Partially Observable Reinforcement Learning for Dialog-based Interactive Recommendation

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

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

ABSTRACT

A dialog-based interactive recommendation task is where users can express natural-language feedback when interacting with the recommender system. However, the users' feedback, which takes the form of natural-language critiques about the recommendation at each iteration, can only allow the recommender system to obtain a partial portrayal of the users' preferences. Indeed, such partial observations of the users' preferences from their natural-language feedback make it challenging to correctly track the users' preferences over time, which can result in poor recommendation performances and a less effective satisfaction of the users' information needs when in presence of limited iterations. Reinforcement learning, in the form of a partially observable Markov decision process (POMDP), can simulate the interactions between a partially observable environment (i.e. a user) and an agent (i.e. a recommender system). To alleviate such a partial observation issue, we propose a novel dialog-based recommendation model, the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP, to effectively incorporate the users' preferences over time. Specifically, we leverage an Estimator to track and estimate users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations, and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. Following previous work, we train our EGE model by using a user simulator which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Thorough and extensive experiments conducted on two recommendation datasets – addressing images of fashion products (namely dresses and shoes) – demonstrate that our proposed EGE model yields significant improvements in comparison to the existing state-of-the-art baseline models.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Theory of computation** → **Reinforcement learning**.

KEYWORDS

interactive recommendation, multimodal, reinforcement learning

ACM Reference Format:
Yaxiong Wu and Craig Macdonald, Iadh Ounis. 2021. Partially Observable Reinforcement Learning for Dialog-based Interactive Recommendation. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3460231.3474256

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

1 INTRODUCTION

Recently, interactive recommender systems (IRS) have received much attention due to their flexible recommendation strategies and their natural multi-step decision-making processes. A typical interactive recommender system continuously recommends items to users and receives various types of users' feedback, such as clicks, ratings, or textual replies [7, 28, 37, 40]. In particular, natural-language feedback allows an interactive recommender system to obtain richer information relating to the users' current preferences, thereby leading to a more suitable recommendation compared to clickthrough data and ratings [36]. Figure 1 shows an example of interactive recommendation based on natural-language feedback, i.e. dialog-based interactive recommendations. In this use case, the user gives natural-language critiques about the system's recommendation at each interaction turn and aims to quickly find the target item, while the system recommends the top-1 item according to the user's natural-language feedback.

Such an interactive recommendation task has been formulated and modelled using reinforcement learning (RL) approaches [3, 9, 21, 26, 34, 36, 38]. In the reinforcement learning framework, the interactive recommendation task is usually formulated as a Markov decision process (MDP) [29] with an assumption that the environment's states (i.e. the users' preferences) are *fully observable*. Such RL-based interactive recommender systems have demonstrated their benefits in fitting the users' dynamic preferences and maximising the expected long-term cumulative rewards from users when achieving the optimal strategies. For instance, a recently proposed Supervised Q-learning (SQN) framework [32] (i.e. a joint learning framework with both a supervised learning layer and a Q-learning layer) was shown to outperform neural recommendation models using supervised learning, such as GRU4Rec [14], Caser [30] and SASRec [17], by taking the Q-learning layer as a regulariser to introduce reward-driven properties (such as long-term user engagement [39]) to the recommendation process.

Despite the expressiveness of natural-language feedback in dialog-based interactive recommendations, the users' feedback can only allow the recommender system to obtain a partial portrayal of the users' preferences. For instance, Figure 1 shows an example of a dialog-based interactive recommendation process between the user (simulator) and the system generated by a RL-based interactive recommender system (called Model-based Policy Improvement (MBPI)) [9]. Each natural-language comment in terms of the current recommendation only contains partial visual features of the target

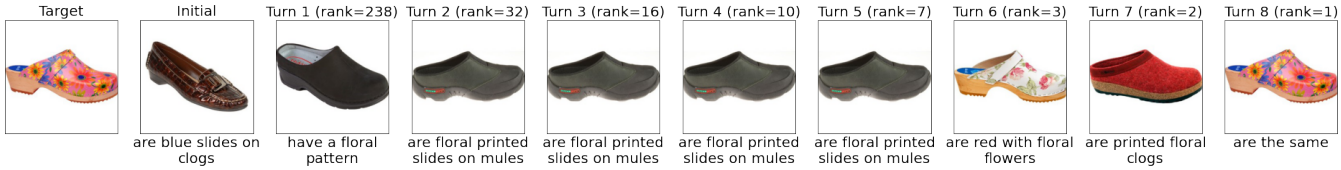


Figure 1: An example of dialog-based interactive recommendations [9]. The first image is the target item desired by the user (labeled with "Target"), while the second image (labeled with "Initial") is the initial recommendation proposed by the system randomly. Then, the user gives natural-language critiques about the recommendation at each turn, while the recommender system updates the ranking list and recommends the top-1 item according to the user's comments. The rank of the target item is also presented above the images at each turn. When the target item is recommended at rank 1, the user will give a comment "are the same".

item, such as "blue slides" at the initial turn and "red with floral flowers" at the 6th turn. Such partial observations of the users' preferences from their natural-language feedback can drive the recommender system towards a degenerate preference estimation that ignores certain features in the historical observations [6], i.e. historical natural-language feedback and historical recommendations. For instance, in Figure 1, "red" clogs are recommended due to the last comment "red with floral flowers" at the 7th interaction turn, while the single "red" colour in the recommended image is violated by the initial comment "blue slides on clogs" and the other comments with "floral". In addition to this so-called violated recommendation issue, *repeated recommendations* can also be observed in the example IRS [9] in Figure 1. Although the rank of the target shoe indicated above each suggested image is increasing from the 2nd turn to the 5th turn, the top-1 recommendation remains the same and hence receives identical feedback from the user (simulator). Both these violated and repeated recommendations can hurt the users' experience thereby increasing their disappointment in the interaction processes with the recommender system. Indeed, such partial observations of the users' preferences from their natural-language feedback make it challenging to correctly track the users' preferences over time, which can result in a poor performance of the recommender system in satisfying the users' information needs when in the presence of limited iterations (as can be observed from the literature [9]). Although Zhang et al. [36] proposed a reward-constrained recommendation (RCR) model with constraint-augmented reinforcement learning that can effectively mitigate the aforementioned violation issue, the utility of their RCR model was limited by the issue that extra rounds of resampling and violated recommendation detection are needed when the previous samplings are detected as violated recommendations. Furthermore, RCR does not address the repeated recommendation issue, and is less useful when well-categorised visual attributes of items are not available.

