Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback

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

Craig Macdonald, Iadh Ounis
University of Glasgow
{firstname.lastname}@research.gla.ac.uk

ABSTRACT

In a real-world shopping scenario, users can express their natural-language feedback when communicating with a shopping assistant by stating their satisfactions *positively* with "I like" or *negatively* with "I dislike" according to the quality of the suggested/recommended fashion products. A multimodal conversational recommender system (using text and images in particular) aims to replicate this process by eliciting the dynamic preferences of users from their natural-language feedback and updating the visual recommendations so as to satisfy the users' current needs through multi-turn interactions. However, the impact of positive and negative natural-language feedback on the effectiveness of multimodal conversational recommendation has not yet been fully explored. Since there are no datasets of conversational recommendation with both positive and negative natural-language feedback, the existing research on multimodal conversational recommendation imposed several constraints on the users' natural-language expressions (i.e. either only describing their preferred attributes as positive feedback or rejecting the undesired recommendations without any natural-language critiques) to simplify the multimodal conversational recommendation task. To further explore the multimodal conversational recommendation with positive and negative natural-language feedback, we investigate the effectiveness of the recent multimodal conversational recommendation models for effectively incorporating users' preferences over time from both positively and negatively natural-language oriented feedback corresponding to the visual recommendations. We also propose an approach to generate both positive and negative natural-language critiques about the recommendations within an existing user simulator. Following previous work, we train and evaluate the two existing conversational recommendation models by using the user simulator with positive and negative feedback as a surrogate for real human users. Extensive experiments conducted on a well-known fashion dataset demonstrate that positive natural-language feedback is more informative relating to the users' preferences in comparison to negative natural-language feedback.

ACM Reference Format:
Yaxiong Wu and Craig Macdonald, Iadh Ounis. 2022. Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543829.3543837

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

1 INTRODUCTION

Conversational recommendation, a type of conversational information seeking task [6, 29], is a newly emerging research area that aims to track/elicit the users' dynamic preferences and take actions (such as recommending items) according to their current needs through multi-turn interactions [9, 13, 14]. Conversations between humans are largely multimodal, including visual, textual, audio, and touch signals [6]. Multimodal conversational recommendation (using text and images in particular) is specifically concerned with a goal-oriented sequence of interactions between users and recommender systems, where the users can receive visual recommendations (i.e. items' images) and express fine-grained natural-language critiques about the recommendations in terms of their preferences [11, 23, 24]. In particular, natural-language feedback corresponding to the visual recommendations allows a conversational recommender system to obtain rich information relating to the users' current preferences, thereby leading to a suitable recommendation [11, 17, 20, 23, 28]. Figure 1 (a) shows an example of multimodal conversational recommendation with natural-language feedback [11, 23, 24] for fashion products (such as shoes). In this use case, the user gives natural-language feedback (critiques) that describe the differences between the users' preferences (i.e. the *target* item they have in mind) and the system's recommendations at each interaction turn, to obtain items with more preferred features. The conversational recommender system recommends the images of 3 items, based on the users' natural-language critiques.

Such a multimodal conversational recommendation task is close to a real-world shopping scenario, where the users generally express their natural-language feedback positively or negatively according to the quality of the recommendations when communicating/interacting with the shopping assistants (who may recommend items). In particular, the users might be asked to state their satisfactions using the sentences with "I like" for positive feedback or "I dislike" for negative feedback. Figure 1 (b) demonstrates an example of both positive and negative natural-language feedback in the multimodal conversational recommendation task. The recommender system is expected to update visual recommendations with more preferred features and to avoid recommendations with undesired features according to the users' positive and/or negative natural-language feedback.

Despite the expressiveness of natural-language feedback in conversational recommendation, the impact of positive and negative natural-language feedback on the effectiveness of multimodal conversational recommendation has not yet been fully explored. Due to the lack of multimodal conversations with both positive and negative natural-language critiques about the visual recommendations in terms of the users' preferences, the existing research on multimodal conversational recommendation imposed several constraints

on the users' natural-language expressions, in order to simplify the multimodal conversational recommendation task. For instance, the users are assumed to either only describe their preferred attributes as positive feedback [11, 19, 23, 24, 26, 27, 32] or just reject the undesired item-level recommendations without any natural-language critiques [2, 16, 25] during the multi-turn interactions. To learn satisfactory recommender systems with enough training data, *user simulators* have been used as surrogates for real human users in the optimisation and evaluation processes [11, 16, 23]. In particular, Guo et al. [11] proposed a user simulator with only positive natural-language feedback for relative captioning [18]. Meanwhile, Lei et al. [16] formulated the conversational recommendation task as answering the questions about the attributes and the recommended items with a binary yes/no response.

