Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation

Yaxiong Wu
University of Glasgow
Glasgow, UK
y.wu.4@research.gla.ac.uk

Craig Macdonald
University of Glasgow
Glasgow, UK
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, UK
iadh.ounis@glasgow.ac.uk

ABSTRACT

Multi-modal interactive recommendation is a type of task that allows users to receive visual recommendations and express natural-language feedback about the recommended items across multiple iterations of interactions. However, such multi-modal dialog sequences (i.e. turns consisting of the system's visual recommendations and the user's natural-language feedback) make it challenging to correctly incorporate the users' preferences across multiple turns. Indeed, the existing formulations of interactive recommender systems suffer from their inability to capture the multi-modal sequential dependencies of textual feedback and visual recommendations because of their use of recurrent neural network-based (i.e., RNN-based) or transformer-based models. To alleviate the multi-modal sequential dependency issue, we propose a novel multi-modal recurrent attention network (MMRAN) model to effectively incorporate the users' preferences over the long visual dialog sequences of the users' natural-language feedback and the system's visual recommendations. Specifically, we leverage a gated recurrent network (GRN) with a feedback gate to separately process the textual and visual representations of natural-language feedback and visual recommendations into hidden states (i.e. representations of the past interactions) for multi-modal sequence combination. In addition, we apply a multi-head attention network (MAN) to refine the hidden states generated by the GRN and to further enhance the model's ability in dynamic state tracking. Following previous work, we conduct extensive experiments on the Fashion IQ Dresses, Shirts, and Tops & Tees datasets to assess the effectiveness of our proposed model by using a vision-language transformer-based user simulator as a surrogate for real human users. Our results show that our proposed MMRAN model can significantly outperform several existing state-of-the-art baseline models.

CCS CONCEPTS

• Information systems → Recommender systems; **• Computing methodologies → Learning from critiques**.

KEYWORDS

multi-modal, interactive recommendation, dialog state tracking

ACM Reference Format:
Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3523227.3546774

1 INTRODUCTION

Interactive recommendation is a recently emerging research area [5, 9, 18, 22, 27]. It aims to satisfy the users' needs by incorporating their dynamic preferences through multi-turn interactions using either only natural language for conversational recommendation [17, 23, 30, 32, 41] or, typically, both vision and natural language for multi-modal interactive recommendation [4, 12, 39, 40, 42–44, 46]. Such a multi-modal interactive recommendation is specifically concerned with a goal-oriented multi-modal sequence of interactions between users and the recommender system, where users can receive visual recommendations (i.e. the items' images) and express fine-grained natural-language critiques about the recommendations based on their preferences [5, 12, 39, 40, 45]. Figure 1 illustrates an example multi-modal interactive recommendation scenario. In particular, both the visual recommendation a_{t-1} and the corresponding natural-language feedback o_t contain rich information relating to the user's current preferences a_{target}, thereby allowing the recommender system to make an improved recommendation a_t.

Such a multi-modal interactive recommendation task has been previously modelled using recurrent neural networks (RNNs, using a gated recurrent unit (GRU) [12, 40, 43] or a long short-term memory (LSTM) [46]) or using a transformer [39] as *a state tracker* for both *multi-modal sequence combination* [1, 10] (i.e. combining the users' natural-language feedback sequence and the systems' visual recommendation sequence) and *dialog state tracking* [8, 24, 32] (i.e. eliciting the users' preferences over time). However, the actual neural networks adopted as the state trackers (such as GRUs [3], LSTMs [15] or transformers [35]) are all originally designed for *single-modal* sequence modelling tasks (such as natural language processing [29]). Therefore, these models typically resort to combining the textual and visual representations with a *concatenation operation* [12, 39, 40, 46], rather than processing the differing multi-modal sequence data separately.

Despite the expressiveness and complementary of visual recommendations and the corresponding natural-language feedback in multi-modal interactive recommendations, the long lengths of the dialog sequences makes it challenging to correctly incorporate the users' preferences over time, thereby resulting in a degraded satisfaction of the users' information needs with inappropriate recommendations. Indeed, the existing formulations of interactive recommender systems suffer from an inability to capture *multi-modal sequence dependencies* between the textual feedback and

RecSys '22, September 18–23, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9278-5/22/09…$15.00
https://doi.org/10.1145/3523227.3546774

visual recommendations using either the GRU/LSTM-based models [11, 12, 46] or the transformer-based model [39]. Specifically, we argue that the inability of these GRU/LSTM-based and transformer-based models at capturing such multi-modal sequence dependencies of the dialog sequences is inherently due to their limitations in *combining multi-modal sequences* or *tracking dialog states* (as we further discuss in Section 2).

In this paper, we alleviate the *multi-modal sequence dependency* issue in multi-modal dialog sequences modelling by addressing the *multi-modal sequence combination* and the *dialog state tracking*, respectively. To better combine the multi-modal dialog sequences than using a concatenation operation, we extend the traditional GRU architecture with an extra *feedback gate* (called a gate recurrent network (GRN), inspired by [7, 28]) to separately process the textual feedback and the visual items in the visual dialog sequences. To better track the users' dynamic preferences across multiple interaction turns, a multi-head attention network (MAN) is placed on top of our proposed GRN component to refine the GRN's hidden states and to further enhance the model's ability in dialog state tracking, inspired by RNN-enhanced transformers [38]. To this end, we propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation to effectively incorporate the users' preferences over time from the multi-modal dialog sequences of the users' natural-language feedback and the systems' visual recommendations. Following previous work, we train and evaluate our MMRAN model by using a vision-language transformer-based user simulator (VL-Transformer) [39], which has been previously shown to be a good surrogate for real users. Our extensive experiments conducted on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees* datasets show that our proposed MMRAN model can significantly outperform several existing state-of-the-art baseline models. The main contributions of this paper are as follows:

- We propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation. Our model separately processes the textual feedback sequences and the visual item sequences for multi-modal sequence combination, and tracks the dialog states using abstract representations of the previous interactions. We show that our proposed MMRAN model is more effective in capturing the dialog sequence information of the natural-language feedback and the visual recommendations compared to the existing baseline models.
- We propose a gated recurrent network (GRN) for extracting the hidden states of the past interactions from the natural-language feedback and the visual recommendations. Our GRN extends the traditional gated recurrent unit (GRU) with a *feedback gate* to capture the correlation between the textual feedback at the current turn and the hidden state of the previous turn.
- We deploy an advanced RNN-enhanced transformer architecture [38] for interactive recommendation, in order to effectively track the dialog states with a multi-head attention network (MAN) using the GRN's abstract representations.
- We perform extensive empirical evaluations with our proposed MMRAN model on the multi-modal interactive recommendation task, demonstrating significant improvements over the existing state-of-the-art approaches.

The remainder of this paper is structured as follows: we first discuss the limitations of the existing multi-modal interactive recommendation models in Section 2. We also review the related work and position our contributions in comparison to the existing literature in Section 3. Then we detail our proposed MMRAN model in Section 4. Afterwards we describe our experimental setup in Section 5 and report our experimental results in Section 6, respectively. Finally, we summarise our findings in Section 7.

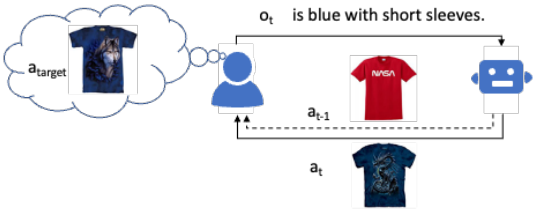


Figure 1: Example multi-modal recommendation scenario.