In our paper, we formulate the dialog-based interactive recommendation task as a partially observable Markov decision process (POMDP) [29] to simulate the interactions between a partially observable environment (i.e. a user) and an agent (i.e. a recommender system). To correctly estimate the users' preferences from such partially observable situations, we extend the SQN framework [32] from a MDP to a POMDP and judge/optimise the quality of the estimated users' preferences with the Q-learning layer (also called an Evaluator). To this end, we propose a novel dialog-based interactive recommendation model, called the Estimator-Generator-Evaluator (EGE) model named after its three distinctive functional components, which apply partially observable reinforcement learning (i.e. Q-learning for POMDP) for dialog-based interactive recommendation to effectively incorporate the users' preferences over time. Specifically, we leverage an Estimator to track and estimate the users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations, and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. To mitigate the impact of repeated recommendations, a post-filter is adopted to remove the repeated recommended items from the ranking list based on the recommendation history. Following previous work [9, 36], we train our EGE model by using a user simulator [9], which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Thorough and extensive experiments conducted on two recommendation datasets – addressing images of fashion products (namely dresses and shoes) – demonstrate that our proposed EGE model yields significant improvements in comparison to the existing state-of-the-art baseline models (i.e. a sequential recommendation model (denoted iGRU) with supervised learning and the Model-based Policy Improvement (MBPI) model). The main contributions of this paper are summarised as follows:

- We propose a novel dialog-based interactive recommendation model, the Estimator-Generator-Evaluator (EGE) model, which formulates the dialog-based interactive recommendation task as a partially observable Markov decision process (POMDP) to address the partial observations issue in the users' feedback. Our proposed EGE model differs from the existing MBPI [9] and RCR [36] models as follows: EGE judges and optimises the quality of the estimated user preferences with a Q-learning layer (i.e. Evaluator) for POMDP based on the users' history feedback, while the MBPI model (with only the Estimator and Generator components) is not able to do so and the RCR model needs to repeatedly sample recommendations and detect violations with extra well-categorised visual attributes of items.
- The EGE model extends the SQN [32] framework from a MDP to a POMDP and is trained with a combination of a supervised learning classification loss and a Q-learning prediction loss.
- Extensive empirical evaluations are performed on the dialog-based interactive recommendation task, demonstrating significant improvements 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 present some necessary background, review the related work, and position our contributions in comparison to the existing literature. Section 3 defines the problem statement and presents our proposed EGE model. Our experimental setup and results are presented in Sections 4 and 5, respectively. Section 6 summarises our findings and provides possible future work.

2 BACKGROUND & RELATED WORK

In this section, we first introduce reinforcement learning, including concepts such as *Markov decision process (MDP)* and *partially observable Markov decision process (POMDP)*. We also introduce dialog-based interactive recommendations and survey related work.

Reinforcement Learning. Reinforcement learning (RL) deals with how *agents* ought to take *actions* in an *environment* with certain *states* in order to maximise the notion of cumulative *rewards*. Basic reinforcement learning is modelled as a Markov decision process (MDP) with an assumption that the complete information of the environment is *fully observable*. There are many variants of RL algorithms, such as Q-learning, SARSA, Policy Gradient, Actor-Critic, and many others [11, 19, 27, 29]. RL has been applied to the recommendation systems field by viewing the recommendation task as sequential interactions between a recommender system (i.e. agent) and users (i.e. environments). Such RL-based recommendation models, such as the Supervised Q-learning (SQN) framework [32], have demonstrated a generally better performance compared to neural recommendation models using supervised learning, such as GRU4Rec [14], Caser [30] and SASRec [17]. In particular, the SQN framework [32] extends existing sequential recommendation models [14, 17] with a Q-learning layer to introduce *reward*-driven properties to the recommendation process.

When complete information about the environment is not available, a reinforcement learning problem is modelled as a partially observable Markov decision process (POMDP) [29]. For instance, given only a single game screen, the game of Pong is a POMDP because a single observation does not reveal the velocity of the ball - indeed it only reveals the location of the paddles and the ball [12]. The estimated states, which characterise the distribution over the latent states in a POMDP, are typically modelled using recurrent neural networks (RNNs), and have been shown to be effective for reinforcement learning in POMDP scenarios [12, 15]. For example, a Deep Recurrent Q-Network (DRQN) [12] was proposed to successfully integrate information (i.e. the location of the paddles and the ball) through time in the game of Pong to detect the ball's velocity, although it was capable of seeing only a single screen at each timestep. The POMDP formulation approach has been shown to be suitable for sequential recommendations [23] and conversational recommendations [28] where it is not possible to fully observe a user's actions on all items in a recommender system, as well as all the desired features expressed by the users' natural-language feedback. For the same reason, the POMDP formulation approach is typically also applicable for dialog-based interactive recommendations.

However, the state estimators in the POMDP formulation might be driven by such partial observations towards a degenerate state that ignores the historical observations [6], which may degrade the performance of recommender systems when recommending items that violate the users' preferences [36]. To this end, inspired by the Supervised Q-learning (SQN) framework [32] with Q-learning as a regulariser, we propose an Evaluator with Q-learning in our proposed EGE model to judge the quality of the estimated states (i.e. the estimated users' preferences) with the users' historical feedback and improve the quality of the following recommendations. The EGE model extends the SQN framework [32] from a MDP to a POMDP to consider the partial observable nature of the users in dialog-based interactive recommendation scenarios.

Dialog-based Interactive Recommender Systems. Textual communications between users and recommender systems have been leveraged to understand the users' preferences and provide recommendations. In particular, deep learning (DL) and reinforcement learning (RL) have been used to understand the users' conversations and make recommendations [7, 16, 20, 28, 33]. However, these conversational recommender systems are purely text-based for both users and recommender systems. Visually-grounded dialog-based recommender systems [9, 34–36] have been proposed recently, which present images as recommendations to the users and obtain natural language feedback. In particular, a dialog-based interactive recommender system (called Model-based Policy Improvement (MBPI)) was introduced by Guo et al. [9] to enable users to provide feedback via natural language, allowing for more effective interaction compared to a supervised learning approach [9]. In addition, a pre-trained user simulator based on relative captioning [9, 25] was used to train and evaluate their proposed MBPI model by generating natural-language feedback that describe the differences between the target users' preferences and the system's recommendations.

However, as shown in Figure 1, recommendations by the MBPI model can violate the users' preferences from previous natural language feedback and can also be repeated. A reward-constrained recommendation (RCR) model with a Discriminator to introduce a penalty (i.e. a probability score of the violation's occurrence) to the users' reward (i.e. a score of the goodness of the latest recommendation) was proposed by Zhang et al. [36] to focus on recommending items that do not violate the user comments in the interactive recommendation through the use of constraint-augmented reinforcement learning. Although such a penalty-based method can effectively mitigate the aforementioned violation issue, the utility of the RCR model can be limited by the issue that extra resamplings and repeated violation detection are needed when the previous samplings are detected as violated recommendations. The reward-constrained approach of the RCR model cannot ensure that the recommended items will not occur in the future recommendation list. In addition, the RCR model heavily relies on extra well-categorised visual attributes of items that are not accessible in many recommendation scenarios. For instance, the *Shoes* dataset [9] for interactive shoe recommendation only contains target-candidate image pairs and their relative captions in natural language. To this end, we propose an approach based on the SQN framework for POMDP that optimises the quality of the estimated states (i.e. the estimated users' preferences). We maintain the gated recurrent unit (GRU) and the top-K-nearest-neighbours (KNNs) sampling as in the existing state-of-the-art RL-based approach [9] for a fair comparison with the existing models. Then we only need to perform the KNN sampling

once for generating the candidate recommendations without the extra resamplings. In addition, different from the RCR model with a Discriminator proposed by Zhange et al. [36], our proposed EGE model only needs natural-language feedback for training without extra well-categorised visual attributes of items.

