Appendix

A. Additional Information on Fashion IQ

Our dataset is publicly available and free for academic use. Figure 8 depicts the empirical distributions of relative caption length and number of attributes per image for all subsets of Fashion IQ. We visualize in Figure 9 the word-frequency clouds of the relative captions in each fashion category. In Figure 10, we show more examples of the original product titles and the derived attributes. In Table 6, we show the detailed statistics for the Fashion IQ dataset.

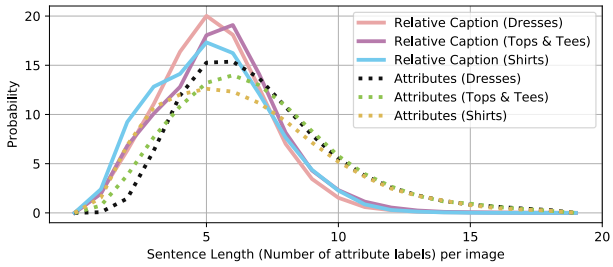

Figure 8: Distribution of sentence lengths and number of attribute labels per image.

Figure 9: Vocabulary of relative captions scaled by frequency

Attribute Prediction. The raw attribute labels extracted from the product websites may be noisy or incomplete, therefore, to address this, we utilize the DeepFashion attributes to complete and de-noise the attribute labels in Fashion IQ. Specifically, we first train an Attribute Prediction Network, based on the EfficientNet-b7 architecture,[5] to predict the DeepFashion attributes, using the multi-label binary cross-entropy loss. After training on DeepFashion labels, we fine-tune the last layer on each of our Fashion IQ categories (namely Dresses, Shirts, and Tops & Tees) with the same loss function. The fine-tuning step adjusts the attribute prediction to our categories' attribute distribution. We then use the attribute network to predict the top attribute labels based on their output values. All images have the same number of attribute labels (that is, 8 attributes per image).

[5]https://github.com/lukemelas/EfficientNet-PyTorch

	#Image	# With Attr	# Relative Cap.
Dresses			
Train	11,452	7,741	11,970
Val	3,817	2,561	4,034
Test	3,818	2,653	4,048
Total	19,087	12,955	20,052
Shirts			
Train	19,036	12,062	11,976
Val	6,346	4,014	4,076
Test	6,346	3,995	4,078
Total	31,728	20,071	20,130
Tops&Tees			
Train	16,121	9,925	12,054
Val	5,374	3,303	3,924
Test	5,374	3,210	4,112
Total	26,869	16,438	20,090

Table 6: Dataset statistics on Fashion IQ.

B. Fashion IQ Applications

The Fashion IQ dataset can be used in different ways to drive progress on developing more effective interfaces for image retrieval (as shown in Figure 2). These tasks can be developed as standalone applications, or can be investigated in conjunction. 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.

Single-shot Retrieval. Single-turn image retrieval systems have now evolved to support multimodal queries that include both images and text feedback. Recent work, for example, has attempted to use natural language feedback to modify a visual query [8, 14, 9]. By virtue of human-annotated relative feedback sentences, Fashion IQ serves as a rich resource for multimodal search using natural language feedback. We provide an additional study using Fashion IQ to train single-shot retrieval systems in Appendix C.

Relative Captioning. The relative captions of Fashion IQ make it a valuable resource to train and evaluate relative captioning systems [26, 57, 43, 15]. In particular, when applied to conversational image search, a relative captioner can be used as a user model to provide a large amount of low-cost training data for dialog models. Fashion IQ introduces the opportunity to utilize both attribute labels and human-annotated relative captions to train stronger user simulators, and correspondingly stronger interactive image retrieval systems. In the next section, we introduce a strong baseline model for relative captioning and demonstrate how it can be leveraged as a user model to assist the training of

