[16] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, 2016. 3

[17] W.-L. Hsiao and K. Grauman. Learning the latent look: Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017. 2

[18] W.-L. Hsiao and K. Grauman. Creating capsule wardrobes from fashion images. In *PCVPR*, 2018. 1, 2

[19] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 1, 2, 3

[20] H. Jhamtani and T. 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. 4

[21] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 7

[23] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, 2013. 2

[24] A. Kovashka and K. Grauman. Discovering shades of attribute meaning with the crowd. In *ECCV Workshop on Parts and Attributes*, 2014. 2

[25] A. Kovashka and K. Grauman. Attributes for image retrieval. In *Visual Attributes*. Springer, 2017. 1, 2

[26] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 1, 2

[27] K. Laenen, S. Zoghbi, and M.-F. Moens. Cross-modal search for fashion attributes. In *KDD Workshop on Machine Learning Meets Fashion*, 2017. 1, 3

[28] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *CVPR*, 2017. 2

[29] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 3

[30] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, 2017. 2

[31] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015. 1, 2, 3

[32] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 2

[33] D. H. Park, T. Darrell, and A. Rohrbach. Robust change captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4624–4633, 2019. 4

[34] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 2

[35] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 1, 2

[36] Y. Rui, T. S. Huang, M. Ortega, and S. 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

[37] A. Saha, M. M. Khapra, and K. Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*, 2018. 2

[38] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *ICCV*, 2013. 3

[39] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 2

[40] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, 2016. 1, 3

[41] Y. Souri, E. Noury, and E. Adeli. Deep relative attributes. In *ACCV*, 2016. 2

[42] F. Strub, H. de Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*, 2017. 2

[43] F. Tan, P. Cascante-Bonilla, X. Guo, S. Wu, G. Hui, S. Feng, and V. Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *NeurIPS*, 2019. 1, 2, 4

[44] H. Tan, F. Dernoncourt, Z. Lin, T. Bui, and M. Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 4

[45] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 2

[46] S. Tellex and D. Roy. Towards surveillance video search by natural language query. In *ACM International Conference on Image and Video Retrieval*, 2009. 2

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

[48] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015. 2

[49] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2, 6, 11

[50] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019. 1, 2, 4, 5, 7

[51] Q. Wu, C. Shen, P. Wang, A. Dick, and A. 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. 1

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2, 6, 8

[53] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 2

[54] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[55] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 1

[56] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 2

[57] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, 2017. 2, 5

[58] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. *arXiv preprint arXiv:1807.01394*, 2018. 2

Appendix

A. Attribute Prediction Network

The process of crawling product information for attributes to associate with individual product images, while automated, can lead to noisy and incomplete attribute features. To alleviate this issue, we introduce an attribute prediction network to infer estimated attributes, which are then used by both the user simulator and the interactive retriever. For each image x in the retrieval database, the AttrNet predicts a set of attribute features $\{\phi^a(x) \in \mathbb{R}^{D_a}\}$, where $a \in \{\text{texture}, \text{fabric}, \text{shape}, \text{part}, \text{style}\}$ is an attribute type indicator, and D_a is the number of attributes within the corresponding attribute type. Specifically, the attribute prediction model is a multi-column neural network with shared lower layers, which takes the image as input, and outputs the attribute tags, as shown in Figure 10. The shared lower layers consists of a pre-trained ResNet-152 network [15] up to the penultimate layer, where the last fully connected layer is replaced by a trainable linear projection, followed by ReLU. We use x to represent both the image and its vector representation $x \in \mathbb{R}^{D_x}$ for notational simplicity. The projected image embedding x is then passed to two independent linear layers with ReLU applied to the hidden layer. The final outputs are rectified by the sigmoid function to generate the attribute features $\phi^a(x)$. The performance of the trained model is shown in Table 5.

B. Dialog-based Interactive Image Retrieval

The role of the dialog manager is to select the best candidate image x_{t+1} from the database, based on the dialog history, H_t, as shown in Figure 11. Following [13], the dialog manager model consists of three main components: a *response encoder*, a *state tracker* and a *candidate generator*. Next, we introduce the design of each of the three components, highlighting the differences between [13] and our framework, which utilizes *attribute-aware* visual representations.

