FashionNTM: Multi-turn Fashion Image Retrieval via Cascaded Memory

Anwesan Pal*[1], Sahil Wadhwa[2], Ayush Jaiswal[2], Xu Zhang[2], Yue Wu[2], Rakesh Chada[2], Pradeep Natarajan[2], and Henrik I. Christensen[1]

[1]UC San Diego, [2]Amazon Alexa Natural Understanding,
{a2pal, hichristensen}@ucsd.edu, {sahilwa, ayujaisw, xzhnamz, wuayue, rakchada, natarap}@amazon.com

Abstract

Multi-turn textual feedback-based fashion image retrieval focuses on a real-world setting, where users can iteratively provide information to refine retrieval results until they find an item that fits all their requirements. In this work, we present a novel memory-based method, called FashionNTM, for such a multi-turn system. Our framework incorporates a new Cascaded Memory Neural Turing Machine (CM-NTM) approach for implicit state management, thereby learning to integrate information across all past turns to retrieve new images, for a given turn. Unlike vanilla Neural Turing Machine (NTM), our CM-NTM operates on multiple inputs, which interact with their respective memories via individual read and write heads, to learn complex relationships. Extensive evaluation results show that our proposed method outperforms the previous state-of-the-art algorithm by 50.5%, on Multi-turn FashionIQ [60] – the only existing multi-turn fashion dataset currently, in addition to having a relative improvement of 12.6% on Multi-turn Shoes – an extension of the single-turn Shoes dataset [5] that we created in this work. Further analysis of the model in a real-world interactive setting demonstrates two important capabilities of our model – memory retention across turns, and agnosticity to turn order for non-contradictory feedback. Finally, user study results show that images retrieved by FashionNTM were favored by 83.1% over other multi-turn models.

1. Introduction

Image retrieval has been extensively studied in the computer vision community, both using classical approaches [10, 52, 25] and recently, using learning-based techniques [2, 15, 41, 43, 23]. Existing works can be grouped based on input queries considered – from image-only queries, commonly known as Content-Based Image Retrieval (CBIR) [35, 39, 48], to attributes [18], sketches [44], and natural language [32, 61]. However, most of these methods do not incorporate interactive user feedback, which is necessary for a personalized task such as fashion retrieval.

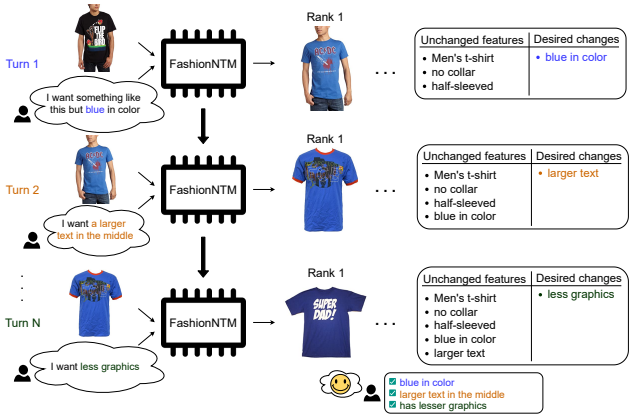

Figure 1: Illustration of multi-turn fashion image retrieval. Initially (Turn 1), the system receives an initial query image, and a textual feedback mentioning the user's desired changes. The model then retrieves a ranked list of closest matching images. Subsequently, the user keeps refining their choice by providing more feedback, while the model retrieves newer images by considering both current and past feedback. This continues until the multi-turn system has successfully obtained the final retrieved image (Turn N) with all the desired properties mentioned across every past turn.

Textual feedback-based fashion image retrieval allows users to refine online shopping search results by providing information about how the results differ from their desired product (e.g., "a dress like this but darker in color"). Several approaches for implementing such a system have been proposed recently [14, 7, 31, 56, 17, 60, 63], which involve learning a joint representation via multi-modal information fusion across the query (reference) image and the associated feedback, and using it to retrieve the closest matching image in the database (product catalog) as the target.

Popular methods for fashion retrieval task [14, 31, 7,

*Work primarily done during internship at Amazon. Additional details are available at https://sites.google.com/eng.ucsd.edu/fashionntm.

4, 23] involve *single-turn* exchange of information, where users provide feedback exactly once to update the search results. However, this is not characteristic of the real-world setting as online shopping customers typically start with a general idea of what they want and iteratively update the requirements until they find something that matches their desired features. This usually involves providing additional attributes, or modifying previously specified features in each turn to refine the search results. The ideal feedback-based fashion image retrieval system is, hence, inherently *multi-turn*, as illustrated in Figure 1.

There are two major challenges associated with multi-turn image retrieval. First, there is a lack of sufficient training and evaluation datasets – despite the abundance of single-turn fashion retrieval datasets, to the best of our knowledge, there is only one publicly available multi-turn fashion image retrieval dataset [60] currently. This is because labeling a sequence of images while ensuring continuity, consistency, and uni-directional information flow is a difficult problem. Thus, to facilitate research in this domain, we created a new fashion image retrieval dataset to allow for further benchmarking. Second, generalizing performance to real-world dynamic user interactive cases is non-trivial – as this is still a relatively new research domain, most existing algorithms do not generalize beyond the training dataset to consider multiple turns of interactive feedback. In this work, we propose a novel memory-based framework to explicitly consider sequential feedback from users across multiple turns to retrieve desired items, both for the static image datasets, as well as real-world dynamic users.

