Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback

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

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

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

ABSTRACT

Interactive recommendation enables users to provide verbal and non-verbal relevance feedback (such as natural-language critiques and likes/dislikes) when viewing a ranked list of recommendations (such as images of fashion products), in order to guide the recommender system towards their desired items (i.e. goals) across multiple interaction turns. Such a multi-modal interactive recommendation (MMIR) task has been successfully formulated with deep reinforcement learning (DRL) algorithms by simulating the interactions between an environment (i.e. a user) and an agent (i.e. a recommender system). However, it is typically challenging and unstable to optimise the agent to improve the recommendation quality associated with implicit learning of multi-modal representations in an end-to-end fashion in DRL. This is known as the coupling of policy optimisation and representation learning. To address this coupling issue, we propose a novel goal-oriented multi-modal interactive recommendation model (GOMMIR) that uses both verbal and non-verbal relevance feedback to effectively incorporate the users' preferences over time. Specifically, our GOMMIR model employs a multi-task learning approach to explicitly learn the multi-modal representations using a multi-modal composition network when optimising the recommendation agent. Moreover, we formulate the MMIR task using goal-oriented reinforcement learning and enhance the optimisation objective by leveraging non-verbal relevance feedback for hard negative sampling and providing extra goal-oriented rewards to effectively optimise the recommendation agent. Following previous work, we train and evaluate our GOMMIR model by using user simulators that can generate natural-language feedback about the recommendations as a surrogate for real human users. Experiments conducted on four well-known fashion datasets demonstrate that our proposed GOMMIR 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, multi-modal, reinforcement learning, relevance feedback

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM Reference Format:
Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2023. Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3604915.3608775

1 INTRODUCTION

Interactive recommendation is a type of interactive information-seeking task [13, 17, 24, 28, 56], which aims to satisfy the users' dynamic information needs by interactively and continuously collecting the users' verbal (such as natural-language critiques) and non-verbal (such as likes/dislikes) feedback in relation to the system's recommendations. In particular, multi-modal interactive recommendation (MMIR) involves information with various modalities, such as natural language and images. In a multi-modal interactive recommendation scenario, users can express their preferences in natural language, and indicate their positive/negative opinions by clicking like/dislike buttons when viewing a ranked list of visual recommendations (such as images of fashion products). Figure 1 shows an example of multi-modal interactive recommendation with both verbal and non-verbal relevance feedback. In this use case, the user indicates the particularly liked item image(s) among the top-K (e.g., $K = 3$) recommended items and provides a natural-language critique at each interaction turn to obtain items with better preferred features, while tagging the other recommendations with a "dislike" if they are less relevant to the user's preferences. Such a multi-modal interactive recommendation task is inherently a "goal-oriented" information-seeking process when a user seeks a target item (i.e. a visual goal) and gives natural-language feedback using the user's preferred features (i.e. textual goals) across multiple interactions.

Interactive recommendation tasks have been typically formulated using deep reinforcement learning (DRL) approaches [14, 21, 29, 30, 39, 45, 53, 57]. Indeed, such approaches have demonstrated an ability to capture the users' preferences and to maximise the expected long-term cumulative rewards (such as fewer efforts/interactions to find the desired items [19, 43]) when deciding what items to recommend to the users (i.e. the environment) at each interaction turn. However, it is typically challenging to learn an effective multi-modal interactive recommendation agent due to the so-called "coupling" of the policy optimisation (for improving the quality of the recommendations) and representation learning (for understanding the visual and textual information) [16]. In particular, prior research often found that learning representations in an end-to-end fashion in DRL is usually unstable [26, 51] due to the coupling issue. Indeed, the policy optimisation processes of the existing DRL-based interactive recommendation models are

Figure 1: An example of multi-modal interactive recommendation with both verbal and non-verbal relevance feedback.

associated with an implicit multi-modal representation learning of discrete actions (i.e. the visual items), relevance feedback (i.e. the natural-language critiques), and their composition of representations (i.e. the estimated preferences). Such implicit multi-modal representation learning cannot guarantee good multi-modal representations, yet the DRL algorithms require good representations to drive the policy learning in a MMIR task. In particular, a simple concatenation operation [19, 43, 45] for multi-modal feature composition between text (encoded with GloVe [36] or BERT [15]) and image (encoded with ResNet [20]) representations does not provide an effective understanding of the users' current information needs at each turn. In addition, more advanced feature composition approaches for combining image and text features (such as Text Image Residual Gating (TIRG) [41] and CLIP for Conditioned image retrieval (CLIP4Cir) [2, 3]) have been recently proposed by various text-image retrieval models [10, 18, 41]. We propose to leverage such approaches as an extra multi-modal composition representation learning task using multi-task learning [27] for decoupling the representation learning from the policy optimisation in the MMIR task.

Along with the coupling issue, an appropriate optimisation objective for learning what to recommend at the next turn is typically important for improving the effectiveness of the interactive recommendation agents [1, 9, 48]. However, the recommendation policy optimisation functions adopted by existing interactive recommendation agents [19, 45, 57] are mainly based on both (1) a sampled softmax [8] with randomly sampled negatives from the whole candidate pool [19, 45], and (2) an uninformative reward function that considers only the critiqued items [19, 45, 57] and/or a sparse reward function defined as a binary credit (success or fail) for reaching the desired item [57]. Due to the "goal-oriented" nature of the multi-modal interactive recommendation task, goal-oriented reinforcement learning (GORL) [11, 33] can be easily adapted to the MMIR task with a goal-oriented policy optimisation function that allows the agents to pursue their own *goals* (i.e. the users' desired items or the users' critiques for acquiring their desired items) and to learn to achieve their goals via goal-oriented rewards. In the multi-modal interactive recommendation task, goals are both the users' target item (i.e. the visual goal) and the corresponding natural-language critiques (i.e. the textual goals) in the multi-turn interactions. These rewards can be formulated by using a distance measure between the achieved textual goals and the desired visual goal without any domain knowledge [33]. In this paper, we leverage a goal-oriented policy optimisation function with hard negative samples obtained iteratively from the disliked items across multiple interaction turns, as well as more informative rewards by measuring the similarities between the retrieved top-K item images (according to the estimated preferences at each turn) and the user's target item image. In addition, the critiqued items and the corresponding natural-language critiques (the textual goals) are collectively taken as the inputs of the interactive recommendation agent for estimating the users' preferences over time.