As a consequence, in this paper, we argue that the existing dialog-based interactive recommendation models are not able to mitigate the partial observation issue effectively, which limits these models' ability of incorporating the users' preferences over time. These models either ignore the partial observation issue or rely on well-categorised visual attributes of items to repeatedly sample recommendations and detect violations, thereby limiting the applicability of these models to datasets without such information. Inspired by these previous dialog-based interactive recommendation works and the advanced RL methods, we propose an approach that judges and optimises the quality of the estimated states (i.e. the estimated users' preferences) with historical feedback to effectively mitigate the partial observation issue.

3 THE EGE MODEL

In this section, we introduce our notations and formulate the problem of the dialog-based interactive recommendation task via partially observable reinforcement learning. Next, we propose an Estimator-Generator-Evaluator (EGE) model and describe each of its components. Finally, we describe training the model using the interactions with simulated users.

3.1 Problem Statement

We study the dialog-based interactive recommendation task in a partially observable reinforcement learning formulation, using user feedback in the form of natural language. We consider the dialog-based interactive recommendation process as a partially observable Markov decision process (POMDP) with a tuple of seven elements $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{R}, \mathcal{T}, \mathcal{U}, \gamma)$, where $\mathcal{S}$ is a set of *states* (i.e. the users' preferences), $\mathcal{A}$ is a set of *actions* (i.e. the items for recommendation), $\mathcal{O}$ is a set of *observations* (i.e. the users' natural-language feedback), $\mathcal{R}$ is the *reward function*, $\mathcal{T}$ is a set of conditional transition probabilities between states, $\mathcal{U}$ is a set of conditional observation probabilities, and $\gamma \in [0, 1]$ is the *discount factor* for future rewards. We denote by $s_t \in \mathcal{S}$ the estimated user preferences at time t. When an item $a_t \in \mathcal{A}$ is recommended, the estimated preferences change according to the transition distribution, $s_{t+1} \sim T(s_{t+1}|s_t, a_t)$. Subsequently, the recommender agent receives a partial observation $o_{t+1} \in \mathcal{O}$ according to the distribution $o_{t+1} \sim U(o_{t+1}|s_{t+1}, a_t)$, and a reward $r_{t+1} \in \mathbb{R}$ according to the distribution $r_{t+1} \sim R(s_{t+1}, a_t)$.

A recommender agent acts according to its policy $\pi(a_t|o_{\leq t}, a_{<t})$, which returns the probability of taking action a_t at time t, and where $o_{\leq t} = (o_1, ..., o_t)$ and $a_{<t} = (a_0, ..., a_{t-1})$ are the observation and action histories, respectively. The recommender agent's goal is to learn a policy π that maximises the expected future return $J = \mathbb{E}_{p(\tau)}[\sum_{t=1}^{T} \gamma^{t-1} r_t]$ over trajectories $\tau = (s_0, a_0, ..., a_{T-1}, s_T)$ induced by its policy. In general, a dialog-based interactive recommender system via a POMDP must condition its actions on the entire history $h_t = (o_{\leq t}, a_{<t}) \in \mathcal{H}$.

3.2 The Model Architecture

In dialog-based interactive recommendations, a recommender agent recommends an item (in particular, an image) and a user provides natural-language feedback. Figure 2 shows our proposed end-to-end Estimator-Generator-Evaluator (EGE) model with partially observable reinforcement learning for dialog-based interactive recommendations to effectively incorporate the user's preferences over time. The user views the recommended item (a single item at each interaction) and gives natural-language feedback by describing their desired features that the current recommended item lacks. The system then incorporates the user's natural-language feedback and recommends (ideally) more-suitable items, until the desired item is found.

Estimator. The goal of the Estimator is to track and estimate the user's preferences (i.e. states) from both the user's natural-language feedback and the latest recommended visual item. The Estimator consists of a text encoder, an image encoder and a gated recurrent unit (GRU) [5] as in [9]. In particular, the text encoder extracts the textual sentence representations of the user's preferences from the current user's natural-language feedback. In the textual sentence representations, each word is represented by a one-hot vector. Similarly, the image encoder extracts image feature representations based on the ImageNet pre-trained ResNet101 [13] as in [9]. Then, both the image feature representations and the textual representations are concatenated as input to a following linear mapping (i.e. a multilayer perceptron (MLP)) and a GRU to obtain the estimated user's preferences. Given a candidate image a_{t-1} 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 $x_{t-1}^{img} = ResNet(a_{t-1})$. The estimated user's preferences can be achieved with $s_t = Linear(GRU(Linear([x_t^{txt}, x_{t-1}^{img}]), s_{t-1}))$. The GRU component of the Estimator allows our EGE model to sequentially aggregate the partially observable information from the user's natural-language feedback to the estimated preferences.

Generator. The goal of the Generator is to recommend a candidate item for the next action according to the estimated state. Considering the large amount of candidate images in the image database, all images are projected into the feature space (ResNet). If K items are recommended at each time t, we select the top K closest images to the estimated state s_t under the Euclidean distance in the image feature (ResNet) space: $a_{t,\leq K} \sim KNNs(s_t)$, where $KNNs()$ is a softmax distribution over the top-K nearest neighbours of s_t and $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$. Furthermore, based on the interaction history $h_t = (o_{\leq t}, a_{<t})$, a post-filter is adopted to remove any candidate items from the ranking list that have previously occurred in the recommendation history $a_{<t}$.

Evaluator. The Evaluator is proposed to judge the quality of the estimated state s_t at time t based on Q-learning. It performs the judgement process with the user's historical natural-language feedback $o_{\leq t} = (o_1, ..., o_t)$ to regularise the Estimator. Given an estimated state s_t and the textual features $x_{\leq t}^{txt} = (x_1^{txt}, ..., x_t^{txt})$ from the user's historical natural-language feedback, the state values in terms of the user's historical natural-language feedback are computed with $V(s_t, o_i) = Linear(Linear(s_t, x_i^{txt}))$, where $i \leq t$. The

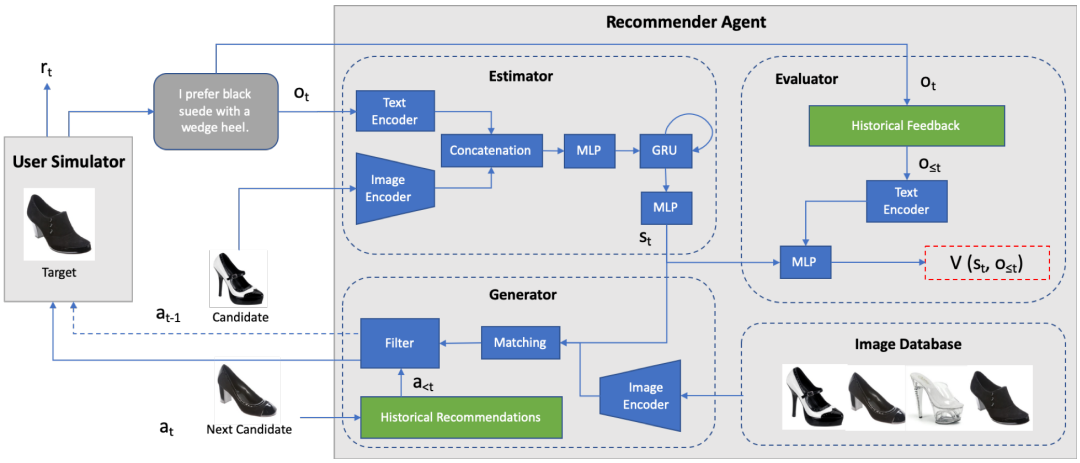


Figure 2: The proposed Estimator-Generator-Evaluator (EGE) model for dialog-based interactive recommendations.

final state value is computed using $V(s_t, o_{\leq t}) = Mean(V(s_t, o_i))$, where $i \leq t$ and $Mean()$ is the average function.