Sequential modeling is a relatively mature field of research. However, a majority of the existing approaches [46, 27, 49, 20, 9] do not maintain an explicit memory, and therefore cannot learn long and complex information. Vanilla memory network-based methods, which explicitly maintain an external memory, could be used for retaining past information, but they do not provide a robust mechanism to iteratively update their memory [55, 51]. In contrast, Neural Turing Machines (NTMs) [16] provide a fully differentiable model with sophisticated read and write operations to extract and update historical information in its explicit memory via an attention mechanism. Therefore, in this work, we build on NTMs to develop a novel framework for the multi-turn retrieval task. We further propose a novel Cascaded Memory Neural Turing Machine (CM-NTM) that allows us to encode multiple relationships from the features of a particular turn and store them over time across multiple memories in a multi-turn setting. This is similar to how multi-head attention (MHA) operates for transformers [53]. To ensure that the individual memories effectively utilize each other's information, we link them together in a *cascaded* fashion. Evaluation results demonstrate that our proposed approach improves the retrieval performance as compared to the previous state-of-the-art by 50.5% on Multi-turn FashionIQ, and by 12.6% on Multi-turn Shoes.

In summary, we make the following contributions. First, we propose a state-of-the-art memory-based framework, called FashionNTM, for multi-turn feedback-based fashion image retrieval, that uses an external memory to learn complex long-term relationships. Second, we develop a novel Cascaded Memory Neural Turing Machine (CM-NTM), that extends NTM to learn relationships across multiple inputs via additional controllers and read/write heads in a cascaded fashion. Third, we conduct experiments to show that the proposed approach outperforms existing state-of-the-art retrieval models by 50.5% on Multi-turn FashionIQ [60], and around 12.6% on the multi-turn version of Shoes dataset [5] respectively. Additionally, by performing an interactive analysis, we demonstrated two important capabilities of our multi-turn system – memory retention across turns, and agnosticity to turn order for non-contradictory feedback. Finally, a user study result shows that on an average, the images retrieved by our model are preferred 83.1% more than those from other multi-turn methods.

2. Related Work

Single turn feedback-based fashion image retrieval - Previous works in feedback-based fashion image retrieval have primarily focused on the single-turn scenario [14, 40, 31, 7, 54, 4, 3, 59, 6, 36, 22], where a model is provided with a reference image along with an associated feedback text highlighting the desired attribute changes. The typical approach is to encode the multi-modal image and text input using pre-trained visual feature extractors [19, 24, 34] and sequential natural language processors [20, 12], respectively. More recently, pre-trained extractors such as Contrastive Language-Image Pre-training (CLIP) [45] have also been used [4, 3]. This is then followed by a transformer-based decoder network [53] for generating information-rich features for image retrieval from the database. Although these methods perform well in single-turn settings, they cannot be directly used for real-world applications that deal with multiple turns of information exchange.

Multi-turn visuo-linguistic methods - A few methods have been proposed recently for fusing visual and textual input across multiple turns of information exchange. A popular application has been the video dialog task [42, 21, 33, 30, 17], where a trained system is asked to answer questions based on an ongoing video dialog. However, these kinds of dialogs are primarily text-based for both the questioner and the answering agent, without any interaction across image and text inputs. In the fashion domain, there have been some early works for the multi-turn retrieval task. Guo *et al.* [17] proposed a model that uses convolutional neural networks (CNNs) for encoding images and text, followed by a recurrent neural network for aggregating sequences. Then,

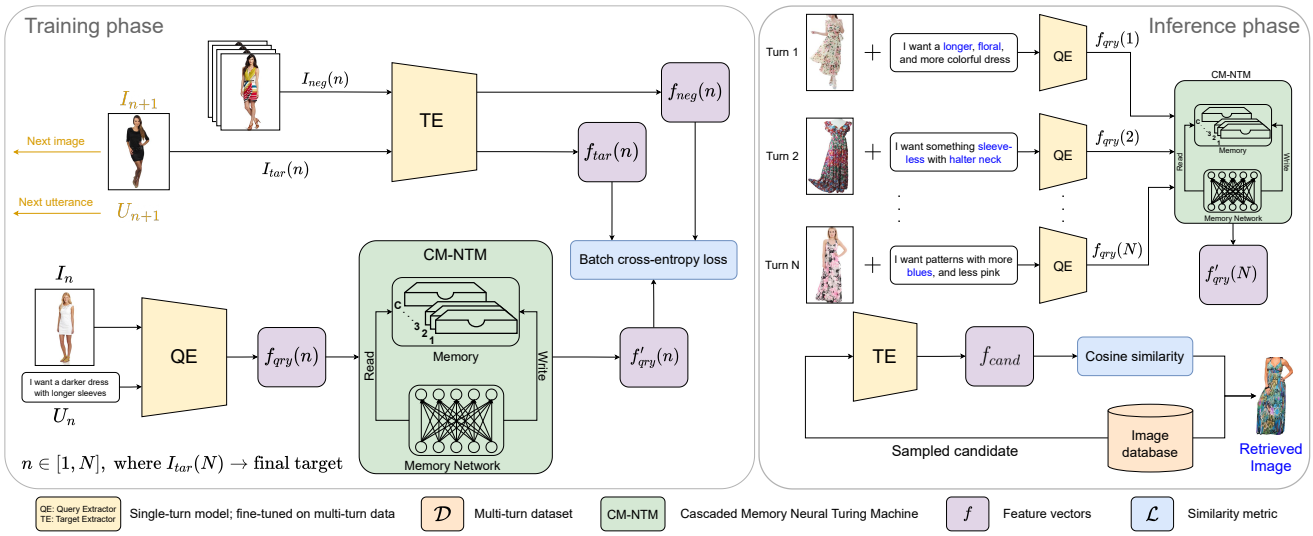

Figure 2: The complete FashionNTM framework. During training, the model receives an input query image, I_n, and associated feedback, U_n. From these, query features $f_{qry}(n)$ are computed using an off-the-shelf single-turn fashion image retrieval model. These features are then fed into the memory network which learns to retain useful information from the current turn, and combine it with information from previous turns by interacting with an external memory bank, via read/write operations. The memory-modified features $f'_{qry}(n)$ are compared with the target index features $f_{tar}(n)$, and other negative samples $f_{neg}(n)$, to compute the training loss. At inference, the model receives a series of multi-modal inputs turn-wise and computes the final modified feature $f'_{qry}(N)$ using the trained memory network. This is compared with features f_{cand} derived from different candidate images in the database to retrieve the closest matching images.