In this paper, we propose a novel goal-oriented multi-modal interactive recommendation (GOMMIR) model for addressing the so-called "coupling" issue, to use both verbal and non-verbal relevance feedback to effectively incorporate the users' preferences over time. In particular, we formulate the MMIR task with goal-oriented reinforcement learning [33] based on a policy gradient method (i.e. REINFORCE [8]) to effectively optimise the recommendation policy using hard negative sampling and goal-oriented rewards for pursuing the textual and visual goals. Different from the existing models, our proposed GOMMIR model adopts a recent unified multi-modal vision and language model (i.e. CLIP) for image and text encoding, as well as a Text Image Residual Gating (TIRG) [41] component for multi-modal feature composition to better understand the users' current information needs at each turn. For the training of our model, we adopt a multi-task learning [27] approach that jointly leverages both a deep reinforcement learning objective for improving the recommendation quality and a supervised learning objective for explicitly learning the multi-modal composition representations. Following previous work [19, 43, 45], we train and evaluate our proposed GOMMIR model by using user simulators that can generate natural-language critiques about the recommendations as a surrogate for real human users. Experiments conducted on four well-known fashion datasets (Shoes, Dresses, Shirts, and Tops & Tees) demonstrate that our proposed model yields significant improvements in comparison to the existing state-of-the-art baseline models. The main contributions of this paper are summarised as follows:

- We propose a goal-oriented multi-modal interactive recommendation (GOMMIR) model for addressing the coupling issue of policy optimisation and representation learning from both the users' verbal and non-verbal relevance feedback. Our model adopts an advanced multi-modal composition model (i.e. TIRG) and a multi-task learning approach to explicitly learn the multi-modal composition representations during the recommendation policy optimisation process using goal-oriented reinforcement learning.
- The GOMMIR model leverages verbal relevance feedback as textual sub-goals and adopts non-verbal relevance feedback for hard negative sampling and the extra visual rewards.
- An extensive empirical evaluation is performed on the multi-modal interactive recommendation task, demonstrating significant improvements with GOMMIR over existing state-of-the-art approaches.

2 RELATED WORK

In this section, we first introduce multi-modal interactive recommendation. Then, we describe goal-oriented reinforcement learning. Next, we discuss the use of verbal and non-verbal relevance feedback in recommendation.

Multi-Modal Interactive Recommendation. Recently, multi-modal interactive recommendation has been intensively investigated in the literature, as it can satisfy the users' information needs by effectively eliciting the users' preferences from the visual recommendations (e.g., images of fashion products) and the corresponding verbal

and/or non-verbal relevance feedback (e.g., natural-language feedback and likes/dislikes) [7, 12, 19, 31, 45–47, 53]. These kinds of interactive recommendations are suited for taste-oriented domains such as fashion, where search-type interaction methods are less useful. Typically, the multi-modal interactive recommendation task focuses on tracking and estimating the users' preferences over time with a state tracker, such as a gated recurrent unit (GRU) [19, 45], a long short-term memory (LSTM) [57], a Transformer encoder [43, 47], or an RNN-enhanced Transformer [46], in an end-to-end fashion with supervised learning (SL) and/or deep reinforcement learning (DRL) approaches. The representations of visual candidate items and natural-language feedback are initially generated with pre-trained models (such as ResNet for image encoding and BERT or GloVe for text encoding), and are then implicitly further tuned along with the recommendation policy optimisation. Most existing multi-modal interactive recommendation models adopt a simple concatenation operation for feature composition. However, learning representations in an end-to-end fashion in DRL is usually unstable [16, 26, 51] due to the previously mentioned coupling issue of policy optimisation and representation learning. Meanwhile, the DRL algorithms require good representations to drive the policy learning in a multi-modal interactive recommendation task. This so-called coupling issue has not been fully explored in the multi-modal interactive recommendation scenario.

Goal-Oriented Reinforcement Learning. Deep reinforcement learning has been widely adopted in recommender systems in order to improve the quality of the recommendations while maximising the users' long-term satisfaction and engagement. Typically, the multi-modal interactive recommendation task has been modelled with reinforcement learning (RL) and formulated as Markov decision processes (MDPs) [19], partially observable Markov decision processes (POMDPs) [45], constrained Markov decision processes (CMDPs) [57] or multi-armed bandits [53] so as to effectively incorporate the users' information needs across multiple turns. However, the policy optimisation adopted by existing interactive recommendation agents [19, 45, 57] is generally ineffective due to random negative sampling [19, 45] and sparse/non-informative rewards (as discussed in Section 1). Compared to the standard RL algorithms that learn a policy solely based on the states or observations, goal-oriented reinforcement learning (GORL) additionally requires the agent to make decisions according to different goals [33]. A goal is defined as "a cognitive representation of a future object" [11], which the agent is committed to achieve or maintain. The goal-oriented reinforcement learning approaches have been shown to improve training sample efficiency by learning from self-generated rewards (i.e. intrinsic rewards) when the external rewards are sparse. For example, Wang et al. [42] proposed a novel model-based model, GoalRec, based on a Dueling Deep Q-Network (DDQN), by designing a disentangled universal value function with the users' desired future trajectory (i.e. goal). In addition, Zhao et al. [60] proposed a novel multi-goals abstraction-based deep hierarchical reinforcement learning algorithm (MaHRL) to generate multiple goals with the high-level agent so as to reduce the difficulty for the low-level agent to approach the high-level goals. The high-level agent catches long-term sparse conversion signals, while the low-level agent captures short-term click signals. However, these existing formulations of recommendation agents with GORL are not suitable for the MMIR task where there is neither a desired future trajectory nor any conversion signals that can be leveraged as a goal or to learn high-level goals. Indeed, to the best of our knowledge, goal-oriented reinforcement learning has not yet been explicitly formulated with the MMIR scenario, which has both visual and textual goals for optimising the recommendation policy.

Relevance Feedback in Recommendation. Relevance feedback provides indications about whether the shown recommendations are relevant to the user's current preferences. Both verbal (e.g., natural-language feedback) and non-verbal (e.g., likes/dislikes, clicks, and skips) relevance feedback have been intensively investigated in the recommendation field [4, 12, 21, 61]. In particular, non-verbal relevance feedback is often used to model the users' behaviours and to indicate their preferences. For instance, Zhao et al. [61] proposed the DEERS model with a Deep Q-Network (DQN) to automatically learn the optimal recommendation strategies through the incorporation of positive (such as purchases) and negative (such as skips) feedback for sequential recommendations. In addition, natural-language feedback has been shown to be more informative about the users' preferences in comparison to non-verbal relevance feedback (e.g., ratings and clicks) [17, 24]. For instance, existing conversational recommendation models either allow the users to describe their preferred attributes as positive feedback [19, 21, 39, 43, 52, 53, 59] (e.g., "I prefer dresses with longer sleeves.") or to provide disliked attributes as negative feedback [47] (e.g., "I dislike shoes with high heels."). In addition, the users can also answer some attribute-level clarification questions (e.g., "Do you like a red colour?") with a binary yes/no response, while rejecting the undesired item-level recommendations [6, 29, 50]. In this paper, we consider both verbal (e.g., natural-language critiques) and non-verbal (e.g., likes/dislikes) relevance feedback from the user's multi-turn interactions to incorporate their preferences in the MMIR task.

We particularly argue that the existing multi-modal recommendation models [19, 46, 57] have not effectively addressed the coupling issue of the policy optimisation and representation learning from both the verbal and non-verbal relevance feedback. Such an issue limits these models' ability at incorporating the users' preferences over time. Our proposed GOMMIR model aims to address the coupling issue by adopting an advanced multi-modal composition model (such as TIRG [41]) and a multi-task learning approach to explicitly learn the multi-modal composition representations during the recommendation policy optimisation process driven by a goal-oriented reinforcement learning.