To summarise, in the EGE model architecture, we maintain the GRU for the state estimation in the Estimator and the KNNs for the candidate matching in the Generator as in the state-of-the-art RL-based approach [9], while we propose a Q-learning layer with the historical feedback as an Evaluator to optimise the quality of the estimated state.

3.3 The Learning Algorithm

In this work, we adopt a multi-task learning [8] approach for POMDP (inspired by [32] and [4]) to optimise the networks with a combination of a supervised learning classification loss and a Q-learning prediction loss.

Given an estimated state s_t, a target image representation $x^{img}_{+,0}$ (i.e. a positive sample) and several representations of randomly sampled images $x^{img}_{-,1}, ..., x^{img}_{-,J}$ (i.e. negative samples), the supervised training loss can be defined as the cross-entropy over the classification distribution:

$$L_s = -\log(\frac{e^{y_0}}{e^{y_0} + \sum_{j=1}^{J} e^{y_j}}) \tag{1}$$

where y denotes the L^2-norm: $y_0 = ||s_t - x^{img}_{+,0}||_2$ and $y_j = ||s_t - x^{img}_{-,j}||_2$. We define the RL loss for the training of the Estimator component based on one-step Temporal Difference (TD) error (i.e. $error = |V(s_t, o_{\leq t}) - (r_t + \gamma V(s_{t+1}, o_{\leq t+1}))|$) using a *Smooth L1 Loss*[1]:

$$L_q = \begin{cases} 0.5(V(s_t, o_{\leq t}) - (r_t + \gamma V(s_{t+1}, o_{\leq t+1})))^2, & \text{if } error < 1 \\ |V(s_t, o_{\leq t}) - (r_t + \gamma V(s_{t+1}, o_{\leq t+1}))| - 0.5, & \text{otherwise} \end{cases} \tag{2}$$

It is desired that the visual appearance of the recommended item becomes more similar to that of the desired item with increasing user interactions. Thus, at time t, given the recommended item a_t and the desired item representation $x^{img}_{+,0}$, we want to minimise the Euclidean distance. That is, we maximise the following visual reward: $r^{vis}_t = -||ResNet(a_t) - x^{img}_{+,0}||_2$. In addition, we expect that the desired item will be placed at higher ranks with more user interactions. Thus, we also model the *ranking percentile* [9] (i.e. the percentage of items with a rank lower than the target item among all items) as a reward r^{per}_t in terms of ranking. We define the reward r_t at time t as $r_t = \alpha r^{vis}_t + (1-\alpha) r^{per}_t$, where $\alpha \in [0, 1]$ is a reward weighting factor.

We jointly train the supervised loss and the RL loss by taking the latter one as a regulariser to introduce reward-driven properties to the recommendation process, in a similar manner to [32]:

$$L_{EGE} = L_s + L_q \tag{3}$$

To train our proposed EGE model, we adopt a user simulator [9] as a surrogate for real human users in the training processes. Further details about the used user simulator are provided in Section 4.3. When we start to train the proposed framework, the network parameters are randomly initialised. To facilitate an efficient exploration during the following reinforcement learning process, we first pre-train the model with a triplet loss objective, L_{tri}, similar to [9]:

$$L_{tri} = max(0, ||s_t - x^{img}_{+}||_2 - ||s_t - x^{img}_{-}||_2 + m) \tag{4}$$

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. Indeed, the rank of the target image can be improved compared to a random initialisation after the appropriate initial supervised learning process with L_{tri}. Based on the pre-trained model obtained with L_{tri}, the joint loss objective L_{EGE} can further ensure proximity between the target and candidate image representations (L_s), as well as maximise the expected future rewards (L_q), while applying smaller learning rates, resulting in better recommendation performances.

4 EXPERIMENTAL SETUP

In this section, we evaluate the effectiveness of our proposed EGE model for dialog-based interactive recommendations in comparison to the existing approaches from the literature. Figure 3 shows an example of a recommendation scenario to illustrate how the users can obtain their target items through interaction with the recommender system in the dialog-based interactive recommendation

[1] https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html

scenario. In particular, the recommender system ranks items based on the ranking scores (i.e. the similarities between the estimated preferences and all items), while the user gives feedback on the single top-ranked recommendation presented to them. Hence, effectiveness can be measured by the percentage of user sessions for which the target item is presented at the top rank by interaction turn M. Furthermore, it is possible that the user may view more of the ranking of items at each interaction turn, down to rank N. Therefore, we define success in the dialog-based interactive recommendation task as higher values in top-heavy metrics such as NDCG@N with a truncation at rank N calculated at the M-th turn, or Success Rate (SR) at the M-th turn.

In our experiments, we address three research questions, which are concerned with ascertaining how the Q-learning for POMDP (i.e. the Evaluator) can help the GRU (i.e. the Estimator) to better incorporate the users' accurate preferences over time with the partial observable preferences from the users' natural-language feedback, so as to make better recommendations with KNNs (i.e. the Generator). In particular, our three research questions relate to the Q-learning for POMDP, the historical information and the rewards in the EGE model – namely, how useful the Q-learning for POMDP is, how much historical information is required, and how the rewards are applied:

• RQ1: Can our proposed EGE model with Q-learning for POMDP outperform the existing state-of-the-art baseline models in the visually-grounded dialog-based interactive recommendation task?
• RQ2: What are the impacts of the historical information in the EGE model on its performance, such as the historical natural-language feedback and the historical recommendations?
• RQ3: What are the impacts of the reward-related hyper-parameters of the EGE model on its performance, such as the reward discount factor γ and the reward weighting factor α?