a k-Nearest Neighbor search is performed across sampled images to get the closest match. The entire model is trained end-to-end using reinforcement learning (RL). Inspired by this, Zhang *et al*. [63, 62] proposed two approaches for enhancing the text-image feature fusion by adding constraints, and using offline interactive recommendation. Recently, Yuan *et al*. [60] released the first multi-turn fashion image retrieval dataset, based on the original single turn FashionIQ [56]. They also proposed a state-of-the-art model, which we directly compare with our approach.

Memory networks for vision and language - Memory networks have been widely used for a number of natural language processing and computer vision applications. Some works [55, 51, 13, 11] utilize it for Sentence Video Questions and Answering (QA) task. Another popular application is video object segmentation [8, 58, 50, 37]. Recently, there have been some works on including memory in transformer architectures [26, 47, 57, 38]. However, these approaches design their memory to be used only for specific tasks, and hence cannot be directly compared with ours. In this work, we propose a memory network based method for the multi-turn fashion image retrieval task.

3. FashionNTM

The multi-turn feedback-based image retrieval task can be viewed as a series of information exchange transactions. We define a transaction as one session of query context comprising a query image and the associated feedback text. Notationally, an N-turn transaction is represented as $T = [(I_1, U_1), (I_2, U_2), \cdots, (I_N, U_N)]$, where I_n and U_n correspond to the query image and feedback utterance respectively, at turn $n \in [1, N]$. Given such a transaction, the aim of a multi-turn model is to iteratively retrieve the final desired target image $I_{tar}(N)$ by ranking candidates in the fashion image database based on a matching score.

The overall pipeline of our approach, called FashionNTM, is illustrated in Figure 2. We start with a single-turn feature extraction module to encode the multi-modal image and text inputs of each turn n in a multi-turn transaction. It comprises two parallel blocks – (i) a query feature extractor (QE), for processing I_n and U_n to generate the joint query representation $f_{qry}(n)$, and (ii) a target feature extractor (TE), for encoding all the images in the database into their corresponding index features. For the ground-truth target image $I_{tar}(n)$, we call these features $f_{tar}(n)$, while for every other sample $I_{neg}(n)$ in the batch, we name it $f_{neg}(n)$. The query feature $f_{qry}(n)$ is fed turn-wise to the Cascaded Memory Neural Turing Machine (CM-NTM) block. CM-NTM first computes several derived features from the original query feature. Subsequently, the original query feature and each of the derived features interact with their own controllers, read/write heads, and sequentially update the memories in a cascaded manner, *i.e*. output of one memory goes as input to the next. Ultimately, we get the enhanced feature $f'_{qry}(n)$ as the final output, which is then compared with the target feature $f_{tar}(n)$ using a similarity score-based batch

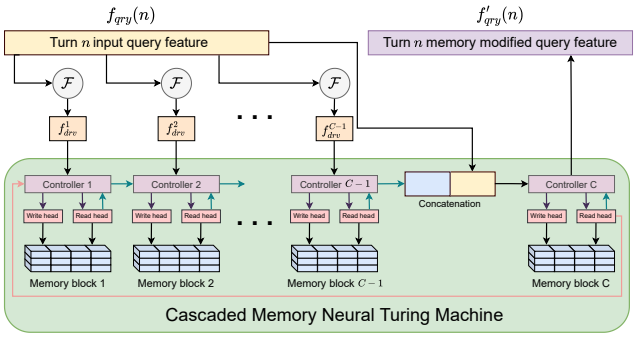

Figure 3: Cascaded Memory Neural Turing Machine (CM-NTM). It consists of C controllers; one for the query feature and $C-1$ for the derived features. Each controller interacts with its own memory using its own read and write heads. The controllers are linked with each other, forming a cascaded chain. The modified features from the last memory are output as the final embedding.

loss function. This loss treats $f_{tar}(n)$ as a positive sample and every other feature in the batch, $f_{neg}(n)$, as a negative sample. We use cosine-similarity [54, 31, 14] between feature vectors as the matching score. During inference, the model receives a sequence of multi-modal turn-wise inputs and computes the final modified feature $f'_{qry}(N)$. This is compared with feature f_{cand} derived from different candidate images in the database, and the closest match is retrieved as output. In the following sections, we describe the key components of our framework.

Single-turn feature extractor - We utilize the state-of-the-art single-turn image retrieval model, FashionVLP [14], for extracting multi-modal text and image features. FashionVLP extracts image embeddings at multiple levels of granularity and incorporates a vision-language pre-trained transformer for fusing these encodings with text feedback to obtain multi-modal query features. It adopts a convolutional neural network (CNN) architecture with contextual attention for fusing fashion-contextual image features of target images. Specifically, for each turn of a multi-turn transaction, we obtain query features f_{qry} and target features f_{tar}. An N-turn retrieval transaction is of the form $T_{retr} = [f_{qry}(1), f_{qry}(2), \cdots, f_{qry}(N)]$, with the target feature representation given by $f_{tar}(N)$.