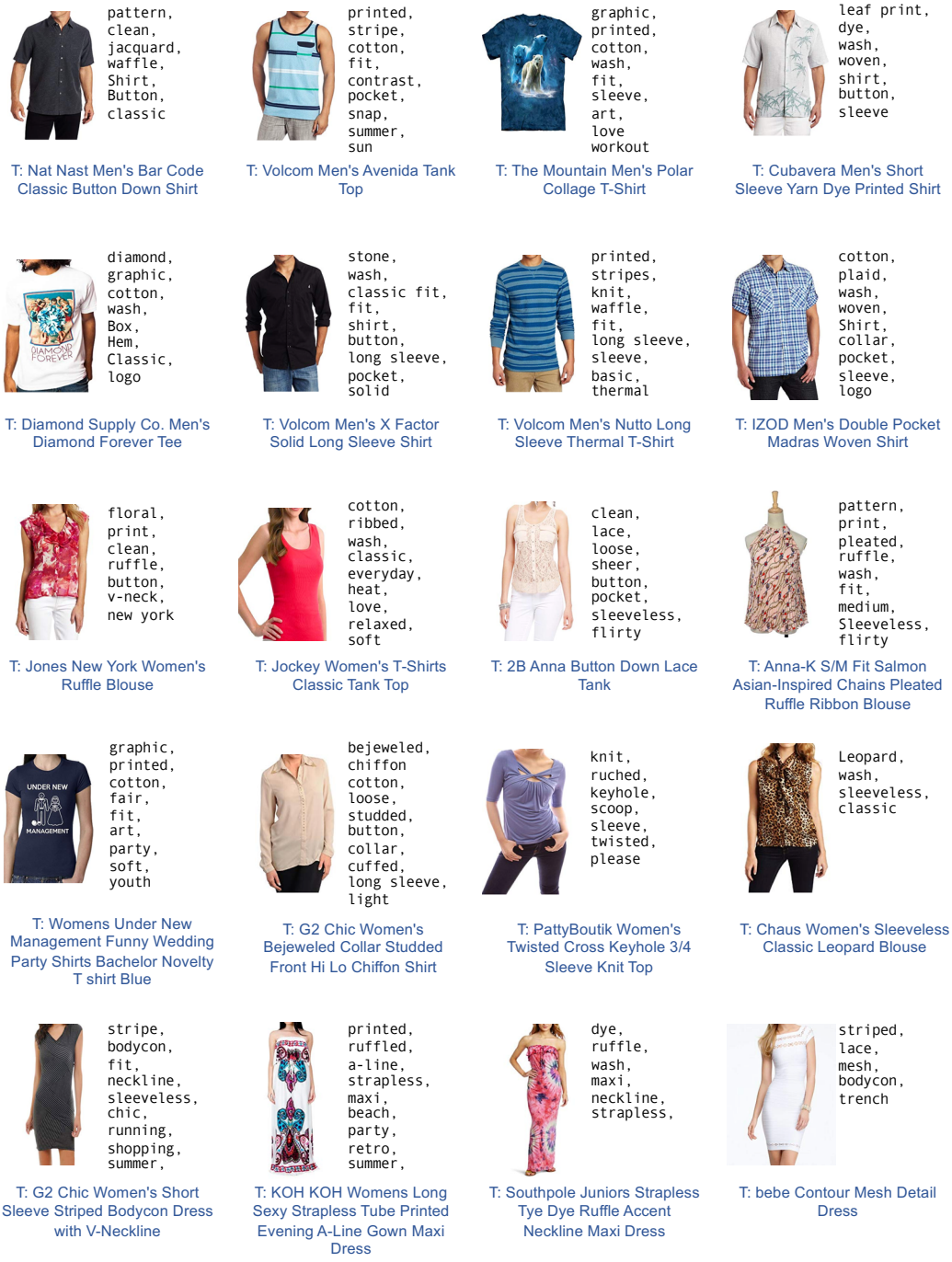

Figure 10: Examples of the original product title descriptions (T) and the collected attribute labels (on the right of each image).

a dialog-based interactive retriever.

Dialog-based Interactive Image Retrieval. Recently, dialog-based interactive image retrieval [18] has been proposed as a new interface and framework for interactive image retrieval. Fashion IQ with the large scale data (∼6x larger), the additional attribute labels, and the more diverse set of fashion categories, allows for more comprehensive research on interactive product retrieval systems. We will show next, how the different modalities available in Fashion IQ can be incorporated together effectively using a mul-

timodal transformer to build a state-of-the-art dialog-based image retrieval model.

C. Single-turn Image Retrieval

As discussed in Appendix B, we identified three main applications for our Fashion IQ dataset and we demonstrated how the dataset can be leveraged to achieve state-of-the-art performance in relative captioning and dialog-based image retrieval. We show here how Fashion IQ can be used in the third task, i.e., single-turn image retrieval.