4.1 Datasets & Measures

Datasets. We perform the experiments on two datasets, namely the *Shoes*[2] dataset [2, 9] and the *Fashion IQ Dress*[3] [10]. In particular, the *Shoes* dataset has previously been used by [9] for a dialog-based interactive recommendation task, and we replicate their setup (relative captioner, user simulator, etc.). Indeed, both datasets are among the few that provide well pre-processed *relative captions* of image pairs that can be used for training and testing the user simulator, as well as the images of the fashion products for training and testing the recommendation models. On both datasets, we apply the same training and testing data split for all recommendation models. In the *Shoes* dataset, there are 10,751 relative captions (with one caption per pair of images about their visual differences) and 3,600 discriminative captions (with one caption per image about their discriminative visual features) for training a user simulator. The *Shoes* dataset also provides 10,000 images for training the recommender systems, and 4,658 images for testing. Meanwhile, in the *Fashion IQ Dress* dataset, there are 7,347 pairs of accessible images with two captions per pair. In particular, the relative captions of the 5,478 pairs from the *Fashion IQ Dress* dataset are used for training a user simulator, and the relative captions of the 1,869 pairs for testing. We also extract 7,182 unique images from the 5,478 pairs for training the recommender systems, and 2,454 unique images from the 1,869 pairs for testing. The recommendation models are evaluated when recommending target images from the test sets, starting from a randomly selected candidate image for the initial dialog turn. Each target image from the test sets represents a user session with the system. Moreover, following [9], as a user simulator, we adopt a *relative captioner* to simulate the user in generating natural-language feedback (described further in Section 4.3), which has been shown to mimic an actual user behaviour/feedback [9].

Metrics. The performances of the dialog-based interactive recommender systems are evaluated with metrics including Normalised Discounted Cumulative Gain (i.e. NDCG@N truncated at rank $N = \{5, 10\}$ calculated at the M-th interaction), Mean Reciprocal Rank (i.e. MRR@N truncated at rank $N = 10$ at the M-th interaction) and Success Rate (SR) at the M-th interaction. In particular, SR is the percentage of users for which the target image was retrieved within M turns among all the users with top-1 recommendation. We use all the evaluation metrics (i.e. NDCG@5, NDCG@10, MRR@10 and SR) at the 10th interaction turn for significance testing.

4.2 Baselines

We compare our proposed EGE model with two existing state-of-the-art baseline models:

• The sequential recommendation model [9] in a supervised-learning setup (which we denote as iGRU, where "i" stands for "interactive") is an approach where the recommender agent (with only a GRU and KNNs) is trained with a triplet loss [9] to maximise the short-term rewards. The iGRU model is close to the well-established GRU4Rec sequential recommendation model [14] but using the users' natural-language feedback and the previous recommended images as the input of the model with an online setup, instead of the logged clickthough data with an offline setup.
• The Model-Based Policy Improvement (MBPI) [9]) model is a RL-based approach where the recommender agent (with only a GRU and KNNs) is pre-trained with a triplet loss, and then further trained with a cross entropy loss. In the second training stage, the MBPI model is optimised by maximising the cumulative future rewards given a known environment (i.e. a user simulator). In particular, the MBPI model explores all possible recommendation trajectories in the future interaction turns with the help of the given user simulator and recommends the items with the maximum cumulative future rewards at each turn during this training process.

These two baseline models are the two existing representative formulations of the dialog-based interactive recommendation task for top-1 recommendation, which are formulated as a sequential modelling problem and a Markov decision process (MDP), respectively. Although there are a few other models with different formulations for the dialog-based interactive recommendation task – such as RCR [36] which is formulated as a constrained Markov decision process (CMDP) [1], the augmented cascading bandit (ACB) [34] or the sleeping pairwise ranking bandit (SPRB) [35], which are formulated as a multi-armed bandit (MAB) problem [29] – these models are not comparable with our scenario due to either requiring extra well-categorised visual attributes of items (RCR) or taking a *category* of the fashion products as the targets (ACB & SPRB).

[2] https://github.com/XiaoxiaoGuo/fashion-retrieval
[3] https://sites.google.com/view/cvcreative2020/fashion-iq

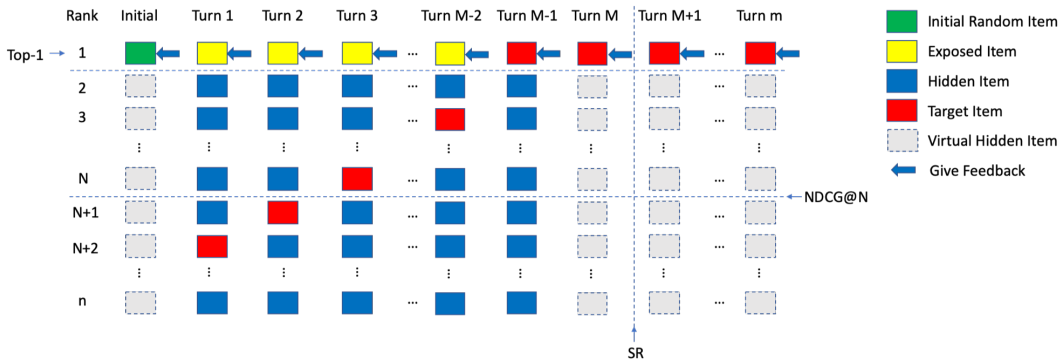


Figure 3: An example of the top-1 recommendation in the dialog-based interactive recommendation scenario.

4.3 Experimental Settings

User Simulator. To tackle the challenge of training an interactive recommender system online, we adopt a user simulator based on relative captioning [25] as in [9], which acts as a surrogate for real human users. The user simulator can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images. Such a natural-language feedback generation process with the user simulator is very similar to a scenario of a shopping conversation session between a shopping assistant and a customer. A user simulator with the *Shoes* dataset was intensively and carefully trained by [9] through crowdsourcing relative expressions about the visual differences of the image pairs and manually removing erroneous annotations. Furthermore, the pre-trained user simulator has previously been thoroughly evaluated via both a quantitative evaluation and a user study, thereby serving as a reasonable proxy for real users in our work. Going further than [9], we also train a user simulator for the *Fashion IQ Dress* dataset following [9]. The user simulator for the *Fashion IQ Dress* dataset is selected with the best prediction performance of the relative captioning task on the caption testing split. The pre-trained user simulators are used for both the training and evaluation of the interactive recommendation models.

Setup for Training. We first train our proposed EGE model with both user simulators on the *Shoes* and *Fashion IQ Dress* datasets, separately. The network parameters are randomly initialised. Following [9], we adopt a two-stage training process to facilitate the efficient exploration during the training with a joint loss L_{EGE}. At the initial stage (i.e. training with a triplet loss objective L_{tri}) and the second stage (i.e. training with a joint loss L_{EGE}) of training, we use Adam [18] as the optimiser on both datasets with an initial learning rate 10^{-3} and 10^{-5} [9, 36], respectively. The embedding dimensionality of the feature space is set to 256 and the batch size is 128, following the setting in [9]. For each batch, we train our model with 10 turns. We consider the top-11 nearest neighbours (considering an initial random item and 10 items during the 10 interactions) for removing the previously recommended items from the ranking list at each interaction with a post-filter, and we pick the top-1 from the post-filtered nearest neighbour list. The number of negative samples (i.e. J) is set at 5, which is considered as a reasonable number for negative sampling, regardless of the dataset size [24]. For our proposed EGE model, if not mentioned otherwise, the reward discount factor γ is set to 0.9 while the reward weighting factor α is set to 0.5 due to the EGE model's general good performances with $\gamma, \alpha \in [0, 1]$ on both datasets (as shown in Section 5.3).