2 MULTI-MODAL INTERACTIVE RECOMMENDATION

In this section, we formulate the problem of the multi-modal interactive recommendation task (Section 2.1). Then, we briefly elicit the limitations of the RNN/transformer-based models in terms of the multi-modal sequence dependency issue (Section 2.2).

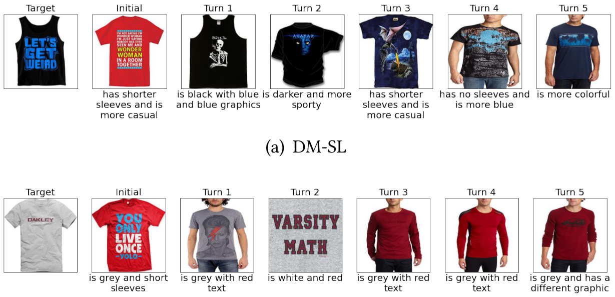
2.1 Preliminaries

We study the multi-modal interactive recommendation task by considering a user interacting with a recommender system using iterative multi-turn interactions through vision and language. At the t-th interaction turn, the recommender system presents a candidate image a_{t-1} selected from a candidate pool $\mathcal{I} = \{a_i\}_{i=0}^{N}$ to the user. The user then provides a natural language critique o_t as feedback, describing the major visual differences between the candidate image and their desired item. Specifically, we assume that the user only gives feedback on the top-ranked candidate item in the ranking list, as in [12, 40]. According to the users' natural-language feedback o_t and the interaction history up to turn t, $\tau_t = (o_{\leq t}, a_{<t}) \in \mathcal{H}$ (i.e. a set of interaction history), where $o_{\leq t} = (o_1, ..., o_t) \in \mathcal{O}$ (i.e. a set of the users' natural-language feedback) and $a_{<t} = (a_0, ..., a_{t-1}) \in \mathcal{A}$ (i.e. a set of items for recommendation), the recommender system selects another candidate image a_t from the candidate image pool. This vision-language interaction process continues until the user's desired target image a_{target} is recommended or a maximum number of interaction turns, M, is reached, leaving the user unsatisfied.

2.2 Multi-Modal Interactive Models

Figure 2 shows examples of interactive recommendations obtained from (a) the Dialog Manager (DM) [12] model, which is based on a gated recurrent unit (GRU) with a supervised-learning setup (which we denote as DM-SL); and (b) the multi-modal interactive transformer (MMIT) [39] model based on a transformer. In each example, the recommender gives a random initial recommendation (denoted "Initial") to the user, while the user with a desired target item (denoted "Target") provides natural-language feedback about the recommendation at each turn. Then, the recommender system updates the ranking list of the candidate items for the next recommendation according to the user's feedback [40].

2.2.1 The GRU/LSTM-based Models. In the GRU/LSTM-based models [12, 40, 46] for multi-modal interactive recommendation,


Target
Initial
Turn 1
Turn 2
Turn 3
Turn 4
Turn 5
has shorter sleeves and is more casual
is black with blue and blue graphics
is darker and more sporty
has shorter sleeves and is more casual
has no sleeves and is more blue
is more colorful
Target
Initial
Turn 1
Turn 2
Turn 3
Turn 4
Turn 5
is grey and short sleeves
is grey with red text
is white and red
is grey with red text
is grey with red text
is grey and has a different graphic

(a) DM-SL

(b) MMIT

Figure 2: Examples of multi-modal interactive recommendations obtained from (a) DM-SL [12] and (b) MMIT [39].

due to the inability of GRUs/LSTMs in processing different multi-modal data separately, the representations of the users' natural-language feedback and the systems' visual recommendations are usually combined with a concatenation operation and a multilayer perceptron (MLP), so as to form a single input of the GRUs/LSTMs at each turn. Such a concatenation operation on the multi-modal sequence data causes the combined textual and visual representations to be memorised or forgotten synchronously at each interaction turn in the GRUs/LSTMs-based state trackers [11, 12, 46] (**Limitation 1**). For instance, in Figure 2 (a), a sleeveless shirt is recommended at the 1st turn by the DM-SL model due to the initial comment, "shorter sleeves", compared to the red T-shirt shown at the initial turn, while the sleeveless feature shown in the image at the 1st turn and similarly conveyed by the natural-language at the initial turn is omitted by the recommender for the following recommendations. However, we argue that the users' feedback should have more effect on the hidden state of the GRUs/LSTMs in addition to the combined textual and visual representations, in that the natural-language feedback explicitly conveys the users' information needs while the rejected visual recommendations can be noisy by also containing the users' undesired features.

2.2.2 *The Transformer-based Model.*

In the transformer-based model for multi-modal interactive recommendation [39], the representations of the users' natural-language feedback and the systems' visual recommendations at all turns are *concatenated* together, while the dialog states (i.e. the estimated users' dynamic preferences) are directly tracked and inferred from all the concatenated textual and visual representations. Although the textual and visual representations at all turns in a multi-modal dialog sequence can fully interact with each other by using a multi-head attention mechanism [35] in the transformers, we argue that the effectiveness of the transformer-based model is limited as it cannot consider the previous inferred hidden states in an iterative manner (i.e. abstract representations of the past interactions) of the multi-modal dialog sequence as performed by the GRU/LSTM models at each turn (**Limitation 2**). For instance, in Figure 2 (b), for the MMIT model, the "red" colour in the comment at the 2nd turn refers to "red text" in the comment at the 1st turn, while it is misunderstood by the recommender system and taken as the colour of the shirt according to the successively recommended "red" shirts from the 3rd turn to the 5th turn.

2.2.3 *Summary of Limitations.*

To conclude, in the above analysis, we have identified two limitations of the existing GRU/LSTM-based and transformer-based models:

Limitation 1: The GRU/LSTM-based models incorporate the multi-modal data with a concatenation operation rather than processing the multi-modal dialog sequences separately for *multi-modal sequence combination.*

Limitation 2: The transformer-based models directly infer the users' preferences from all the concatenated textual and visual representations rather than from the abstract representations of the past interactions for *dialog state tracking*.

In summary, the existing multi-modal interactive recommendation models based on only GRUs, LSTMs or transformers are not able to properly process the multi-modal dialog sequences of the natural-language feedback and recommended visual items, which limits these models' ability to incorporate the users' preferences over time. In Section 4, we propose a model that addresses these limitations. In the next section, we detail related work in multi-modal interactive recommendation and recurrent neural models.

3 RELATED WORK

In this section, we first introduce multi-modal interactive recommendations and survey the recent related work in this area. We then introduce the gating mechanisms in RNNs. We further discuss the RNN-enhanced transformers.