In this paper, we investigate the effectiveness of the recent multimodal conversational recommendation models for effectively incorporating the users' preferences over time from positively and/or negatively natural-language oriented feedback corresponding to the visual recommendations. To make the conversational recommendation task more realistic by supporting both positive and negative natural-language feedback, we propose an approach to generate both positive and negative natural-language critiques about the recommendations with an existing user simulator for relative captioning [11]. Such a user simulator can act as a reasonable surrogate for real human users in the optimisation and evaluation processes, as in [11, 23, 30]. Following previous work, we train and evaluate the two existing multimodal conversational recommendation models (i.e. Dialog Manager (DM) [11] and Multimodal Interactive Transformer (MIT) [23]) by using the user simulator with positive and negative feedback as a surrogate for real human users. Extensive experiments conducted on a well-known fashion dataset demonstrate that positive feedback is more informative relating to the users' preferences in comparison to negative feedback. The main contributions of this paper are summarised as follows:

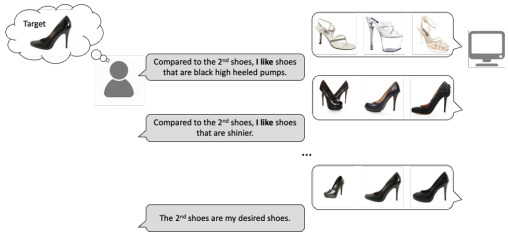
- We first investigate the effectiveness of the multimodal conversational recommendation models with both positive and negative natural-language feedback. Different from the previous work relating to positive and negative feedback, the users are assumed to actively express their satisfactions positively with "I like" or negatively with "I dislike" according to the quality of the recommendations, rather than answering questions passively with "yes" or "no".
- We propose an approach to generate both positive and negative natural-language feedback with a user simulator for relative captioning, which enables our research with various combinations of positive and negative natural-language sentences.
- We investigate the impact of different textual encoding mechanisms (i.e. pre-trained contextual embeddings [7] and one-hot embeddings) on the effectiveness of the multimodal conversational recommendation models.
- Extensive empirical evaluations are performed on the multimodal recommendation task, demonstrating different levels of difficulties for incorporating the users' preferences from positive and negative feedback over existing state-of-the-art approaches while providing directions for future work.

The remainder of the paper is organised as follows: In Section 2, we review the related work and position our contributions in comparison to the existing literature. Section 3 defines the problem statement and extends two recent multimodal conversational recommendation models for top-K recommendations. Section 4 presents the existing user simulator for relative captioning and extends it for generating both positive and negative natural-language feedback. Our experimental setup and results are presented in Sections 5 and 6, respectively. Section 7 summarises our findings and provides possible future work.

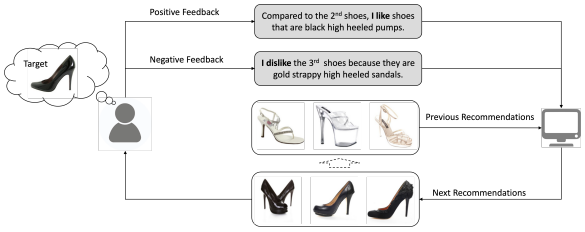
2 RELATED WORK

In this section, we first introduce multimodal conversational recommendations and survey related work. We also introduce positive and negative natural-language feedback in the recommendation field.

Multimodal Conversational Recommendations. Vision-and-language-based interactions between users and recommender systems can be effective for the benefits of both visual information from the recommendations' images and textual information from the users' natural-language feedback [11, 20, 23, 28]. In particular, the users' natural-language critiques about the recommendations can allow the recommender systems to correctly track the users' preferences over time and adapt the systems' instant recommendations, thereby satisfying the users' information needs effectively. Recently, a variety of research in the conversational recommendation field have leveraged the recommendation models' ability in understanding the users' preferences from their natural-language feedback while continuously providing visual recommendations during the multimodal interactions between users and recommender systems [11, 23, 24, 26, 27, 30]. Based on the key component in the state trackers for tracking the states of the visual-language dialog and estimating the users' preferences, the existing multimodal conversational recommendation models can be divided into two major categories: recurrent-neural-network-based (RNN-based) models [11, 24, 30] or transformer-based models [23]. In particular, Guo et al. [11] proposed a Dialog Manager (DM) model based on a gated recurrent unit (GRU) using model-based reinforcement learning to track and estimate the users' preferences from both the users' natural-language feedback and the recommended visual items during the multi-round interactions. Similar to [11], Wu et al. [24] also adopted a GRU component in their Estimator-Generator-Evaluator (EGE) model based on reinforcement learning with a partially observable Markov decision process (POMDP) in a partially observational environment. Meanwhile, Zhang et al. [30] adopted a single long short-term memory (LSTM) component in their reward-constrained recommendation (RCR) model based on constrained-augmented reinforcement learning for mitigating the recommendations that violate the user previous comments. Furthermore, Wu et al. [23] proposed a Multimodal Interactive Transformer (MIT) model to incorporate visual items' features, users' natural-language feedback, and fashion attributes. Despite their general good performances in multimodal conversational recommendation task, these research only focus on positive natural-language feedback with the users' preferred attributes in the top-1 recommendation task. However, the users in the real-world shopping scenario can freely express



(a) Multimodal Conversational Recommendation



(b) Positive & Negative Natural-Language Feedback

Figure 1: An example of multimodal conversational recommendation with positive and/or negative natural-language feedback.

their satisfactions over the top-K recommendations *positively* or *negatively*. To this end, in this paper, we investigate the impact of positive and negative natural-language feedback on the effectiveness of the recent multimodal conversational recommendation models. Both positive and negative natural-language feedback can be directly incorporated into the existing multimodal conversational recommendation models, such as the Dialog Manager (DM) [11] and Multimodal Interactive Transformer (MIT) [23] models.