3.1. Cascaded Memory Neural Turing Machine

To learn relationships across transactions, we propose a novel Cascaded Memory Neural Turing Machine (CM-NTM) module. CM-NTM allows multiple inputs to interact with their own memories using individual read and write heads to learn multiple complex relationships in the input data. We achieve this by deriving several features from the original n-th turn query feature $f_{qry}(n)$ using a projection function $\mathcal{F}$, which comprises a fully-connected (FC) layer with batch normalization. Specifically, for a C-stage cascaded CM-NTM, we obtain the derived feature $f^c_{drv} = \texttt{BatchNorm}(\texttt{FC}_\texttt{c}(f_{qry}(n)))$, where $c \in [1, C-1]$, and $\texttt{FC}_\texttt{c}(\cdot)$ is the FC layer for the c-th stage. Having obtained C inputs comprising the original query and $C-1$ derived features, we pass them sequentially to our memory network. Figure 3 presents our memory network architecture. It consists of three main components[1]: Controller, Read/Write heads, and Memory blocks.

Controller - The vanilla NTM has a single controller which takes the query feature at turn n, along with the previous turn's read vector, and emits an intermediate controller output. This is used by the read and write heads to compute the current turn's attention weights, which are then used to update the memory. In our work, we introduce C different controllers – one for each of our inputs. Each controller, $c \in [1, C]$ can therefore interactively update its memory via individual read and write heads, allowing it to learn multiple complex relationships in each turn.

Read/Write heads - The controller output is fed to these heads. The write head learns to generate erase and add parameters, which are used to update the current memory. Similarly, the read head generates an attention weight vector, which is used to obtain a weighted sum over the memory locations to get the read vector $\mathtt{r}_{out}(n)$. In our work, we have separate read and write heads for each memory.

Memory block - This is represented as a 2-D matrix of the form $N \times M$, where N corresponds to memory locations and M to the vector size at each location. The output of each memory block is a fused representation of the controller output and the read vector. In our cascaded multi-memory setup, the controllers are linked in a chain, such that the memory-modified features are sequentially propagated. Specifically, for controller $c \in [1, C]$ at turn n, the input is given by $\texttt{input}^c(n) = [\mathtt{r}^{c-1}_{out}(n); f^c_{drv}; \mathtt{r}^c_{out}(n-1)]$, where $\mathtt{r}^0_{out}(n) = \mathtt{r}^C_{out}(n-1)$, $f^C_{drv} = f_{qry}(n)$, and ; represents concatenation. The final output, $f'_{qry}(n)$, is the fused representation of the last controller output $\texttt{output}^C_{ctrl}$ and final read vector $\mathtt{r}^C_{out}$. We experiment with a different number of memories C and memory sizes in Section 4.7.

3.2. Loss function

We adopt a batch cross-entropy loss [14], where each entry in a batch acts as a negative sample for all other entries. In the multi-turn setting, we compute the loss function turn-wise. Given a batch size B, the loss between predicted feature $\mathbf{x}_n = [^1 f'_{qry}(n), ^2 f'_{qry}(n), \ldots, ^B f'_{qry}(n)]$, and ground-truth $\mathbf{y}_n = [^1 f_{tar}(n), ^2 f_{tar}(n), \ldots, ^B f_{tar}(n)]$, at turn n is of the form

$$\mathcal{L}(\mathbf{x}_n, \mathbf{y}_n) = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{e^{\kappa(^i f'_{qry}(n), ^i f_{tar}(n))}}{\sum_{j=1}^{B} e^{\kappa(^i f'_{qry}(n), ^j f_{tar}(n))}}$$

[1]We mainly discuss our novel changes to the NTM design. For details regarding the vanilla architecture, please refer to the original paper [16].

Table 1: Statistics of the datasets used in this work. The top 2 rows are for the recently proposed Multi-turn FashionIQ dataset [60]. The bottom two rows are for the Multi-turn version of the Shoes [5] dataset, which we created as part of our work.

Dataset	Split	Number of transactions with 2-turns	3-turns	4-turns	Total images
Multi-turn FashionIQ [60]	Train	6,897	1,733	475	10,438
	Test	1,752	483	165	6,274
Multi-turn Shoes (ours)	Train	1,659	1,036	982	11,030
	Test	296	96	28	4,631

where κ is the cosine similarity metric [14]. In this way, each ${}^{j}f_{tar}(n)$ in the batch, $\forall j \in [1, B]$ and $j \neq i$, serves as a negative sample $f_{neg}(n)$ for a given ${}^{i}f'_{qry}(n)$. The turn-wise retrieval loss function is represented as $\mathcal{L}^{n}_{retr} = \mathcal{L}(\mathbf{x}_n, \mathbf{y}_n)$, with the overall loss function for our proposed multi-turn model given by $\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}^{n}_{retr}$.

4. Experimental Evaluation

4.1. Datasets

Multi-turn FashionIQ [60]: To the best of our knowledge, this is the only existing fashion dataset with multi-turn sessions, where each turn is derived from the original single-turn FashionIQ [56] dataset. It comprises of 11,505 sessions across three clothing types – dress, top-tee, and shirt. The dataset is split into transactions of 2-turns, 3-turns, and 4-turns. In each turn, the data is represented as a pair (I_n, U_n), where I_n and U_n correspond to the query image and the feedback text for turn n.

Multi-turn Shoes: The original Shoes dataset [5] contains images of 10 categories of women's shoes obtained from the web along with automatic labeling of attributes. Guo *et al*. [17] provide additional natural language descriptions of the images to make them suitable for *single-turn* feedback-based image retrieval. This resulted in about 10k training pairs and 4.6k test queries. In this work, we create a multi-turn extension of this dataset to further research in this domain. Like the approach described in [60], we concatenated several single-turn transactions by matching the target image of one session to the query of another. To maintain consistency with Multi-turn FashionIQ, we also developed transactions of 2-turns, 3-turns, and 4-turns. Table 1 provides the statistics of both the datasets.