Multi-Modal Interactive Recommendations. Vision and language-based communications between users and recommender systems have been commonly leveraged to incorporate the users' preferences and continuously provide them with recommendations during their multi-turn interactions [2, 12, 25, 34, 40, 46]. In particular, a multi-modal interactive recommender system (called Dialog Manager) using model-based reinforcement learning was proposed by Guo et al. [12] to allow users to give natural-language critiques on the visual recommendations. To correctly track and estimate the users' preferences over time, a gated recurrent unit (GRU) [3] was adopted in Dialog Manager for generating the users' estimated preference representations from the sequences of the textual representations (i.e. natural-language feedback) and the visual representations (i.e. recommended items). In addition, Wu et al. [40] also adopted a GRU component in their Estimator-Generator-Evaluator (EGE) model as a state tracker for eliciting the users' preferences over time. The EGE model considers the recommendation process as a partially observable Markov decision process (POMDP) that the recommender system can only obtain a partial portrayal of the users' preferences from the users' natural-language feedback at each turn. Meanwhile, Zhang et al. [46] proposed a reward-constrained recommendation (RCR) model using a long short-term memory network (LSTM) component for tracking the users' preferences and constrained-augmented reinforcement learning for mitigating the recommendations that violate the users' comments or feedback. Furthermore, a transformer-based model for multi-modal interactive recommendations (called MMIT) was proposed by Wu et al. [39] to incorporate the visual items' features, the users' natural-language feedback, and the fashion attributes. This transformer-based structure has demonstrated more flexibility in terms of included modalities compared to the RNN-based approaches [12, 39], as well as a more effective recommendation performance. As described in Section 2, these GRU/LSTM-based and transformer-based models suffer from their inability to capture multi-modal sequence dependencies, because of their limitations in either *combining multi-modal sequences* with a concatenation operation or *tracking dialog states* by inferring directly from all the concatenated textual and visual representations at all turns instead of the multi-modal abstract representations of the past interactions.

Gating Mechanisms of Recurrent Models. Traditional RNNs usually suffer from the vanishing gradient problem when processing long sequences [15]. Recurrent units such as a long short-term memory (LSTM) [15] and a gated recurrent unit (GRU) [3] are extensions of traditional RNNs, which use gating mechanisms to control the influence of a hidden state of the previous step. While the GRU and LSTM architectures can alleviate the vanishing gradient problem [3, 15], they cannot process different modalities separately at the same time. The representations of the multi-modal sequences are usually combined with a concatenation operation as a single input of the GRUs/LSTMs at each turn [12, 33, 40, 46]. Many researchers have extended the GRUs/LSTMs to incorporate contextual information associated with the sequence information, such as transition contexts (the time intervals and the geographical distances) by using time and/or spatial-based gates [28, 31, 48]. In the multi-modal interactive recommendation task, the users' natural-language feedback can be taken as contextual information associated with the visual recommendation sequences. However, to the best of our knowledge, such an approach to associate the sources of contextual information has not been investigated for multi-modal interactive recommendation to address **Limitation 1**.

RNN-Enhanced Transformers. In the transformer-based model [39] for interactive recommendation, dialog states (i.e. the estimated users' dynamic preferences) can only be directly tracked and inferred from all the concatenated textual and visual representations instead of being estimated from the abstract representations in the past interactions of the dialog sequences, as performed by the GRU/LSTM models at each turn. To alleviate such inherent limitations of the transformers in state tracking, a number of previous studies [13, 19, 21, 38] have investigated RNN-enhanced transformers for sequence modelling tasks, such as R-Transformer [38], to take the benefits from both the RNNs for abstract representations at each turn and from the transformers for the whole sequence's overall feature interactions in sequence modelling. Wang et al. [38] proposed an RNN-enhanced transformer model with a sliding window (called an R-Transformer) to benefit from the advantages of both an RNN and a transformer's multi-head attention mechanism. Three layers (i.e. RNNs with a sliding window, a multi-head attention layer, and a feedforward layer) were arranged hierarchically. In particular, the RNNs process sequences using a sliding window and generate the hidden states of the past interactions sequentially, while the multi-head attention layer captures the dialog states among the RNNs' hidden states of the previous turns, and the feedforward layer conducts non-linear feature transformation. However, these RNN-enhanced transformers have not yet been investigated for multi-modal interactive recommendation in order to address **Limitation 2**.

As discussed in Section 2, given the limitations of the GRU, LSTM and transformer-based models, we argue that the existing interactive recommendation models based on only GRUs, LSTMs or transformers are not able to properly process the multi-modal dialog sequences of natural-language feedback and recommended visual items. This limits the ability of such models in incorporating the users' preferences over time. To address **Limitation 1** & **2**, we propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation. Specifically, our model separately processes the textual feedback sequences and the visual item sequences for multi-modal sequence combination so as to address **Limitation 1**, while it tracks the dialog states with the abstract representations of the previous interactions in order to address **Limitation 2**. To the best of our knowledge, this novel structure of our MMRAN model constitutes the first work based on a multi-modal recurrent attention network in multi-modal interactive recommendations.

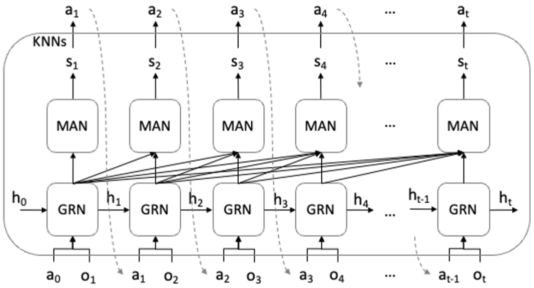


Figure 3: The multi-modal recurrent attention network (MMRAN) model.

4 THE MMRAN MODEL

We now define our proposed **M**ulti-**M**odal **R**ecurrent **A**ttention **N**etwork (MMRAN) model and introduce its components. Figure 3 shows the architecture of MMRAN, which aims to effectively incorporate the users' preferences over time. The architecture consists of three parts: text & image encoders, a gated recurrent network (GRN), and a multi-head attention network (MAN). We also describe the training of the MMRAN model using multi-turn interactions with a user simulator.

Text & Image Encoders. The text encoder (denoted $TxtEnc(\cdot)$) consists of a 1D convolutional layer (1D-CNN) and a subsequent linear layer as in [12, 40], where the user's natural-language feedback o_t (with each word represented by a one-hot vector) is extracted into a textual sentence representation $TxtEnc(o_t)$. Although there are many advanced pre-trained transformer-based language models (such as BERT [6]) for processing the natural-language feedback, we adopt a one-hot vector for each word with a pre-defined vocabulary [12, 40] of fashion-related terms when generating textual sentence representations, thus allowing fair comparisons with existing works [12, 40]. Furthermore, a pre-defined fashion vocabulary is much smaller and is more concentrated on fashion features than BERT. Similarly, the image encoder (denoted $ImgEnc(\cdot)$) consists of the ImageNet pre-trained ResNet101 model [14] and a subsequent linear layer, as in [12, 40], where a candidate image a_{t-1} is extracted into image feature representations $ImgEnc(a_{t-1})$. To simplify the notations, in the following we directly use o_t and a_{t-1} as their representations, respectively. Then, both the visual and textual representations are passed to a gated recurrent network

(GRN) and a multi-head attention network (MAN) to estimate the user's preferences.