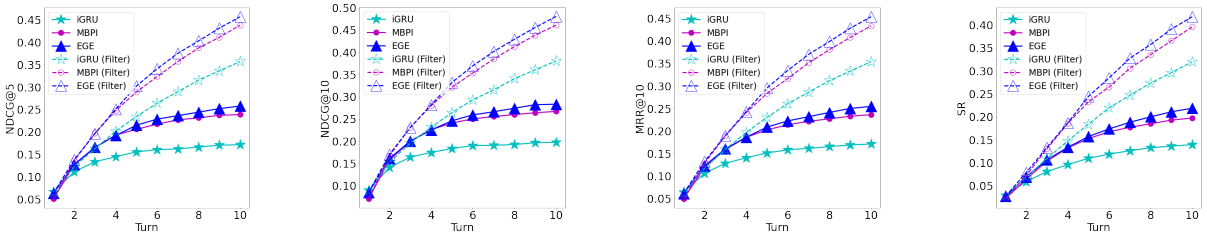
Setup for Evaluation. We consider the top-1 nearest neighbour (i.e. $K = 1$) as a recommendation at each interaction turn with or without a post-filter for testing. In particular, when a post-filter is applied, we pick the top-1 item from the post-filtered nearest neighbour list. For the evaluation metrics, we denote the interaction turn $M \in [1, 10]$. In particular, we mainly compare the performances of the tested models at the 10th turn (i.e. $M = 10$) with significance tests, which is the maximum interaction number in our study. This is smaller than the values adopted in [33, 36] and is more reasonable in the shopping scenario in that the users are more likely to be disappointed if they do not find their desired items after that many turns. If a user obtains the target item in less than 10 turns, we consider the ranking metrics (i.e. NDCG@5, NDCG@10 and MRR@10) for that user to be equal to one for all turns thereafter.

5 EXPERIMENTAL RESULTS

In this section, we analyse the experimental results with respect to the three research questions stated in Section 4, concerning the recommendation effectiveness of our proposed EGE model (Section 5.1), impact of the historical information including the historical natural-language feedback and historical recommendations (Section 5.2), and the impact of hyper-parameters related to the rewards (Section 5.3). We also demonstrate a use case from the logged experimental results to consolidate our findings (Section 5.4).

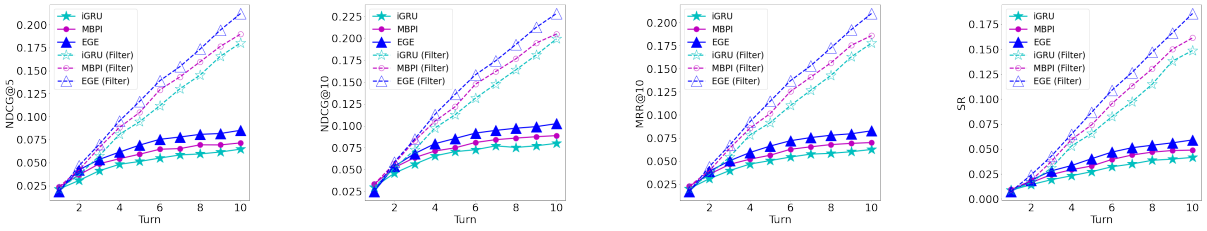
5.1 EGE vs. Baselines (RQ1)

Figures 4 and 5 show the recommendation effectiveness of our proposed EGE model and the existing state-of-the-art baseline models for top-1 recommendation in terms of NDCG@5 (Figure 4 (a) and Figure 5 (a)), NDCG@10 (Figure 4 (b) and Figure 5 (b)), MRR@10 (Figure 4 (c) and Figure 5 (c)) and Success Rate (SR) (Figure 4 (d) and Figure 5 (d)), while varying the number of interaction turns on the *Shoes* and *Fashion IQ Dress* datasets, respectively. The solid lines show the models' performances without a post-filter (which prevents the already recommended items from being recommended),



(a) NDCG@5 (b) NDCG@10 (c) MRR@10 (d) SR

Figure 4: Recommendation effectiveness at various interaction turns with top-1 recommendation on the *Shoes* dataset.



(a) NDCG@5 (b) NDCG@10 (c) MRR@10 (d) SR

Figure 5: Recommendation effectiveness at various interaction turns with top-1 recommendation on *Fashion IQ Dress*.

while the dash lines show performances when a post-filter is applied. When a post-filter is applied, the model is labeled with "(Filter)". Comparing the results in Figure 4 and Figure 5, we observe that our proposed EGE model generally achieves a better overall performance in terms of NDCG@5, NDCG@10, MRR@10 and SR at various interaction turns (except for the initial turn) without/with a post-filter, respectively. In the initial interaction turn, the performance of our proposed EGE model is marginally lower than the iGRU model and marginally higher than the MBPI model on the *Shoes* dataset, while it is marginally lower than the other two on the *Fashion IQ Dress* dataset. As the number of interaction turns increases ($\geq$ 2), the differences between the effectiveness of EGE and iGRU/MBPI on all metrics also increase. The better performance of EGE compared to iGRU can be attributed to the fact that our RL-based EGE model is optimised to maximise the long-term rewards with a Q-learning layer in the Evaluator, while the supervised learning approach (i.e. iGRU) aims to maximise the instant reward. Furthermore, by considering the historical information with Q-learning for POMDP, our proposed EGE model can also outperform the MBPI model with a better recommendation effectiveness, thereby mitigating the partial observation issue.

To quantify the improvements of our proposed EGE model compared to the other two baseline models, we measure their performances at the 10th interaction turn with top-1 recommendation. Table 1 shows the obtained recommendation performances of the models on the user simulator with a test set at the 10th interaction turn. For top-1 recommendation, we compare the performances of our proposed EGE model with the iGRU and MBPI models without/with a post-filter on both the *Shoes* and *Fashion IQ Dress* datasets, respectively. More specifically, Table 1 contains two groups of rows for each dataset. The first group of rows reports the effectiveness of the tested models without a post-filter and the improvements of EGE over the best baseline model. The second group of rows reports the performances of the tested models with a post-filter and shows the improvements, in the same way as the first group. The best overall performing results across the two groups of rows in the table are highlighted in bold in Table 1. * denotes a significant difference in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to EGE/EGE (Filter) in each group, respectively. Comparing the results in the first group of rows in the table, we observe that our proposed EGE model achieves better performances of $6-11\%$ and $15-23\%$ at the 10th turn than the best baseline model (i.e. MBPI) across all metrics without a post-filter on the *Shoes* and *Fashion IQ Dress* datasets, respectively, while achieving improvements of $4-6\%$ and $11-16\%$ with a post-filter, respectively. Indeed, the EGE model is significantly better than the iGRU and MBPI models without/with a post-filter for each metric at the 10th turn with top-1 recommendation, except for MBPI on the *Fashion IQ Dress* dataset in terms of MRR@10 and SR.

In answer to RQ1, the results demonstrate that our proposed EGE model can outperform the state-of-the-art baseline models (i.e. iGRU and MBPI) overall after the first interaction turn. In particular, it is significantly more effective than both the supervised-learning-based approach (i.e. iGRU) and the RL-based approach (i.e. MBPI) without/with a post-filter at the 10th interaction turn with top-1 recommendation. Therefore, our proposed EGE model with Q-learning for POMDP can effectively mitigate the partial observation issue.