3 THE GOMMIR MODEL

In this section, we first formulate the problem of the MMIR task via DRL using goal-oriented partially observable Markov decision processes (GO-POMDP) and introduce our notations. Next, in Section 3.2, we propose a novel goal-oriented multi-modal interactive recommendation (GOMMIR) model to effectively incorporate the users' preferences over time with both verbal and non-verbal relevance feedback. Finally, we define the negative sampling and rewards that are suitable for this MMIR scenario (Section 3.3).

(a) Traditional RL with a MDP/POMDP

(b) GO-POMDP for MMIR

Figure 2: Traditional RL with a MDP/POMDP [22] and GO-POMDP for MMIR.

3.1 Preliminaries

3.1.1 GO-POMDP for MMIR. Figure 2 (a) shows the traditional RL as a Markov decision process (MDP) or a partially observable Markov decision process (MDP) in formulating interactive/sequential recommendations [1, 9, 22, 32]. In this scenario, the users' interactions with the recommended items (actions) are returned as feedback (the so-called observations from the environments, such as views, clicks, skips, purchases, and ratings) to the recommendation agents, which usually convert the users' feedback into a reward signal [22]. The scalar values of the rewards vary based on the different types of feedback (e.g., purchases have high rewards and skips have low rewards). The aim of traditional RL with a MDP/POMDP is to optimise the recommendation agents by maximising the cumulative rewards across the multiple interaction turns. On the other hand, Figure 2 (b) illustrates a goal-oriented partially observable Markov decision process (GO-POMDP) for the MMIR task. Different from the traditional RL with MDPs, the rewards are calculated based on the distances/similarities between the actions (the recommended items) and the goal (the target item). The goal can be either fully represented with an image as a visual goal or partially represented with a natural-language sentence as a textual goal. In particular, users can provide natural-language feedback (critiques), which typically only partially express their preferences [45], by eliciting the missing attributes of the target item (goal) compared to the recommendation items (actions). To this end, the users' natural-language feedback (critiques) can be seen both as an integral part of the environment observations, as well as textual goals towards the users' desired item. The aim of GO-POMDP is to guide the recommendation agents towards the goals (both the textual goals with the critiques and the visual goal with the target item) by taking the critiques (textual goals) as a part of the inputs to the recommendation agents and achieving the maximum cumulative distance-based/similarity-based rewards. Here, we mainly focus on goals in terms of visual features with images and textual inputs due to the limitations of the available datasets. Indeed, we believe that our formulation with GO-POMDP can also be generalised with goals in terms of other non-visual features, such as brands, prices, and functionalities. We leave this as an interesting future work.

3.1.2 Notations. Specifically, we formulate the multi-modal interactive recommendation (MMIR) task as a goal-oriented partially observable Markov decision process (GO-POMDP) with a tuple of seven elements $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{G}, r, \gamma)$ to describe the multi-modal interactive recommendation process, where: $\mathcal{S}$ is a continuous *state* space to describe the user states; $\mathcal{A}$ is a discrete *action* space that contains candidate items for recommendation; $\mathcal{O}$ is a set of *observations*, which are the users' verbal (e.g., the natural-language critiques) and non-verbal (e.g., likes/dislikes) relevance feedback; $\mathcal{T}$ is a set of conditional transition probabilities between states; $\mathcal{G}$ is a set of visual goals (i.e. the users' target items); $R \in \mathbb{R}$ is the *reward function*, where $r(s, a, g)$ is the immediate reward obtained from a user with a desired goal $g \in \mathcal{G}$ by performing action $a \in \mathcal{A}$ at user state $s \in \mathcal{S}$; $\gamma \in [0, 1]$ is the *discount factor* for future rewards.

Figure 3 shows the goal-oriented interactive recommendation process with both verbal and non-verbal relevance feedback for top-K recommendations. During the interaction process (with an initial state s_0), the recommender system suggests a ranking of top-K items $(a_{t,\leq K} = (a_{t,1}, ..., a_{t,K}) \in \mathcal{A})$ at each turn t. Meanwhile, the user provides non-verbal relevance feedback (e.g., likes/dislikes) and gives natural-language feedback ($o_t \in \mathcal{O}$) in terms of the liked item(s) among the current top-K recommendations $a_{t,\leq K}$ by describing the desired features that the current recommended item(s) lack. In this goal-oriented seeking process, we assume that the user gives natural-language feedback on the recommended item that is the most similar item to their perceived target item. Then, the recommender system collects both the top-K recommendations $a_{t,\leq K}$ and the corresponding relevance feedback o_t to track/estimate the user's preferences according to the transition distribution, $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, o_t, a_{t,\leq K})$. The recommender system takes actions according to its policy $\pi(a_{t+1,\leq K}|s_{t+1})$, which returns the probability of taking action $a_{t+1,\leq K}$ at turn $t+1$. Hence, the interactive recommendation process decomposes the long-term, hard-reaching goals (i.e. the users' desired items g) into easily obtained sub-goals expressed by the users' natural-language critiques o_t (i.e. the textual goals).

3.2 The Model Architecture

Figure 4 shows our proposed GOMMIR model for multi-modal interactive recommendations. In particular, we leverage a pre-processing stage for identifying the critiqued items with the non-verbal relevance feedback (i.e. likes/dislikes), a multi-modal encoding stage for extracting textual and visual representations, a composition stage for multi-modal feature composition, a state tracking stage for tracking/estimating the users' preferences over time, and a ranking stage for recommending visual items.

Pre-processing Stage. The goal of the pre-processing step is to identify the critiqued item(s) from the non-verbal relevance feedback (i.e. likes and dislikes), to infer the index numbers of the liked item(s) (i.e. $a_{t,u}$, where $u \in [1, K]$) and the disliked items (i.e. $a_{t,d}$, where $d \in [1, K]$) among the recommendation list $a_{t\leq K}$. The identified liked item(s) are then passed to the subsequent text and image encoders for extracting features, while the disliked items are stored in the set of negative feedback history. The negative feedback history with the disliked items is used as hard negative samples for model optimisation, as described in Section 3.3.

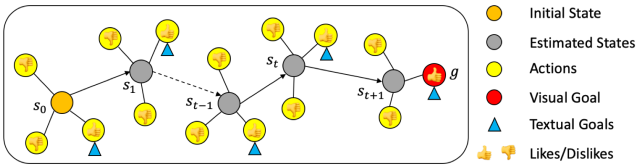


Figure 3: The goal-oriented interactive recommendation process with verbal & non-verbal relevance feedback for top-K recommendations.