The Gated Recurrent Network (GRN). To address **Limitation 1** and effectively incorporate the users' preferences from the multi-modal dialog sequences of the users' natural-language feedback and the recommended visual items, inspired by [7, 28], we propose a gated recurrent network (GRN) with a *feedback gate* for *multi-modal sequence combination.* Figure 4 shows the architecture of our proposed gated recurrent network (GRN). Our GRN extends the traditional gated recurrent unit (GRU) with an extra gate (i.e. a feedback gate β_t) to directly impose more effect on the hidden state in addition to the combined textual and visual representations, in that the natural-language feedback explicitly conveys the users' information needs. The estimated hidden states of the user's preferences can be achieved with $h_t = GRN(h_{t-1}, a_{t-1}, o_t)$. In particular, the proposed feedback gate β_t controls the influences of the current textual feedback o_t at each state as follows:

$$\beta_t = \sigma(W_{\beta,h}h_{t-1} + W_{\beta,o}o_t + b_\beta) \tag{1}$$

where $W_{\beta,h}$, $W_{\beta,o}$ and b are, respectively, the transition matrices and the corresponding bias. Our proposed feedback gate β_t aims to capture the correlation between the current textual feedback o_t and the hidden state of the previous turn h_{t-1}. The feedback gate β_t is activated in case where the natural-language feedback is less informative about the users' preferences compared to the hidden state h_{t-1}. Then, the equations of GRN with the proposed feedback gate β_t are:

$$c_t = W_{c,a}a_{t-1} + W_{c,o}o_t + b_c \tag{2}$$

$$z_t = \sigma(W_z c_t + U_z h_{t-1} + b_z) \tag{3}$$

$$r_t = \sigma(W_r c_t + U_r h_{t-1} + b_r) \tag{4}$$

$$\tilde{h}_t = \tanh(W_h c_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{5}$$

$$h_t = (1 - \beta_t) \odot o_t + [(1 - z_t)h_{t-1} + z_t\tilde{h}_t] \tag{6}$$

where c_t is an initially inferred multi-modal representation of the visual recommendation a_{t-1} and the corresponding natural-language feedback o_t. z_t, r_r are update and reset gates, respectively. $\tilde{h}_t$ is a candidate hidden state. $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and hyperbolic tangent functions, respectively. U_z, U_r and U_h are the weight matrices that capture the recurrent connections between every two adjacent hidden states h_{t-1} and h_t. $\odot$ denotes the element-wise product. W and b with subscripts are, respectively, the transition matrices and the corresponding biases. By including the natural-language feedback o_t through an aggregation operation (Equation (6)), o_t has more effect on the hidden state h_t. In addition, a sliding window with size $N_{sliding\_window}$, as in [38], can be used to limit the length of the multi-modal dialog sequences considered at each turn. We investigate its impact on the model's performance in Section 6.2.

The GRN component allows our MMRAN model to sequentially aggregate the recommendation and feedback information from the recommender system's recommendations and the user's natural-language feedback to the estimated hidden states for *multi-modal sequence combination.* These estimated hidden states can be considered as the representations of the past interactions and are used as inputs to the following multi-head attention network (MAN).

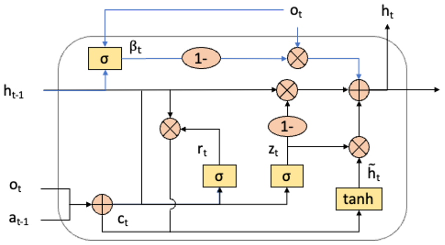


Figure 4: Our proposed GRN architecture.

The Multi-head Attention Network (MAN). To address **Limitation 2** and further track the dialog states among the GRN's hidden states of the previous turns, we adopt a multi-head attention network (MAN) architecture that enables our MMRAN model to consider the entire history of the multi-modal interactions during each interaction turn. The multi-head attention mechanism in transformers has been shown to be extremely effective to learn the long-term dependencies in the sequence modelling, since it allows a direct connection between every pair of its input representations [35, 38]. More specifically, in the multi-head attention mechanism, each input representation at each turn will attend to all the other input representations in the past interactions, thereby obtaining a set of attention scores that are used to refine its representations. In particular, the estimated hidden states of the users' preferences h_i (where $i \in [1, t]$) are further encoded with a multi-layer transformer encoder $TranEnc(\cdot)$ (with N_{layers} layers), which includes the multi-head attention mechanism (with N_{heads} attention heads). The refined hidden states are defined as follows:

$$h'_1, ..., h'_t = TranEnc(h_1, ..., h_t) \tag{7}$$

The estimated final state of the user's preferences is obtained as $s_t = Linear(ReLu(Mean(h'_1, ..., h'_t)))$. For top-$K$ candidate recommendation, the closest images to the estimated state s_t under the Euclidean distance are recommended: $a_t \sim KNNs(s_t)$, where $KNNs(\cdot)$ is a softmax distribution over the K nearest neighbours of s_t.

Overall, our MMRAN model enjoys the advantages of both the feedback gating mechanism when processing multi-modal visual dialog sequence information in the GRN (for *multi-modal sequence combination*), as well as the advantages of the multi-head attention mechanism when tracking the dialog states among the GRN's abstract representations of the users' preferences within the MAN.

Training with A User Simulator. To avoid collecting and annotating entire multi-modal conversations, which is expensive, time-consuming, and does not scale [47], we adopt an existing vision-language transformer-based user simulator (VL-Transformer) [39] as a reasonable proxy for real human users for training and evaluating our proposed MMRAN model. The user simulator considers the differences in the image features of the candidate image $a_{candidate}$ and the target image a_{target} to produce a relative caption:

$$w_{\leq i} = f([ResNet(a_{candidate}), ResNet(a_{target})]) \tag{8}$$

where $w_{\leq i} = (w_0, ..., w_i)$ is the word sequence generated for the caption (i.e. o_t), $f(\cdot)$ is the relative captioning network and $ResNet(\cdot)$ is the ImageNet pre-trained ResNet101 model [14] to obtain the prominent set of visual attributes from each image. The features of the candidate and target image pairs are concatenated to form a

Table 1: Fashion IQ datasets' statistics.

	Dresses			Shirts			Tops & Tees		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Relative Captions	11,970	4,034	-	11,976	4,076	-	12,054	3,924	-
Images	11,452	3,817	3,818	19,036	6,346	6,346	16,121	5,374	5,374

set of relative features, $[ResNet(a_{candidate}), ResNet(a_{target})]$. To train a user simulator, we follow [12, 39, 40] by using relative captioning datasets, described further in Section 5 below. Furthermore, we train our proposed MMRAN model with a triplet loss objective, $L_{triplet}$, similar to [12, 39]:

$$L_{triplet} = \max(0, ||s_t - a_+||_2 - ||s_t - a_-||_2 + m) \tag{9}$$

where a_+ is the representation of the target image as a positive sample, a_- is the representation of a randomly sampled image as a negative sample, $|| \cdot ||_2$ denotes L^2-norm, and m is a constant for the margin.

5 EXPERIMENTAL SETUP

In this section, we evaluate the effectiveness of our proposed MMRAN model in comparison to the existing approaches from the literature. In particular, to address **Limitations 1** & **2**, we answer the following three research questions:

- RQ1: Does our proposed MMRAN model outperform the existing state-of-the-art baseline models in the multi-modal interactive recommendation task with natural-language feedback?
- RQ2: Does the GRN structure address Limitation 1 and thereby improve the MMRAN models' ability to incorporate the users' preferences from the multi-modal dialog sequences?
- RQ3: Does the MAN structure address Limitation 2 so as to improve the MMRAN models' ability to effectively track dialog states?

5.1 Datasets & Measures

We perform experiments on the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets [39]. On these three datasets, both relative captions of image pairs and the images of the fashion products are available for training and testing the user simulators (i.e. relative captioners) and the recommendation models, respectively. The statistics of the Fashion IQ datasets are summarised in Table 1.