4.2. Implementation Details

Single-turn feature extractor pre-training - We use the recently proposed single turn (ST) fashion image retrieval model, FashionVLP [14], to extract the query and target features for both datasets. To ensure a fair comparison with other algorithms, we re-train FashionVLP only on those single-turn queries that are part of the multi-turn dataset. The implementation details and hyperparameters are similar to those mentioned in Goenka *et al*. [14].

CM-NTM training - We build our CM-NTM model using the open-source implementation [28] of NTM [16]. We implement a separate controller for each memory in our cascaded design using Long Short Term Memory (LSTM) networks [20]. The read and write heads are composed of multi-layer perceptrons (MLPs) and attention networks. We use $C = 4$ for Shoes, and $C = 8$ memory stages for FashionIQ, as it is a larger dataset. To ensure equal turn lengths for batch training, we pad the short turn transactions by repeating the last transaction similar to Yuan *et al*. [60]. We train our models for 100 epochs with a batch size of 80 and a learning rate of 1e-4. We used the PyTorch framework with Adam [29] optimizer for training.

4.3. Baselines and Previous State-of-the-art

We compare the performance of our model with six other single-turn and multi-turn approaches.

Single-turn methods - In these methods, we aggregate the multi-turn query data into a *single* feature without iterating over them turn-wise. The first baseline is $\texttt{ST} + \texttt{avg}(\texttt{all turns})$, where we take the mean of all the N query features in a transaction to get a single mean query $\bar{f}_{qry}(N)$. This is compared with the candidate target features $f_{tar}(N)$. Next, we have $\texttt{ST} + \texttt{cat}(\texttt{all captions})$, where all the captions of a multi-turn transaction are concatenated into one long caption, along with initial reference image, to get query features $\hat{f}_{qry}(N)$. This is compared with target features $f_{tar}(N)$ to retrieve the final images.

Table 2: Quantitative results on Multi-turn FashionIQ [60]. We compare with multiple single-turn and multi-turn baselines, and state-of-the-art works [17, 60]. Results show the superior performance of our proposed approach on the popularly used recall rate evaluation metric.

Model	Dress R@5	Dress R@8	Toptee R@5	Toptee R@8	Shirt R@5	Shirt R@8	Overall R@5	Overall R@8	Overall Mean
Single-turn									
ST + avg(all turns)	25.6	32.5	32.1	38.1	27.0	32.3	28.2	34.3	31.3
ST + cat(all captions)	30.2	36.3	36.0	44.1	35.3	42.5	33.8	41.0	37.4
Multi-turn									
Dialog Manager [17]	12.7	16.7	11.6	15.8	13.9	17.7	13.1	15.2	14.2
CFIR [60]	29.8	33.5	29.4	33.6	30.5	34.1	30.3	33.4	31.9
ST + EWMA (ours)	42.0	48.4	43.8	**50.9**	36.9	44.2	40.9	47.8	44.4
ST + LSTM (ours)	47.8	52.5	44.4	50.5	41.6	47.6	44.6	50.2	47.4
FashionNTM (ours)	**48.3**	**52.8**	**45.1**	49.8	**43.8**	**48.8**	**45.7**	**50.4**	**48.1**

Multi-turn methods - In these methods, the multi-turn data is fed turn-wise to the model. The first method is `Dialog Manager` [17], which employs reinforcement learning (RL) framework to learn relationships between turns. Next, we have the previous state-of-the-art Conversational Fashion Image Retrieval `CFIR` method [60], which encodes *only* the text using a transformer network [53], and uses a simple Gated Recurrent Unit (GRU) layer for multi-turn image retrieval. The third baseline is `ST + EWMA`, where we use exponential weighted moving average [1] to *heuristically* aggregate past turns, with more weights given to recent history. Finally, we have the `ST + LSTM` model that uses a single-layer LSTM [20] with a hidden size of 100 for aggregating past information.

4.4. Results

Evaluation Metrics - Following [14, 60, 17], we evaluate models using the standard top-`K` recall (*i.e.* `R@K`) for image retrieval. Overall performance are compared specifically on the average of R@5 and R@8.

Quantitative results on multi-turn datasets - In Table 2, we compare the results of different methods on the multi-turn FashionIQ dataset. We observe that the multi-turn baselines generally perform better than single-turn methods. This is expected as aggregating data by naïvely averaging/concatenating loses feedback content and turn-order information and hence is likely to miss out on important cues. Our memory-based approach outperforms all the other multi-turn baselines by a large margin. This shows that the memory network can store and retrieve useful information to and from the memory between intermediate turns, which allows it to keep track of past information better than other networks that do not use explicit memory. The 50.5% performance gain over the previous state-of-the-art highlights our model's capability to learn meaningful representations over multiple turns of conversational feedback.

In Table 3, we provide a similar comparison for the Multi-turn Shoes dataset. Consistent with results for Multi-turn FashionIQ, the multi-turn baselines perform better than all the single-turn ones, and our memory network-based approach performs the best with a relative improvement of 12.6%. The difference in performance across different models is more pronounced in these results than in Table 2. This is possibly because the annotations are cleaner and more consistent in this dataset as compared to Multi-turn FashionIQ [60]. For instance, multiple images in the FashionIQ dataset can match a particular query, but only one of them is labeled as the ground-truth.

Table 3: Comparison with existing single-turn and multi-turn models on the multi-turn version of the Shoes [5] dataset. We compare with multiple single-turn, and multi-turn baselines. Comparative analysis shows the superior performance of our proposed approach using the popular recall rate evaluation metric.