Multi-Modal Encoding Stage. To represent the textual content related to the users' preferences, both the users' natural-language feedback and the recommender system's visual recommendations are encoded into embedded vector representations, using a text encoder and an image encoder, respectively. In particular, we leverage a pre-trained vision and language model, called CLIP [37], for both image encoding and text encoding. Different from ResNet and GloVe/BERT for image and text encoding [19, 43, 45, 47] used by previous work in this task [19, 43, 47], CLIP can provide unified representation vectors for each modality with the same dimensionality. For instance, an image of red shoes has a similar representation vector to the text "red shoes". Given a user's natural-language feedback o_t at the t-th dialog turn, the encoded textual representation is denoted by $o'_t = Norm(Linear(CLIP^{txt}(o_t)))$. Similarly, given a liked image $a_{t,u}$ at the t-th turn, the encoded image representation is denoted by $a'_{t,u} = Norm(Linear(CLIP^{img}(a_{t,u})))$. For simplicity of notation, we use a_t and o_t directly to denote their representations (i.e. a'_t and o'_t), respectively.

To understand the user's current information needs from the recommendations and the corresponding relevance feedback at each turn, we need to generate a new composed candidate image representation instead of simply concatenating the text and image representations. We adopt a representative composition network ψ (in particular, Text Image Residual Gating (TIRG) [41]) to combine image and text representaions with a gated feature $f_{gate}(a_{t,u}, o_t)$ to establish the input image representation $a_{t,u}$ as a "reference" to the output composition representation and a residual feature $f_{res}(a_{t,u}, o_t)$ to describe the "modification" on the "reference" in the feature space [41]. The multi-modal composition feature $c_t = \psi(a_{t,u}, o_t)$ is computed by:

$$c_t = \psi(a_{t,u}, o_t) = \omega_g f_{gate}(a_{t,u}, o_t) + \omega_r f_{res}(a_{t,u}, o_t) \quad (1)$$

$$f_{gate}(a_{t,u}, o_t) = \sigma(W_{g2} * ReLU(W_{g1} * [a_{t,u}, o_t])) \odot a_{t,u} \quad (2)$$

$$f_{res}(a_{t,u}, o_t) = W_{r2} * ReLU(W_{r1} * [a_{t,u}, o_t]) \quad (3)$$

where ω_g and ω_r are learnable weights. $\sigma(\cdot)$ and $ReLU(\cdot)$ are the Sigmoid and the Rectified Linear Unit (ReLU) functions. W_{g1}, W_{g2}, W_{r1}, and W_{r2} are convolution filters. $\odot$ denotes element-wise product, and * denotes a 2d convolution with batch normalisation.

State Tracking Stage. To incorporate the users' preferences from the combined text and image representations $c_t = \psi(a_{t,u}, o_t)$, we leverage a Transformer encoder $TranEnc(\cdot)$, as in [43, 46, 47], as a state tracker to track/estimate the interaction states. In particular, the Transformer encoder allows our GOMMIR model to sequentially aggregate the recommendation and feedback information from the multi-modal composition feature c_t to attend to the entire feedback history during each interaction turn. The estimated state of the user's preferences can be obtained as follows:

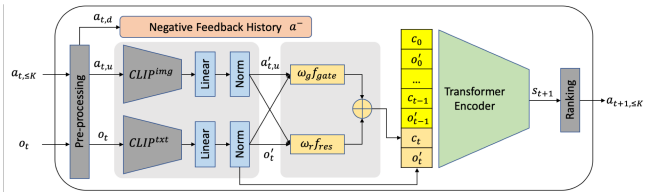


Figure 4: The proposed GOMMIR model for multi-modal interactive recommendations.

$$s_{t+1} = Linear(Tanh(Mean(TranEnc([c_{\leq t}, o_{\leq t}])))) \quad (4)$$

where $c_{\leq t} = (c_0, ..., c_t)$ and $o_{\leq t} = (o_0, ..., o_t)$ are the composition representations and critique histories, respectively.

Ranking Stage. Based on the estimated final state of the user's preferences, we adopt a greedy policy [19, 45] to recommend a candidate item list for the next action. In particular, we select the top-K closest images to the estimated state s_{t+1} under the Euclidean distance in the image feature space: $a_{t+1,\leq K} \sim KNNs(s_{t+1})$, where $KNNs(\cdot)$ 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, based on the interaction history $h_t = (o_{\leq t}, a_{\leq t,\leq K})$, a post-filter is adopted to remove any previously recommended candidate items from the ranking. Indeed, since these items have already been shown to the user, they are assumed to be non-relevant, and do not need to be re-shown again [45].

To summarise, in the GOMMIR model, we maintain the Transformer Encoder for state tracking and the $KNNs(\cdot)$ for sampling as in the state-of-the-art approaches [43, 46, 47]. Meanwhile, we leverage the CLIP-based multi-modal encoders and a composition network (i.e. TIRG [41]) to explicitly learn the multi-modal composition features at each turn and to better incorporate the users' dynamic preferences, rather than using a simple concatenation operation [19, 43, 45] (as described in Sections 1 & 2).

3.3 Learning Algorithm

We adopt a multi-task learning [27] approach for GO-POMDP to optimise the recommendation policy with a policy gradient method (e.g., REINFORCE [8]) learning loss and to explicitly learn good representations of the multi-modal composition features with a supervised learning loss. Although value-based methods (such as DQN [35]) have demonstrated many advantages in solving DRL problems, they are known to be prone to instability with value function approximations [8, 40, 48]. Alternatively, policy-based methods (such as REINFORCE) are more stable given a sufficiently small learning rate [8] compared to value-based methods (such as DQN [35]). Therefore, we rely on a policy gradient method (in particular REINFORCE) and enrich this on-policy method with goals for the MMIR task.

3.3.1 Goal-Oriented Policy Optimisation. The objective of goal-oriented policy optimisation is to reach the goal g via a goal-oriented policy π_θ ($\theta \in \mathbb{R}$ denotes policy parameters) that maximises the expectation of the cumulative return over the goal distribution:

$$\max_\theta J(\pi_\theta) = \max_\theta \mathop{\mathbb{E}}_{\tau \sim \pi_\theta}[R(\tau)] \quad (5)$$

where $R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_{t,\leq K}, g)$ is the discounted cumulative reward, and T is the maximum turn in the interaction trajectory. The expectation is taken over trajectories $\tau = ((o_0, a_{0,\leq K}), ..., (o_T, a_{T,\leq K}))$.

We define the loss for optimising the recommendation policy based on the gradient of $J(\pi_\theta)$ with REINFORCE. Specifically, the gradient of Equation (5) can be computed as follows:

$$\nabla_\theta J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} [\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_{t,\leq K}|s_t) R(\tau)] \tag{6}$$

We define $\log \pi_\theta(a_{t,\leq K}|s_t)$ as a softmax cross-entropy objective to identify the positive sample amongst a set of negative samples:

$$\log \pi_\theta(a_{t,\leq K}|s_t) = \log(\frac{e^{\kappa(s_t,g)}}{e^{\kappa(s_t,g)} + \sum_{j=1}^{J} e^{\kappa(s_t,a_j^-)}}) \tag{7}$$

where $\kappa(\cdot)$ is a similarity kernel that can be the dot product or the negative l_2 distance in our experiments. g is a target image representation, and a_j^- ($j \in [1, J]$) are negative sample representations. The negative samples are usually randomly sampled images from the candidate pool in the previous research [19, 43, 45]. To leverage the benefits from the non-verbal relevance feedback, as hard negative samples, we iteratively consider randomly sampled images from the previously disliked recommendations $(a_{0,d}, ..., a_{t-1,d})$ and the disliked items in the following turn $a_{t,d}$, i.e. $a_{d,j}^-(j \in [1, J])$. Therefore, we optimise the policy after we collect the users' relevance feedback o_t and $a_{t,d}$.