Measures. We measure the effectiveness of the multi-modal interactive recommendation models at the M-th turn interaction with top-heavy metrics, such as NDCG@N (i.e. Normalised Discounted Cumulative Gain truncated at rank $N = 10$), MRR@N (i.e. Mean Reciprocal Rank truncated at rank $N = 10$), and SR (i.e. Success Rate that is the percentage of the succeeded users among all the users with top-1 recommendation), as in [40, 46]. In particular, both NDCG@N and MRR@N measure the quality of the ranking list at each turn, while SR measures the efforts for finding the target items over multi-turn interactions. We apply all the evaluation metrics (i.e. NDCG@10, MRR@10, SR) at the 5th interaction turn for significance testing.

5.2 Baselines

We compare our proposed MMRAN model to three types of the existing state-of-the-art baselines for multi-modal interactive recommendation with different state trackers:

- **RNNs (GRUs/LSTMs)**: The Dialog Manager model [12] is a multi-modal interactive recommendation model based only on a *GRU* as the state tracker. There are two variants of the Dialog Manager model in terms of their learning approaches: Dialog Manager with a supervised-learning setup (denoted DM-SL) and Dialog Manager with a model-based reinforcement learning setup (denoted DM-RL). The DM-SL model is trained with a triplet loss (i.e. Equation (9)) to maximise the short-term rewards, while the DM-RL model is further trained with a cross entropy loss to maximise the cumulative future rewards by exploring all possible recommendation trajectories in the future turns given a known environment (i.e. a user simulator) [12]. In addition, as a further possible baseline, we envisage that an LSTM can also act as a state tracker in the Dialog Manager model with a supervised-learning setup (denoted DM-LSTM). Furthermore, we take the Estimator-Generator-Evaluator (EGE) model [40] as another GRU-based baseline model, which uses reinforcement learning with a partially observable Markov decision process (POMDP).
- **Transformers**: The multi-modal interactive transformer (MMIT) model [39] applies only a *transformer*. The MMIT model directly attends to the entire multi-modal interaction history of both the users' previous textual feedback and the system's visual recommendations. The MMIT model is also trained with a triplet loss as per DM-SL.
- **RNN-Enhanced Transformers**: R-Transformer [38], a typical RNN-enhanced transformer, can be adapted as a strong baseline model based on a GRU and a transformer. There are two variants of the R-Transformer model: R-Transformer with a window size 3 [38] (which we denote as R-T$_{Local}$) and R-Transformer without a sliding window (which we denote as R-T$_{Global}$). The R-T$_{Local}$ and R-T$_{Global}$ models are also trained with a triplet loss similar to DM-SL.

The baseline models based on RNNs (i.e. DM-LSTM, DM-SL and DM-RL) and Transformers (i.e. MMIT)) are the two representative formulations of the existing multi-modal interactive recommendation task, which are formulated as a sequential modelling problem with an RNN (such as a GRU or a LSTM) or a transformer, respectively. The RNN-enhanced transformer models (i.e. R-T$_{Local}$ and R-T$_{Global}$) adapted from the literature [38] can provide stronger baselines using more advanced network structures. In addition, the GRN component in MMRAN can also be adapted as a multi-modal interactive recommendation model to estimate the users' preferences and to make recommendations independently (which we denote as MMRAN w/o MAN).

5.3 Experimental Settings

Setup for User Simulator. We first train the existing VL-Transformer user simulator [39] for relative captioning on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees* datasets, separately. The network parameters are randomly initialised. We use the Adam [20] optimiser with an initial learning rate of 10^{-4}. The batch size is 16, and the maximum number of epochs is 30. The dimensionality of the embeddings and hidden states is 512. Appendix A.1 provides a comparison of the VL-Transformer user simulator with another recent user simulator called Show Tell [37] to demonstrate how close the VL-Transfomer user simulator behaves in comparison to real human captions.

Setup for Recommender Training. Next, we train our proposed MMRAN model using the VL-Transformer user simulator trained on the *Fashion IQ Dresses, Shirts, Tops & Tees* datasets, respectively. The recommendation models' parameters are randomly initialised. We use Adam [20] with a learning rate of 10^{-3} [12, 46]. The embedding dimensionality of the feature space is set to 256 and the batch

Table 2: The multi-modal interactive recommendation effectiveness of our proposed MMRAN model and the baseline models at the 5th turn on the three used datasets. % Improv. indicates the improvements by MMRAN over the best baseline model. The best overall results are highlighted in bold. * and † denote a significant difference in terms of a paired t-test (Holm-Boferroni correction, $p < 0.05$), compared to MMRAN and MMRAN w/o MAN in each dataset, respectively.

Models	State	Dresses			Shirts			Tops & Tees		
	Tracker	NDCG@10	MRR@10	SR	NDCG@10	MRR@10	SR	NDCG@10	MRR@10	SR
DM-LSTM	LSTM	0.1788*†	0.1517*†	0.1163*†	0.0928*†	0.0771*†	0.0573*†	0.1307*†	0.1101*†	0.0835*†
DM-SL	GRU	0.2050*†	0.1756*†	0.1364*†	0.1108*†	0.0923*†	0.0680*†	0.1566*†	0.1337*†	0.1026*†
DM-RL	GRU	0.2339*	0.2047*	0.1621*	0.1274*	0.1065*	0.0798*	0.1654*†	0.1407*†	0.1077*
EGE	GRU	0.2580*	0.2245*	0.1765*	0.1398*	0.1179*	0.0888*	0.1909*	0.1618*	0.1221*
MMIT	Transformer	0.2443*	0.2135*	0.1701*	0.1278*	0.1072*	0.0790*	0.1738*	0.1468*	0.1108*
R-T$_{Local}$	R-Transformer	0.2407*	0.2099*	0.1663*	0.1232*	0.1034*	0.0777*	0.1796*	0.1536*	0.1180*
R-T$_{Global}$	R-Transformer	0.2672*	0.2320*	0.1831*	0.1402*	0.1182*	0.0884*	0.2019*	0.1703*	0.1288*
MMRAN	GRN & MAN	**0.3327**	**0.2918**	**0.2345**	**0.1683**	**0.1414**	**0.1043**	**0.2385**	**0.2041**	**0.1568**
w/o MAN	GRN	0.2477*	0.2162*	0.1751*	0.1244*	0.1058*	0.0808*	0.1823*	0.1545*	0.1178*
% Improv.	-	24.51	25.78	28.07	20.04	19.63	17.45	18.13	19.45	21.74

size to 128 following the setting in [12]. For each batch, we train the model with 10 interaction turns as in [40]. The maximum number of epochs for training is 20. For the recommendation task, early stopping [11] is used to avoid overfitting. The training completes when average NDCG@10 over all the interaction turns on the validation sets stops improving for 5 epochs, or when the maximum number of training epochs is reached. For our proposed MMRAN model, we consider all the previous textual feedback and visual recommendations at each turn.

Setup for Recommender Evaluation. We evaluate the interactive recommendation models for top-K (i.e. $K = 1$) recommendation with multi-turn interactions $M \in [1, 5]$ on the above three datasets, respectively. The previously recommended items are removed from the ranking list at each turn with a post-filter to avoid repeated recommendations, as in [40]. For a fair comparison, we mainly compare the effectiveness of the tested models at the 5th turn (i.e. $M = 5$) using the paired t-test (applying Holm-Bonferroni for multiple comparison correction [16]). When a user successfully finds the target item in less than 5 turns, we consider the ranking metrics (i.e. NDCG@10 and MRR@10) for that user to be equal to one for all turns thereafter.