Positive & Negative Natural-Language Feedback. Positive/negative explicit/implicit feedback (such as ratings, transactions, clicks, and skips) have been intensively investigated in the recommendation field [1, 3, 5, 22, 33, 35]. For instance, Zhao et al. [33] proposed a deep Q-learning network (DQN) based recommender system with GRUs by incorporating both positive implicit feedback (i.e. clicks) and negative implicit feedback (i.e. skips) from the logged implicit interactions datasets. In recent research, natural-language feedback has been proven to be more informative relating to the users' preferences compared to the non-verbal explicit/implicit feedback. Although natural-language feedback has been intensively investigated in the conversational recommendation field [9, 14, 33], these existing research on conversational recommendation imposed several constraints on the users' natural-language feedback to simplify the conversational recommendation task. In particular, the users are assumed to either only describe their preferred attributes as positive feedback [11, 19, 23, 24, 26, 27, 32] or just answer attribute-level questions with a binary yes/no response while rejecting the undesired item-level recommendations without any natural-language critiques [2, 16, 25] during the multi-turn interactions. For instance, the existing multimodal conversational recommendation models based either on a GRU [11, 24, 30] or a transformer [23] only consider the users' positive natural-language feedback for describing their desired features in terms of the recommendations, thereby directing the recommender systems towards obtaining a correct desired item. Meanwhile, Lei et al. [16] formulated the conversational recommendation task as answering the questions about the attributes and the recommended items with a binary yes/no response. Furthermore, a multi-round conversational recommender system (called Feedback-guided Preference Adaptation Network (FPAN)) [25] was recently proposed to consider the relation between attribute-level and item-level positive and negative feedback signals. The users' feedback is constrained to answer the questions asked by the recommender systems and is also simplified by answering "yes" for acceptance and "no" for rejection in terms of the attribute-level clarification questions and the and item-level recommendations from the recommender systems. However, we argue that users should be able to actively express their positive and/or negative critiques about the recommendations via natural language in addition to answering the recommender systems' questions. Such a constraint with only positive natural-language feedback or a simplification with "yes" or "no" is limited by the conversational recommendation datasets available, which makes the research less realistic in the shopping scenario. To this end, we propose an approach to generate both positive and negative natural-language feedback with the existing user simulator for relative captioning.

As a consequence, in this paper, we investigate the effectiveness of the existing multimodal conversational recommendation models with both positive and negative natural-language feedback that describes the users' desired/undesired features in terms of the visual recommendations. To the best of our knowledge, this is the first work for investigating mutlimodal conversational recommendations with both positive and negative natural-language feedback.

3 THE MULTIMODAL CONVERSATIONAL RECOMMENDATION MODELS

In this section, we introduce our notations and formulate the problem of the multimodal conversational recommendation task. Next, we extend two recent multimodal conversational recommendation models for top-K recommendations using both positive and negative textual feedback and describe each of its components. Finally, we describe training the models using the interactions with a simulated user.

3.1 Preliminaries

We study the multimodal conversational recommendation task by considering a user interacting with a recommender system via iterative interaction turns with text and images. At the t-th interaction turn, the recommender system presents K candidate images $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$ 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 differences between the candidate image and the desired image. The natural language feedback can be positive – having the form "Compared to the k-th item, I like ..." (i.e. o_t^+) or negative – such as "I dislike the k-th item because ..." (i.e. o_t^-). Based on the users' positive/negative natural-language feedback and the interaction history up to turn t, $\tau_t = (o_{\leq t}, a_{\leq t}) \in \mathcal{H}$, where $o_{\leq t} = (o_1, ..., o_t) \in \mathcal{O}$ and $a_{\leq t} = (a_1, ..., a_t) \in \mathcal{A}$, the recommender

system selects another candidate image a_t from the candidate image pool. This vision-language interaction process continues until the target image a_{tar} is recommended or the maximum number of interaction turns M is reached.

3.2 The Model Architecture

Figure 2 shows the architectures of two end-to-end models (i.e. Figure 2 (a) Dialog Manager (DM) [11] and Figure 2 (b) Multimodal Interactive Transformer (MIT) [23]) for multimodal conversational recommendations to effectively incorporate the users' preferences over time. The user views the recommended items (K items at each interaction) and provides positive natural-language feedback by describing their desired features that the current recommended items lack. Alternatively, the user can provide negative feedback by describing the undesired features in the current recommended items compared to the user's envisaged target item.

Text & Image Encoders. The multimodal conversational recommendation models track and estimate the user's preferences from both the user's positive/negative natural-language feedback and the latest recommended visual items. The positive/negative natural-language feedback texts are encoded with a text encoder, while the recommended images are encoded with an image encoder [11, 23]. In particular, the text encoder (which consists of a pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) [7], a 1D convolutional layer (1D-CNN) and a subsequent linear layer) encodes the positive and negative natural-language feedback texts into a single textual representation. Alternatively, each word in the sentences can also be represented by a one-hot vector with pre-defined vocabulary [11, 23] of fashion-related terms. We adopt the pre-trained BERT model as our default encoding mechanism, while we investigate the impact of different encoding mechanisms (i.e. the one-hot encoding and the BERT encoding) in Section 6.3. In a similar manner to the text encoder, the image encoder extracts image feature representations based on the ImageNet pre-trained ResNet101 model [12] and subsequently transforms the extracted image feature representations with a linear layer. Then, both the image feature representations and the textual representations are concatenated as input to a subsequent GRU [11] or transformer [23] to model the user's estimated preferences.