We define the goal-oriented reward $r(s_t, a_{t,\leq K}, g)$ as the sum of the similarities between all the top-K candidates and the goal:

$$r(s_t, a_{t,\leq K}, g) = \sum_{i=1}^{K} \kappa(a_{t,i}, g) = \kappa(a_{t,u}, g) + \sum_{d=1}^{K-1} \kappa(a_{t,d}, g) \tag{8}$$

Here, we expect our GOMMIR model to learn from rewards $r_{t,u} = \kappa(a_{t,u}, g)$ on the *critiqued/liked* items, as well as from the extra rewards $r_{t,d} = \sum_{d=1}^{K-1} \kappa(a_{t,d}, g)$ on the *disliked* items. Both the hard negative sampling and the extra visual rewards $r_{t,d}$ on the disliked items provide further information relating to the target item, thereby enhancing the goal-oriented optimisation objective to effectively optimise the recommendation agent.

3.3.2 Composition Representation Learning. To learn the multi-modal composition representation explicitly, we leverage a triplet loss objective for composition representation learning along with the policy optimisation process. Given a multi-modal composition feature $c_t = \psi_\phi(a_{t,u}, o_t)$, a target item (i.e. the goal) g and a negative sample a^-, the composition loss $L(\psi_\phi)$ can be defined as follows:

$$\max_\phi L(\psi_\phi) = \sum_{t=0}^{T} \max_\phi (0, l_2(c_t, g) - l_2(c_t, a^-) + \epsilon_1) \tag{9}$$

where $\phi \in \mathbb{R}$ denotes the parameters of the composition network ψ. $l_2(\cdot)$ denotes the l_2 distance. The negative sample a^- is sampled from $(a_1^-, ..., a_J^-)$ as in Equation (7). ϵ_1 is a constant for the margin to keep negative samples far apart.

Therefore, we jointly train our model with both the goal-oriented policy optimisation objective $J(\pi_\theta)$ and the composition representation learning objective $L(\psi_\phi)$ to mitigate the so-called coupling issue (as described in Sections 1 & 2), as follows:

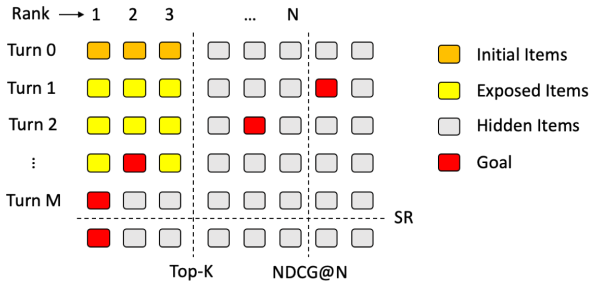


Figure 5: An example of a top-K (e.g., $K = 3$) recommendation in the goal-oriented MMIR scenario.

$$\max \mathcal{L}_{GOMMIR} = \max_\theta J(\pi_\theta) + \max_\phi L(\psi_\phi) \tag{10}$$

3.3.3 Pre-training. To improve the sample efficiency with the policy gradient method, we initialise the GOMMIR model with a supervised pre-training process instead of using a random initialisation. We leverage a triplet loss supervised objective $L(\pi_\theta)$ to pre-train the recommendation policy π_θ, similar to [19]:

$$\max_\theta L(\pi_\theta) = \sum_{t=0}^{T} \max_\theta (0, l_2(s_t, g) - l_2(s_t, a^-) + \epsilon_2) \tag{11}$$

where a^- is a randomly sampled image, and ϵ_2 is a constant for the margin. To learn the composition representation explicitly, we also jointly pre-train the GOMMIR model with both triplet loss objectives (i.e. π_θ and $L(\psi_\phi)$) as follows:

$$\max \mathcal{L}_{Pre-train} = \max_\theta L(\pi_\theta) + \max_\phi L(\psi_\phi) \tag{12}$$

Based on the pre-trained model obtained with $\mathcal{L}_{Pre-train}$, the joint loss objective $\mathcal{L}_{GOMMIR}$ can further improve the composition representations with $L(\psi_\phi)$, as well as maximise the expected future rewards with $J(\pi_\theta)$, thereby addressing the coupling issue.

4 EXPERIMENTAL SETUP

In this section, we evaluate the effectiveness of our proposed GOMMIR model in comparison to the existing approaches from the literature. Figure 5 shows an example of a top-K (e.g., $K = 3$) recommendation in the MMIR scenario. A user browses the exposed items (i.e. the top-K recommendations) and gives likes/dislikes and natural-language critiques on the recommendations at each turn. The figure illustrates how a user can find the desired item (i.e. the goal) through multi-turn interactions. Following the methodology in [45, 46, 57], we measure the effectiveness of the interactive recommendation models at interaction turn M. Meanwhile, the user may examine more items in the ranking list at each turn, down to rank N ($N > K$). In particular, we address three research questions:

- RQ1: Does our proposed GOMMIR model with joint policy and composition representation learning for GO-POMDP outperform the existing state-of-the-art baseline models in the multi-modal interactive recommendation task?
- RQ2: How do the components designed for composition representation learning and goal-oriented policy optimisation in the GOMMIR model affect the performance?

Table 1: Datasets' statistics.

	Shoes		Dresses		Shirts		Tops & Tees	
	Train	Test	Train	Test	Train	Test	Train	Test
Triplets	10,751	-	11,970	4,034	11,976	4,076	12,054	3,924
Images	10,000	4,658	7,182	2,454	8,555	2,966	8,387	2,808

• RQ3: What are the impacts of the introduced hyper-parameters on the performance, such as the reward discount factor γ and the number of recommended items K?

4.1 Datasets & Setup

Our proposed approaches are evaluated on four well-known fashion datasets, namely the *Shoes* [5, 19] and *Fashion IQ Dresses, Shirts, Tops & Tees* [43] datasets, to verify the generalisation of the recommendation performance of our proposed GOMMIR model, following [45]. The statistics of the four datasets are summarised in Table 1. All datasets provide triples (i.e. $\langle a_{target}, a_{candidate}, o_{caption} \rangle$) for training/testing the user simulators (discussed further in Section 4.2). In particular, $o_{caption}$ denotes a relative caption that encapsulates the differences between the target (a_{target}) and candidate ($a_{candidate}$) images. The relative captions of the image pairs have been collected from real users via crowd-sourcing. In addition, all datasets also provide images of the fashion products that can be used for training/testing the recommendation models.