6 EXPERIMENTAL RESULTS

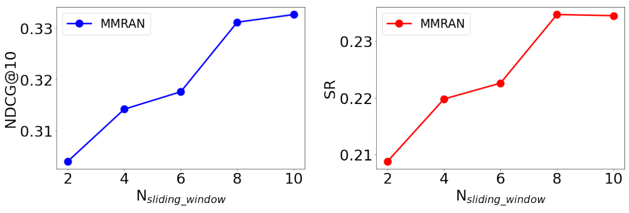
We now analyse the experimental results to answer the three research questions that are stated in Section 5, concerning the effectiveness of our proposed MMRAN model for multi-modal interactive recommendations with natural-language feedback (Section 6.1), the impact of the GRN structure for multi-modal sequence combination (Section 6.2) and the impact of the MAN structure for dialog state tracking (Section 6.3). We also show a use case from the logged experimental results to consolidate our findings (Section 6.4).

6.1 MMRAN vs. Baselines (RQ1)

To answer RQ1, we assess the effectiveness of our MMRAN model by comparing them with seven strong recommendation approaches in the literature. Table 2 shows the obtained recommendation performances of the baseline models (i.e. DM-LSTM, DM-SL, DM-RL, EGE, MMIT, R-T$_{Local}$ and R-T$_{Global}$ in the first part) as well as the MMRAN model variants (in the second part) with the same test sets of the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets at the 5th interaction turn. The best overall performances across the three groups of columns in the table are highlighted in bold in Table 2. In each group, * and † denote, respectively, significant differences compared to MMRAN and MMRAN w/o MAN (i.e. GRN only), in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$) [16]. Comparing the results in the table, we observe that our proposed MMRAN model achieves better performances of 24-28%, 17-20% and 18-22% at the 5th turn than the best baseline model across all metrics on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees*, respectively. Indeed, our proposed MMRAN model is significantly better than DM-LSTM, DM-SL, DM-RL, EGE, MMIT, R-T$_{Local}$ and R-T$_{Global}$ for each metric at the 5th turn with top-1 recommendation. In answer to RQ1, the results demonstrate that our proposed MMRAN model does overall outperform the previous state-of-the-art baseline models. In particular, it is significantly more effective than all of the GRU/LSTM-based models (i.e. DM-LSTM, DM-SL, DM-RL and EGE), the transformer-based model (i.e. MMIT), and the RNN-enhanced transformer models (i.e. R-T$_{Local}$ and R-T$_{Global}$). Therefore, these results demonstrate that our proposed MMRAN model, with the multi-modal recurrent attention network, can effectively incorporate the users' preferences over time.

6.2 Impact of GRN (RQ2)

To address RQ2, the second part of Table 2 examines the comparative performances of the MMRAN and MMRAN w/o MAN models (i.e. GRN only) with different components for tracking and estimating the users' preferences (i.e. state trackers). First, focusing on GRN, we observe that the MMRAN w/o MAN model performs significantly better than the DM-LSTM and DM-SL models in terms of all metrics on the three datasets. The significantly better performance of the GRN component indicates that the extra feedback gate can enhance the GRU's ability in combining the multi-modal sequences (i.e. textual feedback sequences and visual recommendation sequences). In addition, we also observe that MMRAN with a GRN component in the state tracker performs significantly better



(a) NDCG@10

(b) SR

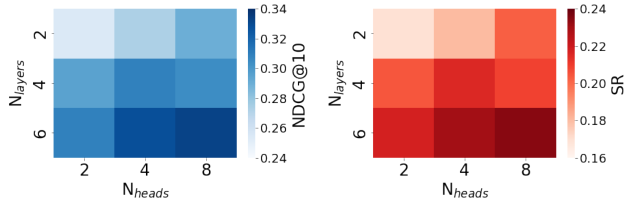
Figure 5: Effects of the sliding window size $N_{sliding\_window}$ over the multi-modal dialog sequences on our proposed MMRAN at the 5th turn on *Fashion IQ Dresses*.

than the R-T$_{Local}$ and R-T$_{Global}$ baselines, which all use a GRU component in the state tracker. This suggests, as we argued in Section 1, that imposing more effect of the users' natural-language feedback on the hidden state of the GRUs in addtion to the combined textual and visual representations can benefit the interactive recommendation model. Furthermore, Figure 5 illustrates the NDCG@10 and SR performances of MMRAN at the 5th turn in top-1 recommendation on the *Fashion IQ Dresses* dataset with different sliding window sizes $N_{sliding\_window}$ that limit the lengths of the multi-modal dialog sequences at each turn. We can see that the performance of MMRAN improves when the value of the sliding window size $N_{sliding\_window}$ increases from 2 to 10, except for $N_{sliding\_window} = 10$ in term of SR. Further ablation studies on the *Fashion IQ Shirts* and *Tops & Tees* datasets also led to similar results and observations. We omit their reporting in this paper because of space constraints.

Overall, for RQ2, we conclude that the GRN component with a natural-language feedback gating mechanism enhances the model's ability to combine the multi-modal sequences so as to address **Limitation 1**, thereby better incorporating the users' information needs from the multi-modal dialog sequences than the traditional GRU network.

6.3 Impact of MAN (RQ3)

To address RQ3, Figure 6 depicts the effects of the MAN's layers N_{layers} and the MAN's attention heads N_{heads} on our proposed MMRAN in terms of NDCG@10 and SR at the 5th turn on *Fashion IQ Dresses*. We can see that the performance of MMRAN improves when the MAN's layers N_{layers} and the MAN's attention heads N_{heads} increase from 2, respectively, except for $N_{layers} = 4$ and N_{heads} from 4 to 8. Meanwhile, our proposed MMRAN model can achieve the best performance with $N_{layers} = 6$ and $N_{heads} = 8$ as in [39]. Moreover, we note that the MMRAN model with both the GRN and MAN components significantly outperforms MMRAN w/o MAN, suggesting that the additional MAN component with the multi-head attention mechanism further refines the hidden states (which are generated by the previous GRN component) by tracking the dialog states of the users' preferences among the multi-modal dialog sequences. The better effectiveness of MMRAN with both GRN and MAN added up suggests that MMRAN can benefit from their joint combination. Furthermore, we also observe that the MMRAN model significantly outperforms the transformer-based MMIT model, suggesting that adopting the hidden states of GRN as the representations of the past interactions is more effective than using the original textual and visual representations as the inputs of the transformer's multi-head attention.



(a) NDCG@10

(b) SR

Figure 6: Effects of the MAN's layers N_{layers} and the MAN's attention heads N_{heads} on our proposed MMRAN model at the 5th turn on *Fashion IQ Dresses*.