The State Trackers. Given a list of candidate images $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$ and a user's corresponding natural-language feedback o_t at the t-th dialog turn, the encoded textual representation is denoted by x_t^{txt} and the encoded image representation is denoted by $x_{t,\leq K}^{img} = ResNet(a_{t,\leq K})$. The concatenated textual and image representations $[x_t^{txt}, x_{t,\leq K}^{img}]$ are further tracked in a gated recurrent unit (GRU) [4] as in [11]. The estimated state of user's preferences can be achieved with $s_{t+1} = Linear(GRU(Linear([x_t^{txt}, x_{t,\leq K}^{img}]), h_t))$, where $h_t = GRU(Linear([x_{t-1}^{txt}, x_{t-1,\leq K}^{img}]), h_{t-1})$ is the estimated hidden states of the user's preferences. The GRU component allows the model to sequentially aggregate the recommendations and positive/negative feedback information from the recommender system's recommendations and the user's natural-language feedback to the estimated hidden states. Alternatively, a transformer-based state tracker enables the recommendation model to attend to the entire history of the multimodal interactions. The estimated state of user's preferences can be achieved with

$$s_{t+1} = Linear(Mean(Transformer([x_{\leq t}^{txt}, x_{\leq t,\leq K}^{img}]))).$$

The Top-K Recommendations. Based on the estimated state of user's preferences, a list of candidate items can be recommended for the next action. If K items are recommended at each turn $t+1$, we select the top-K closest images to the estimated state s_{t+1} under the Euclidean distance in the image feature (ResNet) space: $a_{t+1,\leq K} \sim KNNs(s_{t+1})$, where $KNNs()$ is a softmax distribution over the top-K nearest neighbours of s_{t+1} and $a_{t+1,\leq K} = (a_{t+1,1}, ..., a_{t+1,K})$. Furthermore, to avoid repeated recommendations during the multi-turn interactions, we adopt a post-filter, as in [24], to remove any candidate items from the ranking list that have previously occurred in the recommendation history $a_{\leq t}$.

The Triplet Loss Function. User simulators [8, 11, 23, 31] are generally used as a surrogate for real human users in the training processes. For a fair comparison, we train the above GRU/transformer-based models with a triplet loss objective, L_{tri}, similar to [11, 23]:

$$L_{tri} = max(0, ||s_{t+1} - x_+^{img}||_2 - ||s_{t+1} - x_-^{img}||_2 + m) \quad (1)$$

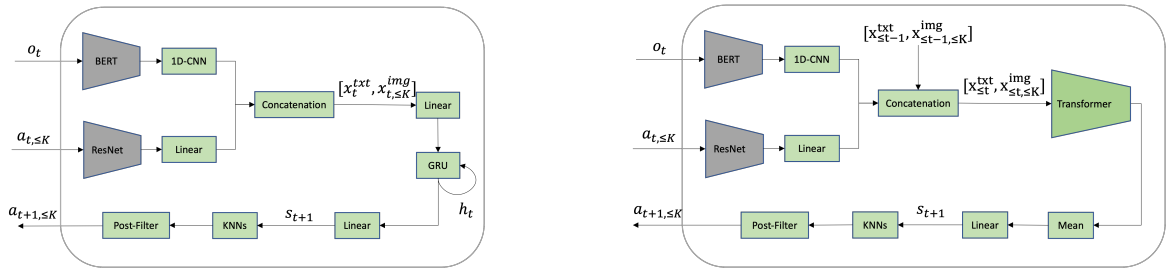
where x_+^{img} and x_-^{img} are respectively the representations of the target image and of a randomly sampled image, m is a constant for the margin and $||.||_2$ denotes L^2-norm.

4 A USER SIMULATOR WITH POSITIVE/NEGATIVE FEEDBACK

To learn satisfactory multimodal conversational recommender systems with enough training data, user simulators based on vision and language (VL) have been considered as surrogates for real human users in the optimisation and evaluation processes [11, 23, 30]. The adoption of such VL-based user simulators helps to avoid collecting and annotating entire multi-modal conversations, which is expensive, time-consuming, and does not scale [31].

User Simulators for Relative Captioning. Such user simulators have been generally formulated as *relative captioners* for fashion recommendation [11, 23] that can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images (i.e. a target representing the user's desired item and the candidate representing a recommendation by the system). For instance, Guo et al. [11] applied long short-term memory network (LSTM)-based models, such as *Show, Tell* [21], to generate the relative captions as natural-language critiques about the recommendations. These user simulators for relative captioning have been thoroughly evaluated via both a quantitative evaluation and a user study, which showed that the user simulators for relative captioning can serve as a reasonable proxy for real users [11].