Response Encoder At the t-th dialog turn, the *response encoder* embeds the candidate image x_t, the candidate image's attribute features $\{\phi^a(x_t)\}$ and the corresponding user feedback o_t into a joint visual semantic representation, $e_t = \mathcal{R}(x_t, \{\phi^a(x_t)\}, o_t) \in \mathbb{R}^{D_e}$. First, the feedback (i.e., a sequence of word indices) o_t is encoded by an LSTM into a vector $e_t^o \in \mathbb{R}^{D_e}$. Then, we consider two ways of combining the image feature and the attribute features to obtain the attribute-aware visual representation, e_t^{x+}. The first approach is based on direct feature concatenation, followed by a linear projection to obtain a vector of length D_e. Alternatively, we adopted the additive attention mechanism [49], where a joint visual representation is obtained by the weighted sum of the image feature and each of the attribute features. The attention weights were computed using the scoring function that takes as input the sum of the projected visual feature and the feedback representation. Finally, given the attribute-aware visual representation, and the feedback representation, e_t^o, the joint visual semantic representation is computed as: $e_t = \sigma(e_t^{x+} + e_t^o)$, where σ is a ReLU layer.

	Dresses		Shirts		Tops&Tees	
	top-3	top-5	top-3	top-5	top-3	top-5
Texture	0.50	0.60	0.69	0.78	0.54	0.65
Fabric	0.45	0.53	0.70	0.76	0.52	0.58
Shape	0.36	0.47	0.69	0.78	0.51	0.61
Part	0.31	0.44	0.51	0.66	0.37	0.49
Style	0.19	0.28	0.26	0.36	0.21	0.28
All	0.36	0.46	0.57	0.66	0.43	0.51

Table 5. Attribute prediction results on top-3 and top-5 recall scores for the five attribute types.

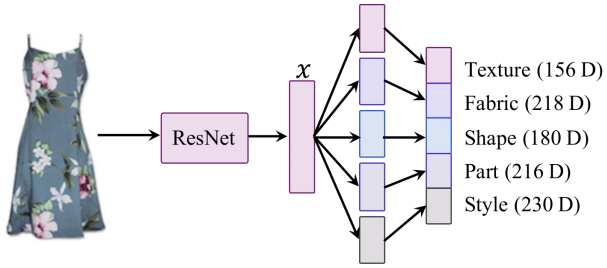

Figure 10. Attribute prediction network.

State Tracker The *state tracker* follows a similar design as in [13], which aggregates the encoded response representation with the dialog history from previous turns, producing a query vector $q_t \in \mathbb{R}^{D_q}$. Specifically, the state tracker is based on a Gated Recurrent Unit (GRU). The forward dynamics of the state tracker are: $h_t = \text{GRU}(e_t, h_{t-1}), q_t = \mathbf{W}^q h_t$, where $h_t \in \mathbb{R}^{D_h}$ and $\mathbf{W}^q \in \mathbb{R}^{D_q \times D_h}$ is a trainable matrix.

Candidate Generator The *candidate generator* searches for a new candidate image, given the aggregated query vector q_t. We represent each candidate image in the retrieval database using the concatenation based attribute-aware visual representation, i.e., $d(x) = \mathbf{W}^q[x, \{\phi^a(x)\}] \in \mathbb{R}^{D_h}$. We then used the $L2$ distance between each database feature $d(x)$ and the query vector q_t to select the candidate image. Given the trainable parameters of the three components, the *response encoder*, *state tracker* and *candidate generator*, we optimized the entire network end-to-end, using the same policy learning procedure as proposed in [13].