5.2 Impact of Historical Information (RQ2)

To address RQ2, we investigate how much historical information is required in our model by considering the users' historical feedback and the agent's historical recommendations. In particular, recall from Section 3 that the users' historical feedback is used as the input of the Evaluator in Figure 2 to judge the quality of the estimated state, while the agent's historical recommendations are used in a post-filter to remove the recommended items from the recommendation list. In summary, compared to the MBPI model, our

Table 1: Recommendation effectiveness of our proposed EGE model and the baseline models at the 10th turn on both the *Shoes* and *Fashion IQ Dress* datasets. % Improv. indicates the improvements by EGE/EGE (Filter) over the best baseline model. The best overall results are highlighted in bold. * denotes a significant difference in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to EGE/EGE (Filter) in each group, respectively.

Models	Post-Filter Applied	Shoes				Fashion IQ Dress			
		NDCG@5	NDCG@10	MRR@10	SR	NDCG@5	NDCG@10	MRR@10	SR
iGRU	No	0.1717*	0.1975*	0.1712*	0.1398*	0.0647*	0.0800*	0.0627*	0.0416*
MBPI	No	0.2389*	0.2671*	0.2363*	0.1977*	0.0715*	0.0888*	0.0702	0.0489
EGE	No	**0.2580**	**0.2834**	**0.2547**	**0.2190**	**0.0852**	**0.1025**	**0.0829**	**0.0591**
% Improv.	-	8.00	6.10	7.79	10.77	19.16	15.43	18.09	22.87
iGRU (Filter)	Yes	0.3574*	0.3807*	0.3544*	0.3201*	0.1802*	0.1994*	0.1777*	0.1487*
MBPI (Filter)	Yes	0.4384*	0.4613*	0.4338*	0.3961*	0.1898*	0.2050*	0.1859*	0.1614*
EGE (Filter)	Yes	**0.4572**	**0.4807**	**0.4541**	**0.4182**	**0.2122**	**0.2284**	**0.2098**	**0.1858**
% Improv.	-	4.29	4.21	4.68	5.58	11.80	11.41	12.86	15.11

EGE model takes the users' historical feedback into consideration during the training process.

Considering the usefulness of the users' historical natural-language feedback, we compare the effectiveness of our proposed EGE model with the MBPI model on both used datasets. In both Figure 4 and Figure 5, we observe that EGE consistently outperforms MBPI in terms of NDCG@5, NDCG@10, MRR@10 and Success Rate through the 2nd turn to the 10th turn. In Table 1, we also observe that the EGE model is consistently and significantly better than the MBPI model in each group for each metric at the 10th interaction turn for top-1 recommendation, except for MBPI on the *Fashion IQ Dress* dataset in terms of MRR@10 and SR. This suggests, as we argued in Section 2, that adopting the users' historical feedback in the judgement of the estimated states can benefit the interactive recommendation model.

Furthermore, we compare the performances of all the tested models (including our proposed EGE model) considering the usefulness of the agent's historical recommendations with a post-filter. In Figure 4 and Figure 5, we observe that all the tested models that apply a post-filter can consistently outperform those without a post-filter after the initial interaction turn. There is a trend that the gap between a model with a post-filter and the one without a post-filter increases at every interaction turn. Thus, this trend indicates that applying the post-filter on the recommendation list using the agent's historical recommendations demonstrates a *cumulative effect*. This suggests that applying a post-filter with the agent's historical recommendation can generally further improve the performances of the interactive recommendation models.

Overall, in response to research question RQ2, we find that our proposed EGE model can benefit from both the users' historical feedback and the agent's historical recommendations.

5.3 Impact of Hyper-Parameters (RQ3)

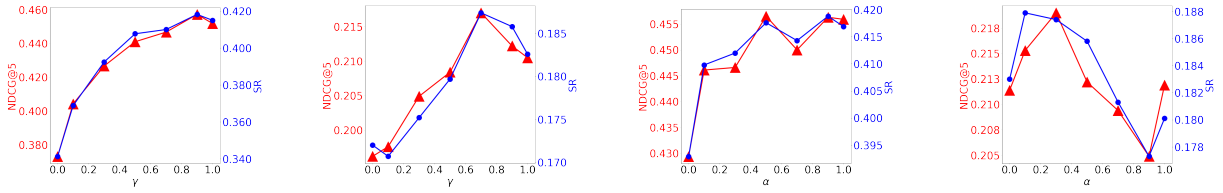
To address RQ3, Figure 6 depicts the effects of the *reward discount factor* γ and the *reward weighting factor* α on our proposed EGE model with a post-filter (i.e. EGE (Filter)) in top-1 recommendation on the *Shoes* & *Fashion IQ Dress* datasets.

Effect of the reward discount factor γ. Figure 6 (a) illustrates the NDCG@5 and SR of EGE (Filter) at the 10th turn in top-1 recommendation with different reward discount factors on the *Shoes* dataset. In particular, $\gamma = 0$ means that the models only consider immediate feedback, while $\gamma = 1$ means that the model weights all future rewards equally. We can see that the performance of EGE (Filter) improves when the reward discount factor γ increases from 0, except for $\gamma = 1$. Figure 6 (b) demonstrates a similar increasing trend on the *Fashion IQ Dress* dataset and both metrics reach a peak at $\gamma = 0.7$. The generally better performance of the model with $\gamma > 0.1$ than the model with $\gamma = 0$ leads to the conclusion that the Evaluator component does help to improve the overall recommendation effectiveness by considering long-term rewards. On the other hand, the decreased performance of the model with $\gamma = 1$ on the *Shoes* dataset and $\gamma > 0.7$ on the *Fashion IQ Dress* dataset shows that the reward discount factor should be set appropriately.

Effect of the reward weighting factor α. Figure 6 (c) illustrates the NDCG@5 and SR of EGE (Filter) at the 10th turn in top-1 recommendation on the *Shoes* dataset with different reward weighting factors α, which weight the contributions of the visual reward r_t^{vis} and the ranking percentile reward r_t^{per} to the final rewards. In particular, $\alpha = 0$ means that the model only considers the ranking percentile reward r_t^{per}, while $\alpha = 1$ means that the model only takes the visual reward r_t^{vis} into consideration. We can see that the performance of EGE (Filter) improves when the reward weigthing factor α increases from 0 to 0.5, and varies slightly with $\alpha > 0.5$. Figure 6 (d) shows a distinctive trend on the *Fashion IQ Dress* dataset in that both the NDCG@5 and SR metrics first increase and reach a peak at $\alpha = 0.3$ and $\alpha = 0.1$, respectively, and then decease when α increases from 0.3 to 0.9. This trend shows that the visual reward r_t^{vis} is more informative than the ranking percentile reward r_t^{per} in the EGE (Filter) model on the *Shoes* dataset, while the ranking percentile reward r_t^{per} is more important than the visual reward r_t^{vis} on the *Fashion IQ Dress* dataset. Such a difference can be attributed to a domain factor from the datasets in that the images from the *Fashion IQ Dress* dataset usually include a human model to display the clothing while the images from the *Shoes* dataset only contain shoes without a model (as can be observed in the image databases for shoes[4] and dresses[5]). The visual features of the human models can confuse the ResNet component when mapping the dress images

[4] http://tamaraberg.com/attributesDataset/attributedata.tar.gz

[5] https://github.com/hongwang600/fashion-iq-metadata/blob/master/image_url



(a) γ for *Shoes* (b) γ for *Fashion IQ Dress* (c) α for *Shoes* (d) α for *Fashion IQ Dress*

Figure 6: Effects of (a) & (b) the *reward discount factor* γ and (c) & (d) the *reward weighting factor* α at the 10th turn in the top-1 recommendation scenario on the *Shoes* and *Fashion IQ Dress* datasets.

to the image feature (ResNet) space. Therefore, the generated dress image embeddings may be affected by the noises from the visual features of the human models, thereby reducing the utility of the visual rewards from the user simulator. To mitigate this issue, our future work will consider more advanced models [22, 31] that aim for effective fashion attribute detection for generating the dress image embeddings.