User Simulators with Positive/Negative Feedback. Here, we propose an approach to generate positive and negative natural-language feedback with the existing user simulators for relative captioning. In the relative captioning task, the model is given a candidate image $a_{t,k}$ ($k \in [1, K]$) and a target image a_{target} and it is tasked with describing the differences of $a_{t,k}$ relative to a_{target} in natural



(a) Dialog Manager (DM)

(b) Multimodal Interactive Transformer (MIT)

Figure 2: Architectures of the multimodal conversational recommendation models: (a) Dialog Manager (DM) and (b) Multimodal Interactive Transformer (MIT).

language. To generate both positive and negative feedback, positive feedback is defined of pairs of images $a_{t,k}$ and a_{target} with a corresponding relative caption $cap_{rel}()$, as follows:

$$o_t^+ = \text{“Compared to the k-th item, I like”} + cap_{rel}(a_{target}, a_{t,k})$$

where each relative caption $cap_{rel}(a_{target}, a_{t,k})$ describes what is missing from candidate image $a_{t,k}$ to obtain a_{target}. We propose that negative feedback can thus be instantiated by reversing candidate and target images:

$$o_t^- = \text{“I dislike the k-th item because”} + cap_{rel}(a_{t,k}, a_{target})$$

and changing the textual prefix from “I like” to “I dislike”. It is worth noting that we adopt templates as wrappers to handle users' positive and negative utterances so as to reduce the errors for language understanding and generation.

5 EXPERIMENTAL SETUP FOR RECOMMENDATION

In this section, we evaluate the effectiveness of the two existing multimodal conversational recommendation models from the literature with different types of natural-language feedback (i.e. positive and/or negative feedback). In particular, we address the three research questions:

- RQ1: Is positive natural-language feedback more informative relating to the users' preferences in comparison to negative natural-language feedback?
- RQ2: Can the combined positive & negative natural-language feedback enhance the ability of the existing GRU/transformer-based models in incorporating the users' preferences?
- RQ3: What is the impact of the natural-language encoding on the models' performances?

5.1 Dataset & Measures

Dataset. We perform experiments on the *Shoes* dataset [11]. The dataset provides 10, 751 pairs of images with relative captions about their visual differences and 3, 600 images with captions about their discriminative visual features for training a user simulator. In addition, the dataset also contains 10, 000 images for training the recommender systems, and 4, 658 images for testing. We apply the same training and testing data splits for the tested recommendation models.

Measures. The effectiveness of the multimodal conversational models is measured by Normalised Discounted Cumulative Gain (i.e. NDCG@N truncated at rank $N = 10$ calculated at the M-th interaction) and Success Rate (SR) at the M-th interaction, as in [24]. In particular, SR is the percentage of users who find their target items in the top-K recommendation lists among all the users within M interactions. Furthermore, it is possible that the user may view more of the ranking of items at each interaction turn, down to rank N. We use the evaluation metrics (i.e. NDCG@10 and SR) at the 5th and 10th interaction turn for significance testing.

5.2 Experimental Settings

Setup for User Simulator. A user simulator with the *Shoes* dataset was intensively and carefully trained by [11] through crowdsourcing relative expressions about the visual differences of the image pairs that are written by real human users in natural language. Furthermore, the pre-trained user simulator has previously been thoroughly evaluated via both a quantitative evaluation and a user study [11], thereby serving as a reasonable proxy for real users in our work. The pre-trained user simulator can generate either positive or negative natural-language feedback with our proposed approach as illustrated in Section 4.

Setup for Recommender Systems. We then train the models (i.e. DM and MIT) with the user simulator on the *Shoes* dataset. The parameters of the models are randomly initialised. We use Adam [15] with a learning rate 10^{-3} [11, 30]. We set the embedding dimensionality of the feature space to 256 and the batch size to 128 as in [11]. For each batch, we train the model with 10 interaction turns as in [24]. We consider the top-K items (K=3) as a recommendation list at each interaction turn for testing. For the evaluation metrics, we denote the interaction turn $M \in [1, 10]$. If a user obtains the target item in less than 10 interaction turns, we consider the ranking metric (i.e. NDCG@10) for that user to be equal to one for all turns thereafter [24].

6 EXPERIMENTAL RESULTS

In this section, we analyse the experimental results respect to the research question stated in Section 5, concerning the effectiveness of the models for multimodal conversational recommendations with positive and negative natural-language feedback (Section 6.1), the

impact of the combined positive and negative feedback (Section 6.2), and the impact of the natural-language encoding mechanisms (Section 6.3). We demonstrate a use case for generating both positive and negative feedback, as well as a use case from the logged experimental results to consolidate our findings (Section 6.4).

6.1 Positive Feedback vs. Negative Feedback (RQ1)

Figure 3 shows the recommendation effectiveness of the DM and MIT models with positive or negative single-sentence feedback for top-3 recommendation in terms of NDCG@10 (Figure 3 (a)) and SR (Figure 3 (b)), while varying the number of interaction turns on the *Shoes* dataset. The solid lines show the models' performances with positive natural-language feedback, while the dashed lines show performances with negative natural-language feedback. Comparing the results in Figure 3, we observe that both DM and MIT models generally achieve a better overall performance with positive feedback than negative feedback in terms of NDCG@10 and SR. The better performance of the tested models with positive feedback compared to those with negative feedback indicates that positive natural-language feedback in more informative relating the users' preferences than negative natural-language feedback. In addition, MIT achieves a better overall performance than DM in terms of NDCG@10 and SR at various interaction turns with positive and negative natural-language feedback. Such an observation is aligned with the results reported in [23] considering positive natural-language feedback only.