In this task, given a reference image and a feedback sentence, we aim to retrieve the target image by composing the two modalities. The retrieval experiments use the portion of the dataset that has relative caption annotations. The two relative caption annotations associated with each image are treated as two separate queries during training and testing. This setting can be thought of as the single-turn scenario in an interactive image retrieval system and has a similar setup as previous work on modifying image query using textual descriptions [65, 8, 14, 9]. In this section, we provide empirical studies comparing different combinations of query modalities for retrieval, including relative feedback, image features, and attribute features. Specifically, the images were encoded using a pre-trained ResNet-101 network; the attributes were encoded based on the output of our Attribute Prediction Network; and the relative feedback sentences were encoded using Gated Recurrent Networks with one hidden layer. We used pairwise ranking loss [29] for all methods with the best margin parameters for each method selected using the retrieval score on the validation set. We include a baseline model from [18], which uses the concatenation of the image feature (after linear embedding) with the encoded relative caption features. We also included two models based on [65], with an additional gating connection, which allows the direct pass of one modality to the embedding space and has been shown to be effective for jointly modeling image and text modalities for retrieval.

We reported the retrieval results on the test set in Table 7. We found that the best performance was achieved by using all three modalities and applying a gating connection on the encoded natural language feedback (Model A). The gating connection on the text feature is shown to be effective for retrieval (comparing B and C), which confirms 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 (model D) significantly outperformed other modalities (models E and F). Finally, Removing attribute features resulted in generally inferior performance (comparing A and B) across the three categories, demonstrating the benefit of incorporating attribute labels, concurring with our observations in user modeling experiments and dialog-based retrieval experiments.

Figure 11: Examples of generated captions from the user model.

D. Additional Results on Interactive Image Retrieval

More extensive results on dialog-based retrieval are shown in Table 8 (including Dialog Turn 3). Additional ablative results on all of the three categories are shown in Table 10.

Initialization using Random vs. Similar Images. In the experiments presented in the main paper, we assume no prior knowledge on the user's search intent, so we start the search with a random image. This is more of a experimental detail and not tied to the dialog-based retrieval paradigm. The difference of starting the search with a similar image (based on product descriptions) as opposed to using a random image is that, the model retrieval performance progresses faster: average ranking percentile of 94.20 (use similar image) vs 93.22 (random start) at the first turn. Detailed comparison of the two initialization settings on our model (with attributes) is shown in Table 9.

		R@10 (R@50)		
		Dresses	Shirts	Tops&Tees
Multi-modality retrieval				
A	Image+attributes+relative captions, gating on relative captions.	11.24 (**32.39**)	**13.73** (**37.03**)	**13.52** (**34.73**)
B	Image+relative captions, gating on relative captions.	**11.49** (29.99)	13.68 (35.61)	11.36 (30.67)
C	Image+relative captions [18].	10.52 (28.98)	13.44 (34.60)	11.36 (30.42)
Single-modality baselines				
D	Relative feedback only.	6.94 (23.00)	9.24 (27.54)	10.02 (26.46)
E	Image feature only.	4.20 (13.29)	4.51 (14.47)	4.13 (14.30)
F	Attribute feature only.	2.57 (11.02)	4.66 (14.96)	4.77 (13.76)

Table 7: Results on single-turn image retrieval.

	Dialog Turn 1			Dialog Turn 3			Dialog Turn 5		
	P	**R@10**	**R@50**	**P**	**R@10**	**R@50**	**P**	**R@10**	**R@50**
Dresses									
Guo et al. [18]	89.45	6.25	20.26	97.49	26.95	57.78	98.56	39.12	72.21
Ours	93.14	12.45	35.21	97.96	36.48	68.13	98.39	41.35	73.63
Ours with Attr	**93.50**	**13.39**	**35.56**	**98.30**	**40.11**	**72.14**	**98.69**	**46.28**	**77.24**
Shirts									
Guo et al. [18]	89.39	3.86	13.95	97.40	21.78	47.92	98.48	32.94	62.03
Ours	92.75	**11.05**	28.99	98.03	30.34	**60.32**	98.28	**33.91**	63.42
Ours with Attr	**92.92**	11.03	**29.03**	**98.09**	**30.63**	60.20	**98.46**	33.69	**64.60**
Tops&Tees									
Guo et al. [18]	87.89	3.03	12.34	96.82	17.30	42.87	98.30	29.59	60.82
Ours	93.03	11.24	30.45	97.88	30.22	59.95	98.22	33.52	63.85
Ours with Attr	**93.25**	**11.74**	**31.52**	**98.10**	**31.36**	**61.76**	**98.44**	**35.94**	**66.56**

Table 8: **Dialog-based Image Retrieval.** We report the performance on ranking percentile (P) and recall at N (R@N) at the 1st, 3rd and 5th dialog turns.