Overall, in response to RQ3, we find that the ranking percentile reward r_t^{per} and the visual reward r_t^{vis} can help our EGE model to improve the recommendation performance.

5.4 A Use Case

To consolidate the results observed in the above sections, we present a use case of the tested models without/with a post-filter in Figure 7: (a) iGRU, (b) iGRU (Filter), (c) MBPI, (d) MBPI (Filter), (e) EGE, (f) EGE (Filter) only on the *Shoes* dataset for top-1 recommendation. In Figure 7 (a-f), the first image is the target item desired by the user (labeled with "Target"), while the second image (labeled with "Initial") is the initial recommendation proposed by the recommender system randomly. For a fair comparison, the initial images are the same across the tested recommender systems given the target image from the testing set. Then, the recommended top-1 items and the user comments in the following turns are presented. The rank of the target item is also presented above the images at each turn (e.g. "Turn 1 (rank=13)" in Figure 7 (a) iGRU, where the target image is ranked at the 13th position in the recommendation list at the first interaction turn). When the target item is recommended, the rank is 1 (e.g. "Turn 5 (rank=1)" in Figure 7 (b)), and the user simulator will give the comment: "are the same". We observe that our proposed EGE model is the most effective recommender system among the tested models. Both EGE and EGE (Filter) only need two interactions to display the desired item, while the other tested models require at least 4 interactions given the same target and initial items. For instance, iGRU fails to recommend the target item within 5 interaction turns and recommends the same items repeatedly, even though the rank of the target item is getting higher. Though MBPI is more effective than iGRU, there is also a repeated recommendation at the 4th interaction turn when the post-filter is not applied. Furthermore, we also observe that the ranks of the target item with iGRU/iGRU (Filter) are much higher than the ranks of MBPI/MBPI (Filter) and EGE/EGE (Filter) at the first interaction turn. One possible reason is that the iGRU model is maximising the instant reward while the RL-based models are maximising the future accumulative rewards. In addition, our proposed EGE model is more effective at making use of the user's natural-language feedback, i.e. "are red shiny high heels". However, both the iGRU (Filter) and MBPI (Filter) models continuously present items that violate the previous user's feedback. Indeed, iGRU (Filter) recommends the red sport shoes that are violated from the "high heels", while MBPI (Filter) recommends the black high heels that are contrary to the "red" colour. Note that a use case on the *Fashion IQ Dress* dataset also led to similar results and observations. We omit their reporting in this paper because of space constraints.

6 CONCLUSIONS

In this paper, we proposed a novel dialog-based recommendation model, denoted by the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP to effectively incorporate the users' preferences over time in a partially observable environment. Specifically, we leveraged an Estimator to track and estimate the users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations (with a post-filter to remove repeated recommendations), and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. Following previous work, we trained our EGE model by using a user simulator, which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Our experiments on the *Shoes* and *Fashion IQ Dress* datasets demonstrated that our proposed EGE model achieves significantly enhanced performances compared to the strongest baseline model (i.e. MBPI) – for instance, improving by $6-23\%$ when a post-filter is not used, and $4-16\%$ when post-filtering is applied, respectively. Our reported results also showed that the EGE model can benefit from the historical information (i.e. the users' historical feedback and the agent's historical recommendations). For future work, we plan to investigate an end-to-end model with historical recommendation by integrating the features of the recommended items into the preferences estimation process. We also plan to adopt more advanced neural models [22, 31] that are more effective in fashion attribute detection when generating the dress image embeddings, thereby mitigating the impact of the noisy features extracted from the human models.

ACKNOWLEDGMENTS

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

(a) iGRU

(b) iGRU (Filter)

(c) MBPI

(d) MBPI (Filter)

(e) EGE

(f) EGE (Filter)

Figure 7: A use case with different recommendation models on the *Shoes* dataset.

REFERENCES

[1] Eitan Altman. 1999. *Constrained Markov decision processes.* CRC Press.

[2] Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proc. ECCV.* 663–676.

[3] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proc. WSDM.* 456–464.

[4] Minmin Chen, Bo Chang, Can Xu, and Ed H Chi. 2021. User Response Models to Improve a REINFORCE Recommender System. In *Proc. WSDM.* 121–129.

[5] 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).

[6] Tanmay Gangwani, Joel Lehman, Qiang Liu, and Jian Peng. 2020. Learning belief representations for imitation learning in pomdps. In *Proc. UAI.* 1061–1071.

[7] 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).

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

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

[10] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019. Fashion IQ: A New Dataset towards Retrieving Images by Natural Language Feedback. *arXiv preprint arXiv:1905.12794* (2019).

[11] Hado Hasselt. 2010. Double Q-learning. *Proc. NeurIPS* (2010), 2613–2621.

[12] Matthew Hausknecht and Peter Stone. 2015. Deep recurrent Q-learning for partially observable mdps. In *Proc. AAAI.* 29–37.

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

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. *Proc. ICLR* (2016).

[15] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. 2018. Deep variational reinforcement learning for POMDPs. In *Proc. ICML.* 2117–2126.

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

[17] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proc. ICDM.* 197–206.

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

[19] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Proc. NeurIPS.* 1008–1014.

[20] 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.

[21] Yu Lei and Wenjie Li. 2019. Interactive recommendation with user-specific deep reinforcement learning. *TKDD* (2019), 1–15.

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR.* 1096–1104.

[23] Zhongqi Lu and Qiang Yang. 2016. Partially observable markov decision process for recommender systems. *arXiv preprint arXiv:1608.07793* (2016).

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NeurIPS.* 3111–3119.

[25] 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.

[26] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. 2005. An MDP-based recommender system. *JMLR* (2005).

[27] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *Proc. ICML.* 387–395.

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

[29] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[30] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proc. WSDM.* 565–573.

[31] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proc. CVPR.* 4271–4280.

[32] Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. 2020. Self-Supervised Reinforcement Learning for Recommender Systems. In *Proc. SIGIR.* 931–940.

[33] 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.

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

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

[36] 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.

[37] 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.

[38] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proc. WWW.* 167–176.

[39] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proc. KDD.* 2810–2818.

[40] 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.