Table 1 shows the obtained recommendation performances of the tested models (i.e. DM and MIT) with the same test sets of the *Shoes* dataset at the 5th and 10th interaction turns. More specifically, Table 1 contains two parts: the first part reports the effectiveness of the models with either positive or negative feedback. The second part reports the effectiveness of the models with different combinations of positive or negative feedback. The best performing results in the first and second parts of the table are underlined, while the best overall performing results are highlighted in bold in Table 1. † and * respectively denote significant differences in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to the best performing results in the first group and the best overall performing results. Comparing the results in the first group of rows in the table, we observe that both DM and MIT achieve a *significant* better overall performance in terms of both NDCG@10 and SR at the 5th and 10th turns with positive feedback (denoted +) than with negative feedback (denoted -) on the *Shoes* dataset, respectively.

In answer to RQ1, the results demonstrate that the tested models with positive feedback are significantly more effective than those with negative feedback. Therefore, it can be inferred that positive feedback is more informative relating to the users' preferences than negative feedback. The DM and MIT models can better incorporate the users' preferences from the recommended visual items with positive natural-language feedback than negative natural-language feedback.

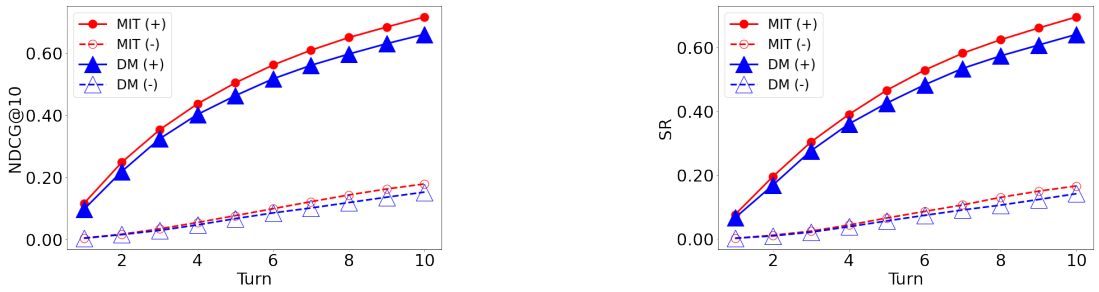
6.2 Impact of the Combined Feedback (RQ2)

Figure 4 (a) and Figure 4 (b) illustrates the SR of DM and MIT with different types of natural-language feedback (i.e. different combinations of positive and negative feedback) at the various interaction turns, respectively. The gray lines show the DM/MIT model's performances with a single sentence at each interaction turn, while the blue/red and green lines show performances with a pair of sentences at each interaction. Comparing the results in Figure 4 (a) and Figure 4 (b), we observe that both DM and MIT achieve a better over performance with paired positive (i.e. + & +) or paired negative (i.e. - & -) natural-language feedback sentences in comparision to the models with a single positive (i.e. +) or single negative (i.e. -) natural-language feedback sentence. Furthermore, the performances of DM and MIT differ with a pair of both positive and negative feedback sentences. In particular, the performance of DM (+) and DM (+ & -) are very close in term of SR at various interaction turns, while MIT (+) outperforms MIT (+ & -) overall except for the initial two interaction turns. The better performance of the models with (+ & +) and (- & -) compared to the models with (+) and (-) can be attributed to the fact that the same type of natural-language feedback at each turn can be aggregated to leverage the information relating to the users' preferences. Meanwhile, the paired positive and negative feedback make it challenging for DM and MIT the elicit the users' preferences from the feedback sentences with opposite sentiments. Furthermore, Table 1 demonstrate that DM (+ & +) and MIT (+ & +) are significantly more effective than those with other types of natural-language feedback at both 5-th and 10-th interaction turns, except for MIT (+) in term NDCG@10 and SR at the 10-th interaction turn.

Overall, in response to RQ2, we find that the single type of natural-language feedback (i.e. either paired positive or paired negative feedback) at each turn can be aggregated to leverage the information relating to the users' preferences, while the paired positive and negative feedback make it challenging for DM and MIT to elicit the users' preferences.

6.3 Impact of Natural-Language Encoding (RQ3)

To address RQ3, Figure 5 depicts the effects of the textual encoding mechanisms on both DM and MIT with different types of natural-language feedback. Figure 5 (a) demonstrates that both DM (+) and MIT (+) with the one-hot encoding using a pre-defined vocabulary of fashion-related terms achieve an overall better performance in comparison to those with the BERT encoding. Figure 5 (b) shows that MIT (-) with the one-hot encoding also outperforms MIT (-) with the BERT encoding, while DM (-) with the one-hot encoding and the BERT encoding are almost the same. The better performance of the models with the one-hot encoding compared to the BERT encoding can be attributed to the fact that the pre-defined fashion vocabulary for the one-hot encoding is much smaller and is more concentrated on fashion features than BERT. Furthermore, Figure 5 (c) shows that the performances of DM (+ & -) and MIT (+ & -) with the one-hot encoding are dramatically degraded compared to those with the BERT encoding that is able to capture the contextual information between sentences with the pre-trained contextual embeddings. Such a difference can be attributed to the