We pre-train our GOMMIR model with a multi-task supervised learning setting (as per Equation (12)) for initialisation, and then further optimise GOMMIR with a joint supervised and reinforcement learning setting with Equation (10)[1]. Following [19], we use Adam [25] with learning rates $\eta_1 = 10^{-3}$ and $\eta_2 = 10^{-5}$ with Equation (12) and Equation (10), respectively, for optimising the GOMMIR model's parameters. The similarity kernel $\kappa(\cdot)$ in Equation (7) is set to be the dot product by default. Unless mentioned otherwise, the discount factor γ is set to 0.2 due to the generally good performance. The embedding dimensionality of the feature space is set to 512 with the pre-trained CLIP model using the "RN101" checkpoint[2]. The batch size is set to 128 and the number of negative samples (i.e. J) is set to 5, following [45]. The maximum number of epochs for training is 20, as in [46]. We consider the top-K (i.e. $K = 3$) items as a recommendation at each interaction turn for both training and testing. Due to the lack of the users' profiles in the datasets, the recommendation models make an initial random recommendation for each user with a fixed random seed (i.e. 42). We expect the recommendations to become more similar to the target item with more interactions. The maximum number of interaction turns is set to 10 as in [45, 46].

4.2 Online Evaluation

An interactive recommender system is a type of closed-loop system [38] in which the inputs (i.e. the users' relevance feedback) of the recommender system are fully or partially determined by the outputs (i.e. the recommendations). When we evaluate the interactive recommendation models, it is challenging to know the users' real-time feedback on the recommendations at each interaction turn. To alleviate this issue, we adopt relative captioning models[3] (i.e. the Show, Attend, & Tell [49] model on *Shoes* and the VL-Transformer [43, 46] model on *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees*) as a surrogate for real human users (a.k.a. user simulators), as in [19, 45, 57]. Indeed, there is a growing interest in user simulation for optimisation and evaluation purposes, such as in conversational recommendation [17, 23, 58] and news recommendation [34]. We assume the user desires a visual item and gives verbal and non-verbal relevance feedback on the recommendations. To properly simulate the user's behaviour, we assume that the user simulator can observe a ranked list of visual recommendations at each interaction turn. Then, the user simulator gives a "like" on the item that is the most similar to the target image, while it gives "dislikes" on other items, and provides a natural-language critique (i.e. a relative caption) to describe the attributes missing from the liked item. The non-verbal relevance feedback (i.e. "likes" and "dislikes") reflects the users' relative preferences among the recommendations at each turn, while the verbal relevance feedback (i.e. natural-language critiques) illustrates the users' evolving dynamic preferences initiated by themselves. Note that we directly use the user simulator checkpoint[4] [5, 19] for *Shoes* provided by Guo et al. [19], following the setting in [19, 45], while we use the user simulator checkpoints for *Fahion IQ Dresses, Shirts*, and *Tops & Tees* provided by Wu et al. [46] following the setting in [43, 46]. It is worth noting that in the real world, the situation of interactive recommendation can be much more complicated in terms of both verbal and non-verbal relevance feedback. For instance, the user may give "likes" on more than one item in the recommendation list and may also give free-form natural-language feedback even on "disliked" items. We leave the handling of such more complex situations in the interactive recommendation task as interesting future work. Note also that our simplification is necessitated by the existing datasets and the availability of accurate user simulators.

4.3 Evaluation Metrics

We measure the effectiveness of different interactive recommendation models under the two evaluation metrics:

• **Normalised Discounted Cumulative Gain (NDCG)**. NDCG measures the quality of the ranking lists by emphasising the importance of higher ranks in relation to the lower ones. In our experiments, we consider NDCG@N, which is truncated at rank $N = 3$ and $N = 10$ as in [45, 46].

• **Success Rate (SR)**. SR considers the percentage of users for which the target image was successfully retrieved with top-K recommendations within M interactions. We report the interaction turn $M \in [1, 10]$ as in [54, 57].

If a user obtains the target item in less than 10 interaction turns, we consider the ranking metrics (i.e. NDCG@3 and NDCG@10) for that user to be equal to one for all turns thereafter. We conduct significance testing in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction for all evaluation metrics (i.e. NDCG@3, NDCG@10 and SR) at the 10th interaction turns.

[1] The code and datasets for this paper are publicly available in https://github.com/yashonwu/gommir [2] https://github.com/openai/CLIP

[3] These user simulators were used by the original authors - we replicate their user simulator setups. [4] https://github.com/XiaoxiaoGuo/fashion-retrieval

4.4 Baselines

We compare our GOMMIR model with three groups of representative baseline models for the MMIR task.

Interactive Recommendation Models with a Single Modality. We first consider two representative interactive recommendation (IR) models, each with a single modality, using a Transformer-based state tracker for sequential modelling as in Section 3.2.

• **IR**$_{img}$: IR$_{img}$ estimates the users' preferences through the sequences of their liked images only.

• **IR**$_{txt}$: IR$_{txt}$ estimates the users' preferences through the sequences of their natural-language critiques only.

Text-Image Retrieval Models. We next consider two representative text-image retrieval models that explicitly learn the composition representations from both the text and image modalities. These models are extended to the MMIR task by incorporating the current recommendations and the corresponding natural-language feedback at each turn. However, due to their lack of a state tracker, they ignore the users' interaction histories.

• **TIRG**[5] [41]: TIRG was the first model proposed for the composition of text and image features in the context of text-image retrieval through a gating and a residual connection. We also use TIRG as a composition network in our GOMMIR model in Section 3.2.

• **CLIP4Cir**[6] [2, 3]: CLIP4Cir adopts a Combiner network [3] with the CLIP image and text encoders to understand the images content, integrate the textual descriptions and provide a combined feature for text-image retrieval. CLIP4Cir obtains a state-of-the-art performance in the context of text-image retrieval on *Fashion IQ*.

Multi-Modal Interactive Recommendation Models. We now consider multi-modal interactive recommendation baseline models with both image and text modalities. These baseline models learn the multi-modal composition representations implicitly. In particular, both EGE [45] and DEERS [61] are the two baseline models that use DRL algorithms.

• **DM**[7] [19]: In the Dialog Manager (DM) model, the image and text representations are concatenated and embedded through a linear transformation layer to obtain a composed feature. The state tracker is based on a GRU for tracking and estimating the users' preferences with the composed representation and the history representation of previous interaction turns.

• **MMT**[8] [43]: The Multi-Modal Transformer (MMT) model directly attends to the entire interaction history of both the users' previous textual feedback and the system's visual recommendations.

• **MMRAN** [46]: The Multi-Modal Recurrent Attention Network (MMRAN) model leverages a gated recurrent network (GRN) with a feedback gate for combining the image and text representations and further uses a multi-head attention network (MAN) for tracking the users' dynamic preferences over time.

