FashionNTM: Multi-turn Fashion Image Retrieval via Cascaded Memory Supplementary Material

Overview

In this supplementary material, we present some additional illustrations, and extended experiments from those we have in the main paper. Specifically, we show the following:

- In the main paper, we showed the ablation study on Shoes dataset. Here, we show similar results for the FashionIQ dataset.
- We provide some illustrations for the multi-turn version of the existing Shoes [1] dataset that we contributed to in our work. We further show how the annotations in this dataset are more specific, and consistent than multi-turn FashionIQ [2].
- We also add some details, and illustrations of the user study experiment that we conducted.
- Finally, we attach some video results where we demonstrate the interactive capability of our approach by taking feedback from real dynamic users at run-time.

1. Ablation Study on FashionIQ

In the main paper, we conduced ablation study on the multi-turn Shoes dataset. Here, we show results of similar experiments on the multi-turn FashionIQ dataset. In Table 1, we show an ablation over different number of memories of our proposed FashionNTM approach, by fixing the memory size to 8×768.

Table 1. Different number of memories for the proposed approach by fixing the memory size to 8×768 for FashionIQ dataset. We select the mean value for comparison and pick the best one.

Model	Number of memories (C)	R@5	R@8	Mean	% increase
ST+v-NTM	1	44.8	50.0	47.4	-
FashionNTM	2	45.0	50.1	47.5	0.2
	4	45.3	50.3	47.8	0.8
	8	**45.7**	**50.4**	**48.1**	**1.5**
	16	44.9	50.0	47.4	0

This is analogous to Table 5 in the main paper, where we conducted a similar experiment for multi-turn Shoes dataset. We observe a similar trend here too, as the performance gradually increases with more cascaded stages, before peaking at $C = 8$, and then reduces as the memory network becomes larger at $C = 16$. An interesting observation in Table 1 is that the relative improvement is not as high as that for Shoes dataset (see Table 5 of main paper). This is primarily due to the differences between the annotation quality of the two datasets.

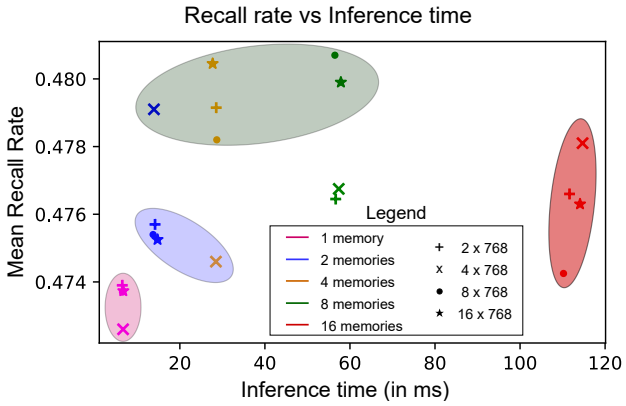

Figure 1. Scatter plot showing recall versus inference time for multiple memories in our CM-NTM for the Multi-turn FashionIQ dataset. The configurations belonging to the green cluster give the best recall overall, while having a high inference time. The configurations in the blue cluster provide a suitable alternative with quicker inference at the cost of lower recall. The magenta and red clusters are undesirable configurations due to poor performance and long inference time, respectively.

In Figure 1, we plot the relative performance of all the experiments that we conducted on the multi-turn FashionIQ dataset. This is analogous to Figure 8 of the main paper. We observe the same type of trend for both datasets – inference time increases as we add more memories. Configurations in the green and blue clusters are desirable, as they provide a good trade-off between recall performance and computation time, while the pink and red clusters are undesirable due to poor performance and longer inference time, respectively.

2. Difference between FashionIQ and Shoes

During our experiments, we observed that descriptions in the multi-turn FashionIQ [2] dataset feedbacks are very generic, and therefore can correspond to multiple target images. An example of this is shown in Figure 2, where we demonstrate additional examples of the top-5 retrieval experiments, that we conducted in Figure 4a of the main paper. As seen from all three cases, there are *multiple* target images, that match the description as provided by the feedback (e.g. in the top row, all 5 retrieved images are "purple and flowing", and 4 of them are also "long sleeved"). However, only one of them is labeled as the correct target (ground-truth) for a particular transaction. This makes it difficult to evaluate multi-turn systems on this dataset, as the performance (recall rate) might be low, even though the requirements are satisfied.