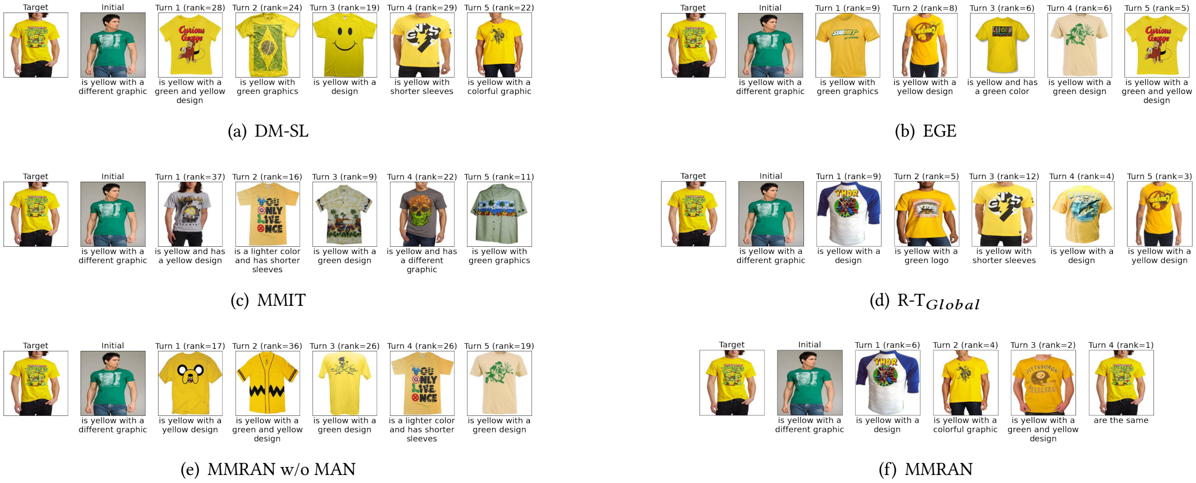
Overall, for RQ3, we conclude that the MAN component of MMRAN allows to effectively track the users' preferences with the GRN's abstract representations of the multi-modal dialog sequences, while addressing **Limitation 2**.

6.4 A Use Case

To consolidate the results observed in the paper, we present a use case of multi-modal interactive recommendation in Figure 7 for the *Fashion IQ Shirts*. Figure 7 shows the interaction process for top-1 recommendation across the six tested models: (a) DM-SL, (b) EGE, (c) MMIT, (d) R-T$_{Global}$, (e) MMRAN w/o MAN and (f) MMRAN. For a fair comparison, the initial images are the same across the tested models given the target image from the testing set. When the target item is recommended, the rank is 1 (e.g. "Turn 1 (rank=1)" in Figure 7 (f)), and the user simulators will give the comment: "are the same". We observe that our proposed MMRAN model is the most effective among the tested models. In particular, DM-SL/EGE with only a GRU, MMIT with only a transformer and R-T$_{Global}$ based on a GRU and a transformer all fail to recommend the target item within 5 interaction turns, while our proposed MMRAN model needs only 4 interaction turns to recommend the target item. Furthermore, we also observe that the rank of the target item with the MMRAN w/o MAN model is relatively higher than the rank of DM-SL at the 5th interaction turn. In addition, both the MMRAN and MMRAN w/o MAN models can generally maintain a reasonable recommendation during the multi-turn interaction process with a shirt that is "yellow with a design". Though the recommendation with MMRAN at the 1st turn is not a really "yellow" shirt, it contains features from both the initial recommendation (i.e. the "green" colour) and the corresponding natural-language feedback (i.e. the "yellow" colour) in its "different graphic", while maintaining the highest rank of the target item (i.e. "rank=6") among all the tested models at the 1st turn. Indeed, the MMRAN model can better capture the "green and yellow design" features from the users' feedback than the other tested models. Similar results and observations were seen for the *Fashion IQ Dresses* and *Tops & Tees* datasets, but are omitted for reasons of space.

7 CONCLUSIONS

In this paper, we proposed a novel multi-modal recurrent attention network (MMRAN) model for multi-modal interactive recommendation to effectively incorporate the users' preferences over time. Specifically, we leveraged a gated recurrent network


Target
Initial
Turn 1 (rank=28)
Turn 2 (rank=24)
Turn 3 (rank=19)
Turn 4 (rank=29)
Turn 5 (rank=22)
is yellow with a different graphic
is yellow with a green and yellow design
is yellow with green graphics
is yellow with a design
is yellow with shorter sleeves
is yellow with a colorful graphic
Target
Initial
Turn 1 (rank=9)
Turn 2 (rank=8)
Turn 3 (rank=6)
Turn 4 (rank=6)
Turn 5 (rank=5)
is yellow with a different graphic
is yellow with green graphics
is yellow with a yellow design
is yellow and has a green color
is yellow with a green design
is yellow with a green and yellow design
Target
Initial
Turn 1 (rank=37)
Turn 2 (rank=16)
Turn 3 (rank=9)
Turn 4 (rank=22)
Turn 5 (rank=11)
is yellow with a different graphic
is yellow and has a yellow design
is a lighter color and has shorter sleeves
is yellow with a green design
is yellow and has a different graphic
is yellow with green graphics
Target
Initial
Turn 1 (rank=9)
Turn 2 (rank=5)
Turn 3 (rank=12)
Turn 4 (rank=4)
Turn 5 (rank=3)
is yellow with a different graphic
is yellow with a design
is yellow with a green logo
is yellow with shorter sleeves
is yellow with a design
is yellow with a yellow design
Target
Initial
Turn 1 (rank=17)
Turn 2 (rank=36)
Turn 3 (rank=26)
Turn 4 (rank=26)
Turn 5 (rank=19)
is yellow with a different graphic
is yellow with a yellow design
is yellow with a green and yellow design
is yellow with a green design
is a lighter color and has shorter sleeves
is yellow with a green design
Target
Initial
Turn 1 (rank=6)
Turn 2 (rank=4)
Turn 3 (rank=2)
Turn 4 (rank=1)
is yellow with a different graphic
is yellow with a design
is yellow with a colorful graphic
is yellow with a green and yellow design
are the same

(a) DM-SL

(b) EGE

(c) MMIT

(d) R-T$_{Global}$

(e) MMRAN w/o MAN

(f) MMRAN

Figure 7: A use case for multi-modal interactive recommendation with different models on *Fashion IQ Shirts*.

(GRN) with a feedback gate to separately process the natural-language feedback and visual recommendations into hidden states (i.e. representations of the past interactions) for multi-modal sequence combination, as well as a multi-head attention network (MAN) to refine the previously generated hidden states by the GRN component to further track the dialog states of the users' preferences. Following previous work, we trained our MMRAN model by using a vision-language transformer-based user simulator (VL-Transformer), which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Our experiments on three *Fashion IQ* datasets demonstrated that our proposed MMRAN model achieves significantly enhanced performances compared to the strongest baseline models on each used dataset - for instance, improvements of 24-28%, 17-20% and 18-22%, respectively. Our reported results showed that the MMRAN model benefits from the capability of GRN in combining multi-modal dialog sequences and from the MAN's structure to effectively track the dialog states.

A APPENDIX

A.1 User Simulator Comparison

To demonstrate how close the VL-Transfomer user simulator behaves in comparison to real human captions, we provide a quantitative analysis of the VL-Transformer user simulator for relative image captioning. We evaluate the relative captioning models (i.e. user simulators) on the validation set due to the fact that the test sets for relative captioning are not released in the *Fashion IQ* datasets. Table A1 shows the relative captioning effectiveness of the VL-Transformer model and another existing state-of-the-art baseline user simulator model, Show Tell [12, 37], for generating natural-language critiques given a pair of images. Effectiveness is measured in terms of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [26] and Consensus-based Image Description Evaluation (CIDEr) [36], on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees* datasets. The best overall performing results for each dataset are highlighted in bold in Table A1. Comparing the results in the table, we observe that, overall, the VL-Transformer model achieves better performances than the LSTM-based model (i.e. Show Tell) across all metrics on all the *Fashion IQ* datasets.

Table A1: The relative captioning effectiveness of the VL-Transformer relative captioning model compared to the Show Tell baseline model on the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets. The best results for each dataset and measure are in bold. * denotes a significant difference compared to the VL-Transformer in terms of a paired t-test ($p < 0.05$).