• **EGE** [45]: The Estimator-Generator-Evaluator (EGE) model is another GRU-based model, which uses a multi-task learning approach for POMDP to optimise the model, combining a supervised learning classification loss and a Q-learning prediction loss.

• **DEERS** [61]: The DEERS model leverages a Deep Q-Network (DQN) to automatically learn the optimal recommendation strategies by incorporating positive and negative feedback. It adopts two GRU-based state trackers to track the users' positive and negative states, respectively. We extend this model for the multi-modal interactive recommendation task by incorporating both images and natural-language feedback as inputs.

In addition to the above baseline models for the MMIR task, the GOMMIR variants used for the ablation studies (in Section 5.2) can also act as strong baselines. For fair comparisons, all of the tested baseline models use CLIP (using the "RN101" checkpoint) for providing the texts and image representations (as described in Section 3.2). Although there are a few more other models with different formulations for the interactive recommendation task, these models are not comparable with our scenario due to them being unable to incorporate both the textual and visual modalities during the recommendation process [29, 39], requiring additional attributes of items for learning [54, 55, 57] or requiring multi-modal knowledge graph for reasoning [44].

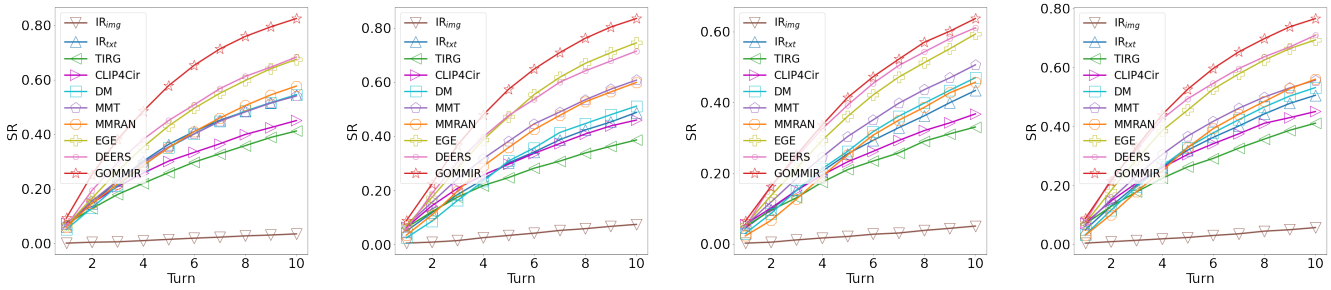
5 EXPERIMENTAL RESULTS

In this section, we analyse the experimental results with respect to the three research questions stated in Section 4 to gauge the effectiveness of our proposed GOMMIR model. Specifically, we address the overall effectiveness of our proposed GOMMIR model for the MMIR task (RQ1, Section 5.1), the impact of the goal-oriented policy optimisation and composition representation learning (RQ2, Section 5.2), and the effects of the hyper-parameters (RQ3, Section 5.3). To consolidate our findings, we provide a use case from the logged experimental results in Section 5.4.

5.1 Performance Comparison (RQ1)

Figure 6 shows the effectiveness of our proposed GOMMIR model in comparison to the baseline models for top-3 recommendation in terms of SR while varying the number of interaction turns on the *Shoes*, *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets. Comparing the results in Figure 6, we observe that our proposed GOMMIR model generally achieves a better overall performance in terms of SR at various interaction turns. As the number of interaction turns increases, the magnitude of the differences between the effectiveness of GOMMIR with the baseline models on SR also increases. Similar trends are also observed with other metrics (i.e. NDCG@3 and NDCG@10) – we omit their reporting due to space constraints. The better overall performance of our proposed GOMMIR model indicates that learning the composition representations explicitly with goal-oriented policy optimisation can better incorporate the users' preferences from the recommended visual items and the corresponding verbal and non-verbal relevance feedback. To quantify the improvements of our proposed GOMMIR model compared to the other nine baseline models, Table 2 reports their performances at the 10th interaction turn. The best results of the baseline models and the best overall results are underlined and highlighted in bold, respectively. Analysing the results in the table, we observe that our proposed GOMMIR model achieves better performances at the 10th turn than the best baseline model on all metrics on *Shoes*, *Dresses*, *Shirts*, and *Tops & Tees* by a margin of 19-21%, 10-12%, 3-4%, and

[5] https://github.com/google/tirg
[6] https://github.com/ABaldrati/CLIP4Cir
[7] https://github.com/XiaoxiaoGuo/fashion-retrieval
[8] https://github.com/XiaoxiaoGuo/fashion-iq



(a) *Shoes* (b) *Dresses* (c) *Shirts* (d) *Tops & Tees*

Figure 6: Comparison of the recommendation effectiveness at various interaction turns with top-3 recommendation.

Table 2: The effectiveness of the tested models at the 10th turn. The best results of baseline models and the best overall results are underlined and highlighted in bold, respectively. % Improv. indicates the improvements by our GOMMIR model over the best baseline model. * denotes a significant difference in terms of paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to GOMMIR.

Models	Shoes			Dresses			Shirts			Tops & Tees		
	NDCG@3	NDCG@10	SR	NDCG@3	NDCG@10	SR	NDCG@3	NDCG@10	SR	NDCG@3	NDCG@10	SR
IR_{img}	0.0339*	0.0366*	0.0350*	0.07272*	0.0780*	0.0746*	0.0490*	0.0526*	0.0506*	0.0549*	0.0590*	0.0566*
IR_{txt}	0.5365*	0.5556*	0.5451*	0.4784*	0.4984*	0.4878*	0.4240*	0.4448*	0.4336*	0.4973*	0.5189*	0.5053*
TIRG	0.4067*	0.4226*	0.4124*	0.3803*	0.3934*	0.3863*	0.3248*	0.3400*	0.3304*	0.4049*	0.4237*	0.4106*
CLIP4Cir	0.4438*	0.4566*	0.4506*	0.4527*	0.4735*	0.4597*	0.3608*	0.3754*	0.3675*	0.4437*	0.4610*	0.4501*
DM	0.5374*	0.5571*	0.5453*	0.5022*	0.5225*	0.5110*	0.4598*	0.4811*	0.4697*	0.5226*	0.5419*	0.5313*
MMT	0.5336*	0.5521*	0.5406*	0.5981*	0.6194*	0.6072*	0.4945*	0.5124*	0.5061*	0.5501*	0.5697*	0.5563*
MMRAN	0.5680*	0.5879*	0.5771*	0.5887*	0.6099*	0.5986*	0.4484*	0.4692*	0.4568*	0.5508*	0.5710*	0.5598*
EGE	0.6657*	0.6880*	0.6750*	0.7353*	0.7559*	0.7449*	0.5826*	0.6044*	0.5931*	0.6868*	0.7059*	0.6930*
DEERS	0.6749*	0.6940*	0.6831*	0.7083*	0.7250*	0.7143*	0.6027	0.6215	0.6106	0.6989*	0.7144*	0.7090*
GOMMIR	**0.8173**	**0.8297**	**0.8248**	**0.8255**	**0.8385**	**0.8346**	**0.6275**	**0.6440**	**0.6369**	**0.7582**	**0.7706**	**0.7653**
% Improv.	21.10	19.55	20.74	12.27	10.93	12.04	4.11	3.62	4.31	8.48	7.87	7.94

7-8%, respectively. Indeed, our proposed GOMMIR model is significantly better than the baseline models (except for DEERS on *Shirts*) for each metric at the 10th turn in top-3 recommendation.