Model	R@5	R@8	Mean
Single-turn			
ST + avg(all turns)	12.4	17.6	15.0
ST + cat(all captions)	10.7	13.6	12.2
Multi-turn			
ST + EWMA (ours)	18.3	23.8	21.1
ST + LSTM (ours)	23.3	32.1	27.7
FashionNTM (ours)	**26.7**	**35.7**	**31.2**

An important property of a good multi-turn system is that performance should be robust to the length of the historical information (number of past turns). For example, even for a large number of previous turns considered, the model should efficiently retain desirable details, while filtering out unnecessary information. In Table 4, we analyze this property for models with and without memory. A single-turn model that assumes ground-truth information for all past turns, and evaluated only on the final-turn, is taken as reference. This is expected to be the upper-bound on the performance in a single-turn setting, as perfect information about the history is guaranteed. We vary the number of past turns included in the input transaction history (versus treated as ground-truth) for the multi-turn models (with or without memory) in order to evaluate their effectiveness at capturing and utilizing past information. As seen from the table, for a model without memory, the performance depreciates significantly with each additional turn from the

Table 4: Importance of aggregating historical data using memory-based vs non-memory approach. In the first row, we show the result of a model using only the final turn information of a multi-turn transaction. This assumes the groundtruth retrieval for all previous turns, and therefore provides the upper-bound on single-turn performance for final retrieval. Subsequently, we include additional information from the history, and compare performance across models with and without memory. As seen from the non-memory case, the performance depreciates a lot ($\approx 64.9\%$ difference compared to the final turn's performance), as we try to aggregate longer historical information. In contrast, for the memory network model, it can be seen that the performance is fairly consistent ($\approx 0.5\%$ difference) across the turn length.

Memory usage	Turn configuration	Dress		Toptee		Shirt		Overall			Difference from final turn
		R@5	R@8	R@5	R@8	R@5	R@8	R@5	R@8	Mean	
-	only final turn	77.9	77.9	84.0	84.0	74.1	77.8	78.7	79.9	79.3	-
Experiments with data aggregated across multiple-turns											
Only single turn (w/o memory)	last two turns	51.3	58.4	56.0	76.0	44.4	48.1	50.6	60.9	55.8	−29.6%
	last three turns	33.6	43.4	32.0	44.0	29.6	40.7	31.8	42.7	37.3	−53.0%
	all turns	18.6	28.3	32.0	32.0	25.9	29.6	25.5	30.0	27.8	−64.9%
FashionNTM (with memory)	last two turns	76.1	77.9	84.0	84.0	77.8	77.8	79.3	79.9	79.6	+0.4%
	last three turns	77.0	77.9	84.0	84.0	77.8	77.8	79.6	79.9	79.8	+0.6%
	all turns	76.1	77.9	84.0	84.0	77.8	77.8	79.3	79.9	79.6	+0.4%

(a) Multi-turn FashionIQ. From left to right, the images belong to the Dress, Shirt, and Toptee categories respectively.

(b) Multi-turn Shoes. From left to right, we have three different samples of shoes from the dataset.

Figure 4: Top-5 final image retrievals on the evaluated multi-turn datasets. The top row illustrates 3 sets of multi-turn query session. We consider three different multi-turn models - `EWMA`, `LSTM`, and our proposed `FashionNTM`. As seen from the retrievals, our proposed model correctly predicts the target image in all 3 cases for the FashionIQ dataset, and in 2 out of 3 cases for the Shoes dataset.

past treated as input rather than ground-truth. As we go further back in history, the performance consistently reduces. However, for our multi-turn model with memory, the performance does not change appreciably as the number of turns change. This shows that our proposed approach can successfully retain/filter out past data based on their relevance, across various lengths of history.

An interesting observation is that having only the final turn with memory does not yield a good result. This is possibly due to the initialization method of the memory network, which is random. Hence, in absence of a history (only single turn case), the only past features to be aggregated are the random initialization features.

Qualitative results - In addition to the quantitative experiments described above, we also present some qualitative final image retrieval results of our evaluated models on both the multi-turn datasets. The first set of results are shown in Figures 4a and 4b for the Multi-turn FashionIQ [60] and the Multi-turn Shoes datasets, respectively[2]. We compare the top-5 predicted results from our proposed model with two other multi-turn baselines, `ST` + `EWMA` and `ST` + `LSTM`. As seen in the figures, our approach can correctly predict the desired target image for both datasets with higher ranks as compared to other baselines. More specifically, we see that in Figure 4a, even though all the three multi-turn models can infer the general sense of desired attributes, such as "is longer" in the left block and "has short sleeves" in the right, only our model can capture complex and detailed properties, *e.g.*, "gray and not red" in the left block, and "red color with center logo" in the middle block. Furthermore, our model can retrieve multiple desirable products, as seen by the first four images in the middle block, and four out of five images in the right block. Similar results are observed for Shoes in Figure 4b, where our model correctly predicts

[2]Please refer to the supplementary material for additional results, along with the differences in annotation quality between the two datasets.

Figure 5: Memory retention capability of our proposed approach. Given an initial query image, we take two interactive user feedbacks in turns. On the left side, we have the single-turn model which only learns to retrieve an image using a single dialog exchange. As a result, none of the retrieved images in turn 2 are "green in color", which was desired in turn 1. In contrast, our proposed approach on the right can learn to retain data from both the turns, and therefore retrieves desirable product in 2 out of 3 cases.

the desired target as rank-1 in two out of three examples, whereas the other models fail to retrieve meaningful results.

4.5. Model Analysis in Interactive Settings

In addition to the results above for the static dataset, we also performed some interactive experiments to evaluate whether our model can adapt to real-world dynamic use cases beyond the trained datasets.