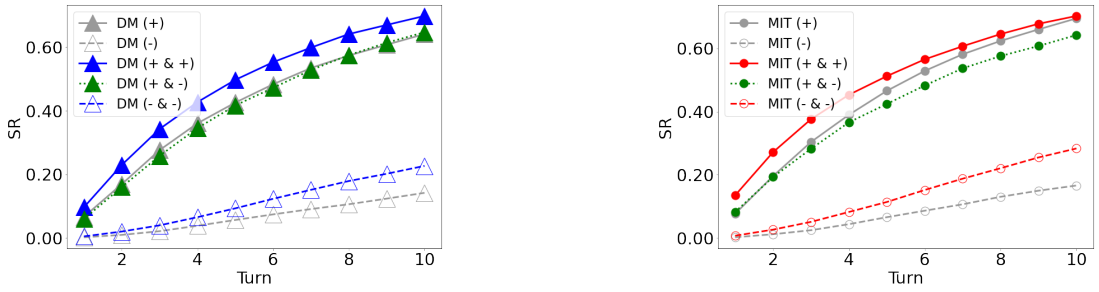


(a) NDCG@10 (b) SR

Figure 3: Comparison of the recommendation effectiveness of DM and MIT with single-sentence feedback at various interaction turns with top-3 recommendation on *Shoes*. + & - denote positive and negative natural-language feedback, respectively.

Table 1: Multimodal conversational recommendation effectiveness of the tested models at the 5th and 10th turns on the *Shoes* dataset. The best overall results are highlighted in bold. The best performing results in the first and second parts of the table are underlined, while the best overall performing results are highlighted in bold. † and * respectively denote significant differences in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to the best performing results in the first group and the best overall performing results. + and - denote positive and negative natural-language feedback, respectively.

Models →	DM				MIT			
Feedback Type ↓	Turn 5 NDCG@10	Turn 5 SR	Turn 10 NDCG@10	Turn 10 SR	Turn 5 NDCG@10	Turn 5 SR	Turn 10 NDCG@10	Turn 10 SR
+	0.4627*	0.4253*	0.6602*	0.6404*	0.5039*	0.4657*	0.7158	0.6949
-	0.0675*†	0.0567*†	0.1527*†	0.1419*†	0.0771*†	0.0659*†	0.1791*†	0.1662*†
+ & +	**0.5330**	**0.4966**	**0.7157**	**0.6973**	**0.5471**	**0.5122**	**0.7210**	**0.7027**
+ & -	0.4524*	0.4163*	0.6650*	0.6462*	0.4628*	0.4242*	0.6638*	0.6423*
- & -	0.1111*	0.0932*	0.2450*	0.2265*	0.1362*	0.1140*	0.3023*	0.2834*

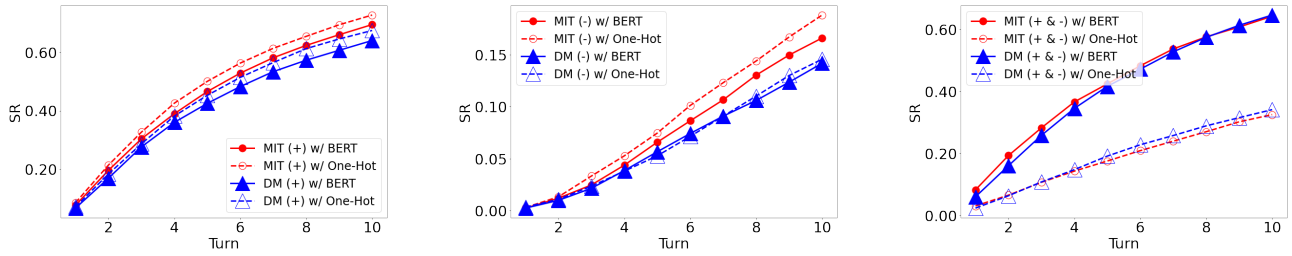


(a) DM (b) MIT

Figure 4: Comparison of the recommendation effectiveness of DM and MIT with various types of natural-language feedback at various interaction turns with top-3 recommendation on *Shoes*. + & - denote positive and negative natural-language feedback, respectively.

inability of the one-hot encoding in capturing the relations between the positive and negative natural-language feedback.

Overall, in response to RQ3, we find that the BERT encoding is surprisingly important to capture the contextual information with the pre-trained contextual embeddings when there are both positive and negative feedback, while the one-hot encoding can enhance the models' performance by using a pre-defined fashion vocabulary that is more concentrated on fashion features than BERT.



(a) Positive (b) Negative (c) Positive & Negative

Figure 5: Effects of the textual encoding mechanisms on both DM and MIT with different types of natural-language feedback.

6.4 Use Cases

A Use Case for Generating Positive & Negative Feedback. Table 2 demonstrates an example of the generated natural language critiques given a target image and a candidate image on the *Shoes* dataset. There are two ground truths for each pair of images, while following the generated positive and negative feedback by a user simulator for relative captioning.

A Use Case for Multimodal Conversational Recommendation. To consolidate the results observed in the the above sections, we present a use case of multimodal conversational recommendation in Table 3 and Table 4 on the *Shoes* dataset. Table 3 and Table 4 show the interaction process for top-3 recommendation with the DM model over positive feedback (i.e. DM (+)) and negative feedback (i.e. DM (-)), respectively. For fair comparison, the initial images are the same across DM (+) and DM (-) given the target image from the testing set. We observe that DM with positive feedback is more effective than negative feedback. In particular, DM with positive feedback only needs 2 interactions to display the desired item in addition to the initial random recommendation by capturing the key features from the user's positive feedback, such as "gold", "open-toed", "high heels", and "straps". However, DM with negative feedback fails to recommend the user's desired shoes within 5 interaction turn. Although, DM with negative feedback can successfully capture the "open toe" feature from the rejection of the "closed toe" feature, it is still struggling with the decisions of the colours and the thickness of the platform.