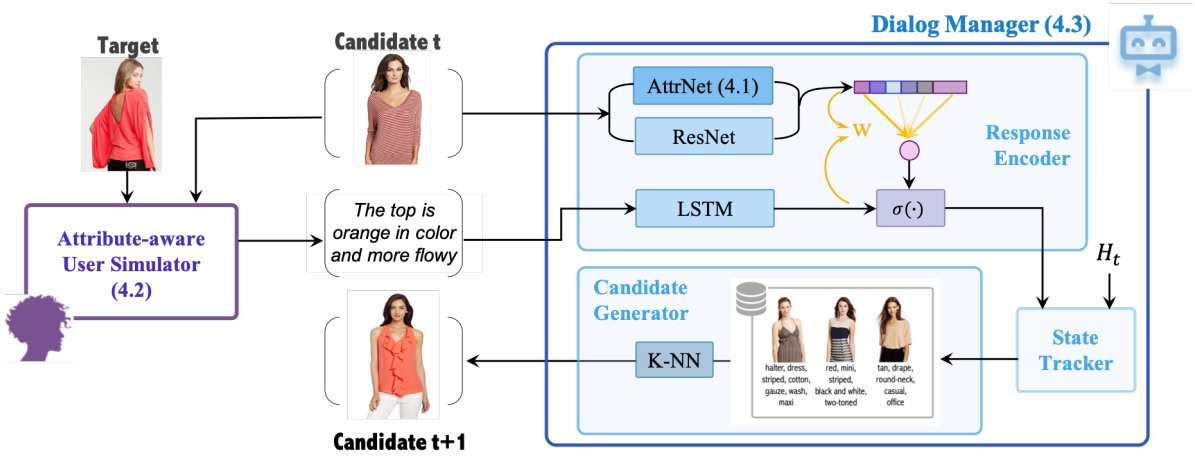

Figure 11. Framework of our attribute-aware dialog-based interactive image retrieval system.

	Dialog Turn 1				Dialog Turn 3				Dialog Turn 5			
	P	R@5	R@10	R@50	P	R@5	R@10	R@50	P	R@5	R@10	R@50
Attribute-aware (D)	**90.52**	**4.74**	**7.73**	23.94	**98.09**	26.45	**36.19**	**67.72**	98.92	40.71	**52.43**	79.91
with Attention (S)	**90.87**	**2.88**	**4.96**	**17.32**	**98.02**	**18.95**	**27.33**	**55.49**	**98.87**	**29.49**	**40.07**	**69.71**
(T)	**90.37**	3.07	5.16	17.27	**98.04**	**21.93**	**30.18**	59.06	99.03	**36.97**	**47.87**	**77.30**
Attribute-aware (D)	90.39	4.52	7.48	**24.14**	98.00	**26.65**	36.05	65.60	**98.95**	**40.88**	52.37	**79.99**
via Concatenation (S)	89.93	2.41	4.09	14.86	97.55	16.15	23.63	50.60	98.55	27.21	36.44	65.25
(T)	90.34	**3.22**	**5.39**	**17.75**	98.03	20.78	29.02	**59.57**	**99.07**	35.37	46.41	76.58
Image-Only (D)	89.45	3.79	6.25	20.26	97.49	19.36	26.95	57.78	98.56	28.32	39.12	72.21
(S)	89.39	2.29	3.86	13.95	97.40	14.70	21.78	47.92	98.48	23.99	32.94	62.03
(T)	87.89	1.78	3.03	12.34	96.82	10.76	17.30	42.87	98.30	20.57	29.59	60.82

Table 6. Dialog-based interactive image retrieval performance on ranking percentile (P) and recall at N (R@N) at the 1st, 3rd and 5th dialog turns. D / S / T indicate the Dresses / Shirts / Tops&Tees datasets. The highest scores per dataset are highlighted.

C. Experimental Settings

Attribute Prediction Network The image embedding size (D_x) is 1024 in the attribute prediction network. For each attribute-specific column, the penultimate layer size for each attribute group is twice the number of attribute labels for that group (i.e., $2 \times D_a$). In training, we used binary cross entropy loss and Adam with an initial learning rate of 0.001.

Attribute-aware User Simulator The word embedding dimension and the decoder LSTM configuration are the same for all methods. Specifically, the word embedding size is 512-D, the decoder LSTM hidden state is 512-D and the input dimension is 1024-D. For the image-only and the attribute-aware concatenation captioning models, the image embedding is 1024-D. The attribute-aware concatenation captioning model linearly projects the concatenated attribute and image features to 1024-D. For attribute-aware attention captioning model, the image embedding is 512-D, and the projected attribute vectors are also 512-D. After concatenated with the word embedding, the input to the decoder LSTM is thus 1024-D, which is consistent with the other two models.