In contrast, for the multi-turn Shoes dataset, feedback

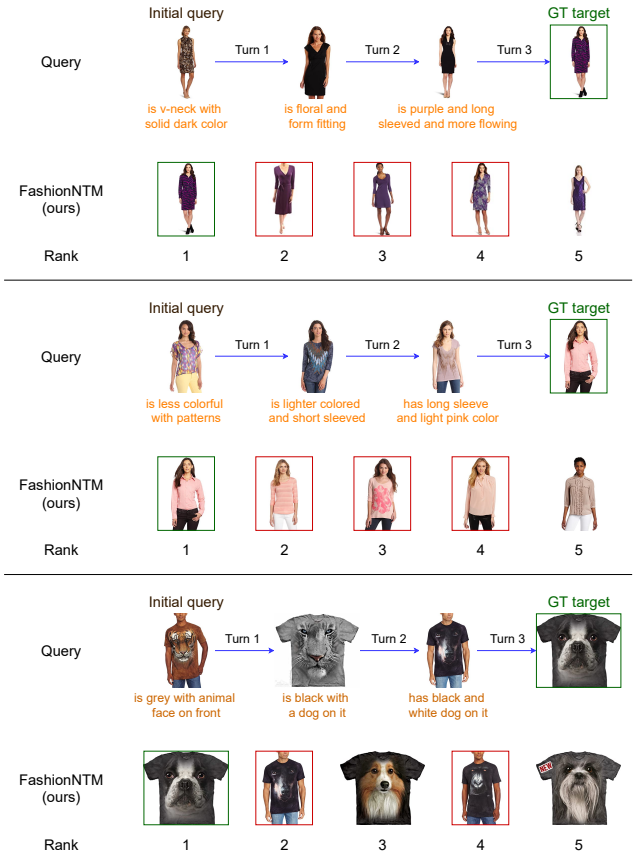

Figure 2. Top-5 retrievals of our method on FashionIQ dataset. The images with green bounding box correspond to ground-truth target. The images with red bounding box have similar properties as those of the GT, but are not marked. This is because the annotations in FashionIQ dataset are too generic, and hence can correspond to multiple images.

Figure 3. Top-5 retrievals of our method on Shoes dataset. The images with green bounding box correspond to ground-truth target. Due to more consistent and accurate descriptions, the feedback text in this dataset typically corresponds to only one image.

texts are more concise, and hence there are usually very few target images ($\approx$ 1 or 2) that match the description given in the feedback. This makes the Shoes dataset more suitable for evaluating multi-turn systems. Some more top-5 retrieval results for multi-turn Shoes dataset, in addition to those in Figure 4b of the main paper, are shown in Figure 3.

3. User study details

In the main paper, we showed the results of a user preference study conducted among 5 human participants. In Figure 4, we show some examples of the actual interface that was shown to the users. Each participant was shown 45 such images, and asked a simple question about "Which model's performance are you most satisfied with?" Based on their understanding, the users had to choose between the top images retrieval of 3 models. To maintain fairness, we did not disclose the identity of the models to the participants. At the end of the study, we aggregated the votes given to the models by each user, and plotted it as a histogram as shown in Figure 7 of the main paper.

4. Video demonstration of interactive qualitative results

We attach a supplemental video, named `iccv2023_supplementary_video.mp4` as part of our submission package. In the first part of the video, we do a walk-through of the overall pipeline of our proposed multi-turn fashion image retrieval algorithm, FashionNTM. In the next part, we perform a demonstration of two interactive user experiments, where we take user input across multiple turns to retrieve fashion images.

The first experiment highlights the memory retention capability of our approach. We show that using a memory-based model, we can learn information across multiple turns, while that is not the case with a non-memory retrieval model, which only considers a single turn feedback into account for retrieval.

The second experiment highlights an important property of a multi-turn system, in that it should be independent of the order in which feedbacks are provided, as long as they

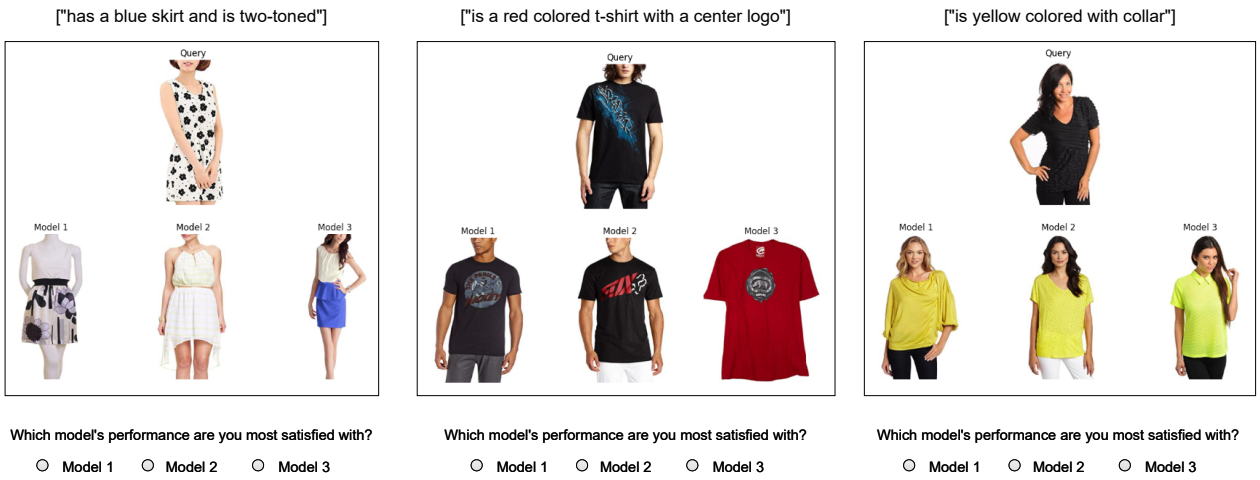

Figure 4. User interface shown to the participants for the human preference study. Here, we show three examples of top-1 retrievals of 3 best multi-turn models on the multi-turn FashionIQ dataset. Model 1 corresponds to ST + EWMA, while Model 2 to ST + LSTM. Finally, our proposed FashionNTM model is Model 3. For fairness, we did not disclose the identity of the models to the participants.

are non-contradictory. We demonstrate this property by taking two user inputs for a given query image, and then run our retrieval system twice – once with the original sequence, and then with the reversed sequence order. We show that in both the cases, the final retrieved product look similar, even though the intermediate outputs are very different.

References

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 1

[2] Y. Yuan and W. Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 839–848, 2021. 1