Memory retention - In this experiment, we demonstrate the memory retention capability of our proposed model by comparing it with a single-turn network, which does not retain historical information. We start with an initial query image retrieved via a single-turn model. Subsequently, we take user input for the next two turns to retrieve newer sets of images. As seen in Figure 5, for the single-turn model (left side), none of the retrieved images in turn 2 are *green* in color. This is because the "green in color" attribute was a desired property in turn 1, which the model without memory could not recover. In contrast, for our memory network approach (right side) both the top-2 retrieved images for turn 2 are "green in color" in addition to having "a solid color and small image". Thus, our model can learn to retain information from previous turns.

Agnosticity to turn order - Ideally, a deployed multi-turn image retrieval system should be independent of the order of feedbacks provided, as long as they are non-contradictory. This is demonstrated in the experiment conducted in Figure 6. In the first case, we take two text inputs as feedbacks from a user and present them to the model. In the second case, we reverse the order of the feedbacks. As shown in the figure, for the flipped case, our proposed FashionNTM model retrieves similar looking final products, even though the intermediate retrievals are very different.

It is to be noted that for both the experiments depicted in Figures 5 and 6, the feedback is taken from a dynamic user, thereby establishing the interactive capability of our proposed model beyond the training dataset.

Figure 6: Turn order independence feature of a memory network. In this experiment, we start with an initial query image, and take two *non-contradictory* user feedbacks. In one case (left), we let the model retrieve images based on the original order of the feedbacks, whilst in the other (right), we flip the order of the feedbacks. Our proposed approach adapts to presented input, and retrieves similar looking final results for both the cases, even though the intermediate outputs are quite different.

4.6. User Study

Fashion image retrieval is inherently a subjective task, where the task success heavily relies on the satisfaction of a customer. Thus, we conducted a small human-preference survey among 5 participants (not associated with the paper in any way). To each user, we showed the final top-1 retrieved image by the 3-best multi-turn models on the FashionIQ [60] dataset from Table 2 – `EWMA`, `LSTM`, and our proposed `FashionNTM`. To ensure consistency, we generated 45 queries for this study, whose results are shown in Figure 7. The y-axis shows the number of preferred retrieval results for each user. Results show that each of the 5 users preferred images retrieved by the proposed FashionNTM model, on an average 83.1% more as compared to other multi-turn methods. In the supplementary material, we include some examples of the user interface shown to

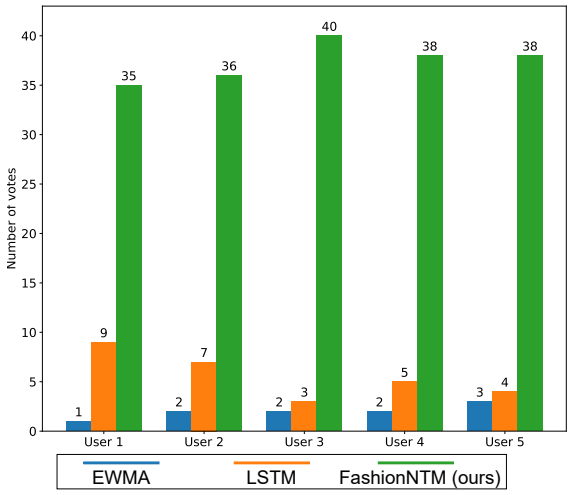

Figure 7: Human preference study of top multi-turn systems.

Table 5: Different number of memories for the proposed approach by fixing the memory size to 4×768. 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	24.5	30.5	27.5	-
FashionNTM	2	26.4	33.8	30.1	9.5
	4	**27.6**	**35.5**	**31.5**	**14.5**
	8	26.9	35.2	31.1	13.1
	16	26.4	33.8	30.1	9.5

the participants during this study.

4.7. Ablation Studies

We perform multiple ablation studies to gain insights on how changing different configurations of the memory network affect the overall performance. We perform these ablations on the multi-turn Shoes dataset as it contains more realistic and consistent feedback texts.[3]

Number of memories in CM-NTM - This experiment involves varying the number of memories C in our cascaded memory architecture. In Table 5, we observe that the cascaded memory CM-NTM models perform significantly better than vanilla NTM, which has only one memory. We hypothesize that having inputs from multiple turns interacting with the same memory could eventually lead to saturation as we get additional data which could be alleviated if there are multiple memories in the network to recover the past which might help in capturing multiple complex relationships in multi-turn interactions better. Additionally, multiple memory networks can help in learning diverse representations of the input using derived features, which is not possible with a single memory network. As seen in Table 5 for Shoes dataset, the performance increases with the number of memories, peaking at $C = 4$, and then gradually decreases as the model starts to overfit.

Inference time with multiple memories - In this experiment, we study the performance of FashionNTM in terms of the mean of R@5 and R@8 recall rates along with the time taken to process one multi-turn transaction. For each C, we evaluate the performance across four different memory sizes. As seen from Figure 8, the inference time increases with additional 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.

5. Conclusion and Future Work

In this paper, we presented a novel cascaded Neural Turing Machine-based approach, called FashionNTM, for multi-turn feedback-based fashion image retrieval. Multiple

[3]For a similar study on the Multi-turn Fashion-IQ dataset, please refer to the supplementary material.