7 CONCLUSIONS

In this paper, we first investigated the effectiveness of the multimodal conversational recommendation models with both positive and negative natural-language feedback. To make the conversational recommendation task more realistic with both positive and negative natural-language feedback, we proposed an approach to generate both the positive and negative natural-language critiques about the recommendations with the existing user simulator for relative captioning. Following previous work, we trained and evaluated the two existing conversational recommendation models by using the user simulator with positive and negative feedback as a surrogate for real human users. Our experiments on the *Shoes* dataset demonstrated that positive feedback is more informative relating to the users' preferences in comparison to negative feedback. Our reported results also showed that the types of users' natural-language feedback (i.e. different combinations of positive and negative feedback) and the types of textual encoding mechanisms (i.e. pre-trained contextual embeddings and one-hot embeddings) can greatly affect the performance of the both tested models (i.e. DM & MIT). For future work, we plan to investigate an end-to-end model with pre-trained transformers in Fashion (such as FashionBERT [10] and Kaleido-BERT [34]) to better incorporate the users' preferences from positive and negative natural-language feedback and visual recommendations. In addition to visual and textual modalities, an increasing number of live shopping videos with audio signals are available to display and introduce fashion products in a more vivid and detailed way owing to the flourishing live stream shopping (such as TikTok Live Shopping). We also plan to incorporate such more advanced modality (i.e. live shopping videos) into the multimodal conversational recommendation.

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] Zeynep Batmaz, Ali Yurekli, Alper Bilge, and Cihan Kaleli. 2019. A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review* 52, 1 (2019), 1–37.
[2] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. 2019. Conversational product search based on negative feedback. In *Proc. CIKM*. 359–368.
[3] Samit Chakraborty, Md Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, Edgar Lobaton, et al. 2021. Fashion Recommendation Systems, Models and Methods: A Review. *Informatics* 8, 3 (2021), 49.
[4] 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).
[5] 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).
[6] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multimodal conversational information seeking. In *Proc. SIGIR*. 1577–1587.
[7] 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
[8] Michael D Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. 2021. SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research. In *Proc. RecSys*. 803–805.
[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).

Table 2: An example of positive and negative feedback with relative captioning on the *Shoes* dataset.

Pair	Target	Candidate	Feedback
1			**Ground Truths:** • Compared to the candidate shoes, the target shoes are shoes with red trim. • Compared to the target shoes, the candidate shoes are brown, not red. **Positive Feedback:** Compared to the candidate shoes, I like shoes that are red and black. **Negative Feedback:** I dislike the candidate shoes because they are brown.
2			**Ground Truths:** • Compared to the candidate shoes, the target shoes are the same design but are brown. • Compared to the target shoes, the candidate shoes are black, not brown. **Positive Feedback:** Compared to the candidate shoes, I like shoes that are the same design but are brown. **Negative Feedback:** I dislike the candidate shoes because they are black, not brown.
3			**Ground Truths:** • Compared to the candidate shoes, the target shoes are all white. • Compared to the target shoes, the candidate shoes have pink accents and more lace eyelets. **Positive Feedback:** Compared to the candidate shoes, I like shoes that are solid white. **Negative Feedback:** I dislike the candidate shoes because they have pink accents and more eyelets.
4			**Ground Truths:** • Compared to the candidate shoes, the target shoes are almost identical. • Compared to the target shoes, the candidate shoes have more athletic soles. **Positive Feedback:** Compared to the candidate shoes, I like shoes that are almost identical. **Negative Feedback:** I dislike the candidate shoes because they have a chunkier sole.

[10] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proc. SIGIR*. 2251–2260.

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

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

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

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

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

[16] 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.

[17] 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.

[18] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proc. CVPR*. 7008–7024.

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

[20] 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.

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

[22] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: challenges, progress and prospects. *Proc. IJCAI*, 6332–6338.

[23] 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.

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

Table 3: An example use case for multimodal conversational recommendation for the Dialog Manager model with positive natural-language feedback on the *Shoes* dataset.

Turn	Top-3 Recommendations	Positive Feedback
0		Compared to the 3rd shoes, I like shoes that are gold open toe high heels.
1		Compared to the 3rd shoes, I like shoes that are open-toed with straps.
2		The 1st shoes are my desired shoes.

[25] 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.

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

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

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

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

[30] 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.

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

[32] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proc. CIKM*. 177–186.

[33] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proc. KDD*. 1040–1048.

[34] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proc. CVPR*. 12647–12657.

[35] Lixin Zou, Long Xia, Yulong Gu, Xiangyu Zhao, Weidong Liu, Jimmy Xiangji Huang, and Dawei Yin. 2020. Neural Interactive Collaborative Filtering. In *Proc. SIGIR*. 749–758.

Table 4: An example use case for multimodal conversational recommendation for the Dialog Manager model with negative natural-language feedback on the *Shoes* dataset.

Turn	Top-3 Recommendations	Negative Feedback
0		I dislike the 1st shoes because they are colorful and white running shoes.
1		I dislike the 1st shoes because they are black with a closed toe.
2		I dislike the 2nd shoes because they are red and have a pattern.
3		I dislike the 2nd shoes because they are beige open toed pumps.
4		I dislike the 2nd shoes because they are black strappy high heeled shoes
5		I dislike the 2nd shoes because they have a higher platform.