	Dialog Turn 1			Dialog Turn 3			Dialog Turn 5		
	P	**R@10**	**R@50**	**P**	**R@10**	**R@50**	**P**	**R@10**	**R@50**
Dresses									
Similar	93.87	13.74	37.45	98.62	40.19	74.41	98.94	46.52	79.39
Random	93.50	13.39	35.56	98.30	40.11	72.14	98.69	46.28	77.24
Shirts									
Similar	93.42	10.55	30.56	98.34	30.85	61.69	98.59	34.50	66.20
Random	92.92	11.03	29.03	98.09	30.63	60.20	98.46	33.69	64.60
Tops&Tees									
Similar	95.29	23.67	40.22	97.00	35.63	63.36	97.43	38.49	67.24
Random	93.25	11.74	31.52	98.10	31.36	61.76	98.44	35.94	66.56

Table 9: Dialog-based image retrieval results when started with random or similar initial image pairs.

Visualization. Figure 11 shows examples of generated relative captions from the user model, which contain flexible expressions and both single and composite phrases to describe the differences of the images. Figure 12 shows examples of the user model interacting with the dialog based retriever. In all examples, the target images reached final rankings within the top 50 images. 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.

	Dialog Turn 1			Dialog Turn 3			Dialog Turn 5			Average		
	P	**R@10**	**R@50**	**P**	**R@10**	**R@50**	**P**	**R@10**	**R@50**	**P**	**R@10**	**R@50**
Dresses												
Retriever (R) + User (R)	89.45	6.25	20.26	97.49	26.95	57.78	**98.56**	39.12	72.21	95.17	24.11	50.08
Retriever (R) + User (T)	89.10	7.00	21.28	97.16	29.07	59.16	98.18	41.57	70.93	94.81	25.88	**59.46**
Retriever (T) + User (R)	92.29	11.61	33.92	**98.12**	36.18	**69.34**	98.52	**42.40**	**74.78**	96.31	30.06	59.35
Retriever (T) + User (T)	**93.14**	**12.45**	**35.21**	97.96	**36.48**	68.13	98.39	41.35	73.63	**96.50**	**30.09**	58.99
Shirts												
Retriever (R) + User (R)	89.39	3.86	13.95	97.40	21.78	47.92	**98.48**	32.94	62.03	95.09	19.53	41.3
Retriever (R) + User (T)	90.45	4.77	16.45	97.14	20.52	46.60	98.15	30.12	58.85	95.25	18.47	40.63
Retriever (T) + User (R)	91.77	9.33	27.15	98.02	27.25	57.68	98.41	30.79	62.53	96.07	22.46	49.12
Retriever (T) + User (T)	**92.75**	**11.05**	**28.99**	**98.03**	**30.34**	**60.32**	98.28	**33.91**	**63.42**	**96.35**	**25.10**	**50.91**
Tops&Tees												
Retriever (R) + User (R)	87.89	3.03	12.34	96.82	17.30	42.87	98.30	29.59	60.82	94.34	16.64	38.68
Retriever (R) + User (T)	90.31	5.75	18.10	97.73	27.72	56.42	**98.33**	**36.20**	**65.45**	95.46	23.22	46.66
Retriever (T) + User (R)	92.24	10.67	29.97	**97.90**	29.54	58.86	98.26	33.50	63.49	96.13	24.57	50.77
Retriever (T) + User (T)	**93.03**	**11.24**	**30.45**	97.88	**30.22**	**59.95**	98.22	33.52	63.85	**96.38**	**24.99**	**51.42**

Table 10: Detailed ablative studies on **Dialog-based Image Retrieval.** We report the performance on ranking percentile (P) and recall at N (R@N) at the 1st, 3rd and 5th dialog turns. R / T indicate RNN-based and Transformer-based models.

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