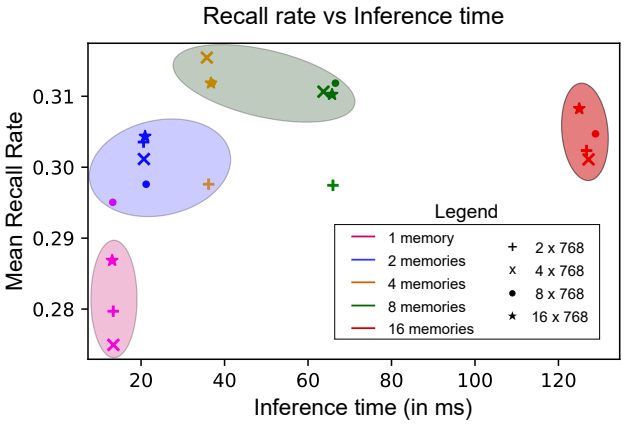

Figure 8: Scatter plot showing recall versus inference time for multiple memories in our CM-NTM for the Multi-turn Shoes dataset. The configurations belonging to the green cluster give the best recall overall, while having a high inference time. The configurations in the blue cluster provide a suitable 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.

memories in our model allow it to effectively retain and recall a number of complex relationships across transactions in the multi-turn setting, and multiple controllers help in assigning relative importance to each feature stored in the memory. This aids in attending to different parts of the input at different levels, thus leading to better performance across datasets. We also performed extensive experiments to compare our performance with baselines and previous state-of-the-art and observed that our multi-memory model significantly outperforms previous works [60], with up to 50.5% relative improvement on Multi-turn FashionIQ, and by 12.6% on the multi-turn Shoes dataset, which we created in this work. We further demonstrated that our model can generalize beyond the trained setting to dynamically interact with real-world users to retrieve meaningful final product images. Finally, a user preference study reveals that our model is preferred by human participants on an average 83.1% more as compared to other multi-turn methods.

Despite promising results, there are a few limitations that make multi-turn image retrieval a hard problem to solve. Firstly, there is dearth of high quality and diverse multi-turn image retrieval datasets in the fashion domain which hinders comprehensive studies in this field. Additionally, deploying current approaches to real-world scenarios (*e.g.*, virtual private assistants) becomes a challenge due to computational requirements. Lastly, selecting the right configuration for different components such as memory size, number of memories, *etc*. in a memory-based network is not a trivial task and depends a lot on the use case. Nevertheless, for future work, our approach could be further extended to non-fashion domains where multi-turn feedback-based information retrieval solutions are required.

References

[1] Moving Average. https://en.wikipedia.org/wiki/Movingaverage. 6

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 1

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4955–4964, 2022. 2

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21466–21474, June 2022. 2

[5] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 1, 2, 5, 6

[6] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *European Conference on Computer Vision*, pages 136–152. Springer, 2020. 2

[7] 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 (CVPR)*, June 2020. 1, 2

[8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 3

[9] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 2

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 1

[11] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19119–19128, June 2022. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 3

[14] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14105–14115, June 2022. 1, 2, 4, 5, 6

[15] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 1

[16] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2, 4, 5

[17] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018. 1, 2, 5, 6

[18] 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 *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 5, 6

[21] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE, 2019. 2

[22] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3596–3605, 2020. 2

[23] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *The International Conference on Computer Vision (ICCV)*, October 2021. 1, 2

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[25] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008. 1

[26] Haozhe Ji, Rongsheng Zhang, Zhenyu Yang, Zhipeng Hu, and Minlie Huang. Lamemo: Language modeling with look-ahead memory. *arXiv preprint arXiv:2204.07341*, 2022. 3

[27] Michael I Jordan. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier, 1997. 2

[28] Clement Joudet. ntm - neural turing machines in pytorch. https://github.com/clemkoa/ntm, 2019. 5

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[30] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues. *arXiv preprint arXiv:2010.10095*, 2020. 2

[31] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, June 2021. 1, 2, 4

[32] Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. Text-based image retrieval using progressive multi-instance learning. In *2011 international conference on computer vision*, pages 2049–2055. IEEE, 2011. 1

[33] Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. *arXiv preprint arXiv:1908.08191*, 2019. 2

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[35] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1

[36] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 2

[37] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *European Conference on Computer Vision*, pages 661–679. Springer, 2020. 3

[38] Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. ∞-former: Infinite memory transformer, 2021. 3

[39] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018. 1

[40] Suvir Mirchandani et al. FaD-VLP: Fashion vision-and-language pre-training towards unified retrieval and captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10484–10497, Dec. 2022. 2

[41] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 1

[42] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video dialog as conversation about objects living in space-time. *arXiv preprint arXiv:2207.03656*, 2022. 2

[43] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision – ECCV 2016*, pages 3–20, 2016. 1

[44] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *Proceedings of the european conference on computer vision (eccv)*, pages 751–767, 2018. 1

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[46] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 2

[47] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12155–12164, 2022. 3

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[49] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 2

[50] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision*, pages 629–645. Springer, 2020. 3

[51] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015. 2, 3

[52] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 1

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2, 6

[54] 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. 2, 4

[55] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2, 3

[56] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317, June 2021. 1, 3, 5

[57] Qingyang Wu, Zhenzhong Lan, Jing Gu, and Zhou Yu. Memformer: The memory-augmented transformer. *arXiv preprint arXiv:2010.06891*, 2020. 3

[58] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, 2021. 3

[59] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *ECCV*, 2020. 2

[60] Yifei Yuan and Wai 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, 2, 3, 5, 6, 7, 8, 9

[61] Chen Zhang, Joyce Y Chai, and Rong Jin. User term feedback in interactive text-based image retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, 2005. 1

[62] Ruiyi Zhang, Tong Yu, Yilin Shen, and Hongxia Jin. Text-based interactive recommendation via offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11694–11702, 2022. 3

[63] Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, and Changyou Chen. Text-based interactive recommendation via constraint-augmented reinforcement learning. *Advances in neural information processing systems*, 32, 2019. 1, 3