Simulators	Dresses		Shirts		Tops & Tees	
	ROUGE	CIDEr	ROUGE	CIDEr	ROUGE	CIDEr
Show Tell	0.3105*	0.5165*	0.3030*	0.5640*	0.3074*	0.5801*
VL-Transformer	**0.3225**	**0.6346**	**0.3198**	**0.6489**	**0.3266**	**0.7006**

Figure A1 presents an example of the generated natural-language critiques given a target image and a candidate image on each dataset: (a) Dresses, (b) Shirts, and (c) Tops & Tees. There are two shown ground truths[1] (i.e. GT-1 and GT-2) for each pair of images, each followed by the generated captions by Show Tell and VL-Transformer. From the generated captions on each dataset, it can be observed that the relative caption generated by the VL-Transformer model is more expressive and more close to the ground truths compared to the other model. These results demonstrate that the VL-Transformer user simulator can generate expressive natural-language feedback via relative captioning that is close to the ground truths. Therefore, the use of the VL-Transformer for relative captioning can act as a reasonable surrogate for real human users in generating natural-language feedback.

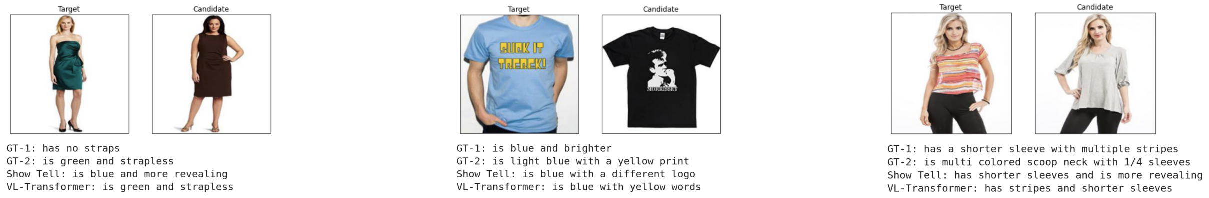
ACKNOWLEDGMENTS

The authors acknowledge support from EPSRC grant EP/R018634/1 entitled Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics.

REFERENCES

[1] Rory Beard, Ritwik Das, Raymond WM Ng, PG Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. 2018. Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proc. CoNLL*. 251–259.

[1] https://github.com/XiaoxiaoGuo/fashion-iq



```
GT-1: has no straps
GT-2: is green and strapless
Show Tell: is blue and more revealing
VL-Transformer: is green and strapless
```

(a) Dresses

```
GT-1: is blue and brighter
GT-2: is light blue with a yellow print
Show Tell: is blue with a different logo
VL-Transformer: is blue with yellow words
```

(b) Shirts

```
GT-1: has a shorter sleeve with multiple stripes
GT-2: is multi colored scoop neck with 1/4 sleeves
Show Tell: has shorter sleeves and is more revealing
VL-Transformer: has stripes and shorter sleeves
```

(c) Tops & Tees

Figure A1: Examples of different user simulators.

[2] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. 2021. Ask&Confirm: Active Detail Enriching for Cross-Modal Retrieval With Partial Query. In *Proc. ICCV*. 1835–1844.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[4] Yashar Deldjoo, Fatemeh Nazary, Arnau Ramisa, Julian Mcauley, Giovanni Pellegrini, Alejandro Bellogin, and Tommaso Di Noia. 2022. A Review of Modern Fashion Recommender Systems. *arXiv preprint arXiv:2202.02757* (2022).

[5] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multi-modal conversational information seeking. In *Proc. SIGIR*. 1577–1587.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:cs.CL/1810.04805

[7] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2017. Sequential user-based recurrent neural network recommendations. In *Proc. RecSys*. 152–160.

[8] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on Conversational Recommendation Systems. In *Proc. RecSys*. 751–753.

[9] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *arXiv preprint arXiv:2101.09459* (2021).

[10] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. 2021. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion* 66 (2021), 184–197.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[12] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Proc. NeurIPS*. 678–688.

[13] Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019. Modeling recurrence for transformer. *arXiv preprint arXiv:1904.03092* (2019).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. 770–778.

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

[16] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[17] Dietmar Jannach and Li Chen. 2022. Conversational Recommendation: A Grand AI Challenge. *arXiv preprint arXiv:2203.09126* (2022).

[18] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. *arXiv preprint arXiv:2004.00646* (2020).

[19] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. 2018. Recurrent transformer networks for semantic correspondence. *arXiv preprint arXiv:1810.12155* (2018).

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. ICLR*.

[21] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402* (2020).

[22] Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *Proc. SIGIR*. 2425–2428.

[23] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proc. WSDM*. 304–312.

[24] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *Proc. SIGIR*. 675–684.

[25] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proc. MM*. 801–809.

[26] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. of ACL Workshop on text summarization branches out*. 74–81.

[27] Joao Magalhaes, Tat-Seng Chua, Tao Mei, and Alan Smeaton. 2021. The Next Generation Multimodal Conversational Search and Recommendation. In *Proc. MM*. 953–954.

[28] Jarana Manotumruksa, Craig Macdonald, and Iadh Ounis. 2018. A contextual attention recurrent architecture for context-aware venue recommendation. In *Proc. SIGIR*. 555–564.

[29] Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624.

[30] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to Ask Appropriate Questions in Conversational Recommendation. *arXiv preprint arXiv:2105.04774* (2021).

[31] Elena Smirnova and Flavian Vasile. 2017. Contextual sequence modeling for recommendation with recurrent neural networks. In *Proc. of the 2nd workshop on deep learning for recommender systems*. 2–9.

[32] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proc. SIGIR*. 235–244.

[33] Zhi-Xuan Tan, Arushi Goel, Thanh-Son Nguyen, and Desmond C Ong. 2019. A multimodal lstm for predicting listener empathic responses over time. In *Proc. FG*. 1–4.

[34] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2021. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion* 77 (2021), 149–171.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*. 5998–6008.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. CVPR*. 4566–4575.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. CVPR*. 3156–3164.

[38] Zhiwei Wang, Yao Ma, Zitao Liu, and Jiliang Tang. 2019. R-transformer: Recurrent neural network enhanced transformer. *arXiv preprint arXiv:1907.05572* (2019).

[39] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proc. CVPR*. 11307–11317.

[40] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2021. Partially Observable Reinforcement Learning for Dialog-Based Interactive Recommendation. In *Proc. RecSys*. 241–251.

[41] Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. 2021. Adapting User Preference to Online Feedback in Multi-round Conversational Recommendation. In *Proc. WSDM*. 364–372.

[42] Tong Yu, Yilin Shen, and Hongxia Jin. 2019. A visual dialog augmented interactive recommender system. In *Proc. KDD*. 157–165.

[43] Tong Yu, Yilin Shen, and Hongxia Jin. 2020. Towards Hands-Free Visual Dialog Interactive Recommendation. In *Proc. AAAI*, Vol. 34. 1137–1144.

[44] Yifei Yuan and Wai Lam. 2021. Conversational Fashion Image Retrieval via Multiturn Natural Language Feedback. *arXiv preprint arXiv:2106.04128* (2021).

[45] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *arXiv preprint arXiv:2201.08808* (2022).

[46] Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, and Changyou Chen. 2019. Text-based interactive recommendation via constraint-augmented reinforcement learning. In *Proc. NeurIPS*. 15214–15224.

[47] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proc. KDD*. 1512–1520.

[48] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2017. What to Do Next: Modeling User Behaviors by Time-LSTM. In *Proc. IJCAI*. 3602–3608.