Image Retrieval Experiments For composing text and image features for retrieval, the network embedding is 1024-D, which we found performed well for all methods. All dialog-based interactive image retrieval methods share the same model configuration. The response encoding (D_e) is 512-D. The state tracker GUR hidden state (D_h) is 256-D. The query embedding (D_q) is 512-D.

D. Additional Results on Interactive Image Retrieval

Figure 12 shows examples of the attribute-aware user simulator interacting with the dialog manager. In all examples, the target images reached final rankings within top 50 after 5 dialog turns. The target images ranked incrementally higher during the dialog and the candidate images were more visually similar to the target images. These examples show that the dialog manager is able to refine the candidate selection given the user feedback, exhibiting promising behavior across different clothing categories.

The image retrieval performance is quantified by the av-

Figure 12. Examples of the simulator interacting with the dialog manager system. The right-most column shows the target images.

erage ranking percentile of the target image on the test data set and the recall of the target image at top-N (R@N) in Table 6. Our attribute-aware method outperforms the image-only baseline. Additionally, the attention-enabled model produced better retrieval results overall, suggesting that more advanced techniques for composing side information, relative feedback and image features could lead to further performance gains.

pattern, clean, jacquard, waffle, Shirt, Button, classic

T: Nat Nast Men's Bar Code Classic Button Down Shirt

printed, stripe, cotton, fit, contrast, pocket, snap, summer, sun

T: Volcom Men's Avenida Tank Top

graphic, printed, cotton, wash, fit, sleeve, art, love workout

T: The Mountain Men's Polar Collage T-Shirt

leaf print, dye, wash, woven, shirt, button, sleeve

T: Cubavera Men's Short Sleeve Yarn Dye Printed Shirt

diamond, graphic, cotton, wash, Box, Hem, Classic, logo

T: Diamond Supply Co. Men's Diamond Forever Tee

stone, wash, classic fit, fit, shirt, button, long sleeve, pocket, solid

T: Volcom Men's X Factor Solid Long Sleeve Shirt

printed, stripes, knit, waffle, fit, long sleeve, sleeve, basic, thermal

T: Volcom Men's Nutto Long Sleeve Thermal T-Shirt

cotton, plaid, wash, woven, Shirt, collar, pocket, sleeve, logo

T: IZOD Men's Double Pocket Madras Woven Shirt

floral, print, clean, ruffle, button, v-neck, new york

T: Jones New York Women's Ruffle Blouse

cotton, ribbed, wash, classic, everyday, heat, love, relaxed, soft

T: Jockey Women's T-Shirts Classic Tank Top

clean, lace, loose, sheer, button, pocket, sleeveless, flirty

T: 2B Anna Button Down Lace Tank

pattern, print, pleated, ruffle, wash, fit, medium, Sleeveless, flirty

T: Anna-K S/M Fit Salmon Asian-Inspired Chains Pleated Ruffle Ribbon Blouse

graphic, printed, cotton, fair, fit, art, party, soft, youth

T: Womens Under New Management Funny Wedding Party Shirts Bachelor Novelty T shirt Blue

bejeweled, chiffon cotton, loose, studded, button, collar, cuffed, long sleeve, light

T: G2 Chic Women's Bejeweled Collar Studded Front Hi Lo Chiffon Shirt

knit, ruched, keyhole, scoop, sleeve, twisted, please

T: PattyBoutik Women's Twisted Cross Keyhole 3/4 Sleeve Knit Top

Leopard, wash, sleeveless, classic

T: Chaus Women's Sleeveless Classic Leopard Blouse

stripe, bodycon, fit, neckline, sleeveless, chic, running, shopping, summer,

T: G2 Chic Women's Short Sleeve Striped Bodycon Dress with V-Neckline

printed, ruffled, a-line, strapless, maxi, beach, party, retro, summer,

T: KOH KOH Womens Long Sexy Strapless Tube Printed Evening A-Line Gown Maxi Dress

dye, ruffle, wash, maxi, neckline, strapless,

T: Southpole Juniors Strapless Tye Dye Ruffle Accent Neckline Maxi Dress

striped, lace, mesh, bodycon, trench

T: bebe Contour Mesh Detail Dress

Figure 13. Examples of the original product title descriptions (T) and the collected attribute labels (on the right of each image).