Therefore, in answer to RQ1, the results show that the GOMMIR model can outperform the existing state-of-the-art baseline models. In particular, it is significantly more effective than the state-of-the-art baseline models at the 10th turn. Therefore, we conclude that our proposed GOMMIR model, which addresses the coupling issue, can better incorporate the users' preferences for an improved top-3 recommendation. In the next section, we analyse the impact of the coupling issue and demonstrate how they are addressed with our proposed GOMMIR model.

5.2 Impact of Components (RQ2)

To address RQ2, we investigate the impact of the components designed for both composition representation learning and goal-oriented policy optimisation to tackle the coupling issue. Table 3 reports the performances of our GOMMIR model with different ablations in terms of SR considering the original setting in the top part of the table, the composition representation learning in the second part of the table, and the goal-oriented optimisation in the last part of the table. The same trends can be also observed on NDCG@3 and NDCG@10 – we omit their reporting due to space constraints.

Composition Representation Learning. We investigate the impact of the explicit composition learning on the performance of our proposed GOMMIR model in terms of four aspects: the whole composition network ψ, the gated feature f_{gate}, the residual feature f_{res}, and the triplet loss for the composition representation learning $L(\psi_\phi)$. Table 3 (second part of the table) reports the performances of our GOMMIR model with different ablations considering the aforementioned four aspects at the 10th interaction turn. The reported results in Table 3 show that the full GOMMIR model (i.e. considering the above four aspects in the second part of Table 3) can outperform "GOMMIR w/o ψ", "GOMMIR w/o f_{gate}", "GOMMIR w/o f_{res}", and "GOMMIR w/o $L(\psi_\phi)$". These results suggest that our proposed GOMMIR model can benefit from both the composition network (i.e. TIRG) with both gated and residual features and the composition learning loss $L(\psi_\phi)$. In particular, the composition learning loss $L(\psi_\phi)$ contributes the most to the GOMMIR model's performance on all four datasets, while the gated feature f_{gate} contributes the least on *Dresses* and *Tops & Tees*, and the residual feature f_{res} contributes the least on *Shoes* and *Shirts*. Therefore, it is necessary to explicitly learn the multi-modal composition representations with an advanced composition network (such as TIRG).

Goal-Oriented Policy Optimisation. We now investigate the impact of goal-oriented policy optimisation on the performance of our proposed GOMMIR model in terms of four aspects: the hard

Table 3: Ablation study at turn 10 in terms of SR. w/o denotes that component is removed from GOMMIR. * denotes a significant difference in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to GOMMIR.

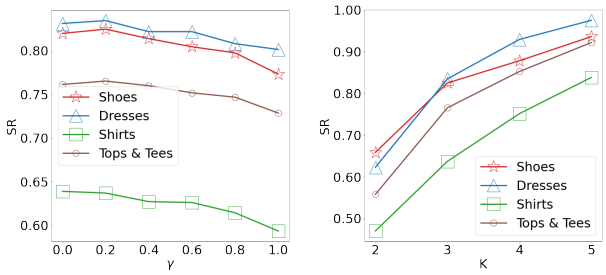
Models	Shoes	Dresses	Shirts	Tops & Tees
GOMMIR	**0.8248**	**0.8346**	**0.6369**	**0.7653**
Composition Representation Learning				
1. w/o ψ	0.7428*	0.7384*	0.5850*	0.7001*
2. w/o f_{gate}	0.7168*	0.7816*	0.5792*	0.6948*
3. w/o f_{res}	0.7863*	0.7384*	0.5890*	0.6595*
4. w/o $L(\psi_\phi)$	0.6932*	0.7115*	0.4589*	0.6528*
Goal-Oriented Policy Optimisation				
5. w/o $a^-_{d,j}$ in Eq. (7)	0.7231*	0.7649*	0.6177	0.7279*
6. w/o $a_{t,d}$ in $a^-_{d,j}$	0.8010*	0.8329	0.6274	0.7546
7. w/o $r(s_t, a_{t,\leq K}, g)$	0.7799*	0.7991*	0.6001*	0.7350*
8. w/o $r_{t,d}$ in Eq. (8)	0.8128*	0.8305	0.6369	0.7614

negative sampling $a^-_{d,j}$ in Equation (7), the following relevance feedback $a_{t,d}$ in hard negative sampling $a^-_{d,j}$, the goal-oriented rewards $r(s_t, a_{t,\leq K}, g)$ in Equation (6), and the extra rewards of the disliked items $r_{t,d}$ in Equation (8). Table 3 (last part) reports the performances of the GOMMIR variants considering the aforementioned four aspects. In particular, within the table, "GOMMIR w/o $a^-_{d,j}$ in Equation (7)" selects negative samples randomly from the candidate pool rather than sampling from the negative feedback history (i.e. the disliked items $(a_{0,d}, ..., a_{t,d})$). "GOMMIR w/o $r_{t,d}$ in $a^-_{d,j}$" samples hard negatives from the previously disliked recommendations $(a_{0,d}, ..., a_{t-1,d})$. "GOMMIR w/o $r(s_t, a_{t,\leq K}, g)$ in Equation (6)" optimises the recommendation policy using supervised learning without the goal-oriented rewards. "GOMMIR w/o $r_{t,d}$ in Equation (8)" only considers the visual reward for the critiqued/liked item rather than all the rewards for both the liked and disliked recommendation items. The results reported in Table 3 show that the full GOMMIR model (i.e. considering the above four aspects) can outperform the above four variants on all four datasets, except for "GOMMIR w/o $r_{t,d}$ in Equation (8)" on *Shirts*. These results suggest that it is necessary to consider non-verbal relevance feedback in the hard negative sampling and the reward function during the goal-oriented policy optimisation process. In addition, we can also observe that GOMMIR can gain more improvements with the explicit composition loss $L(\psi_\phi)$ compared to using the goal-oriented rewards $r(s_t, a_{t,\leq K}, g)$.

In response to RQ2, we find that our proposed GOMMIR model can benefit from explicitly learning the composition representation with an advanced composition network (i.e. TIRG) and optimising the recommendation policy with hard negative sampling and rewards based on the non-verbal relevance feedback.

5.3 Impact of Hyper-Parameters (RQ3)

To address RQ3, Figure 7 depicts the impact in terms of SR of the reward discount factor γ and the number of recommended items K when training the GOMMIR model on all four datasets, respectively.



(a) γ for *SR* (b) K for *SR*

Figure 7: Comparison of the recommendation effectiveness at 10th turn with different γ and K values.

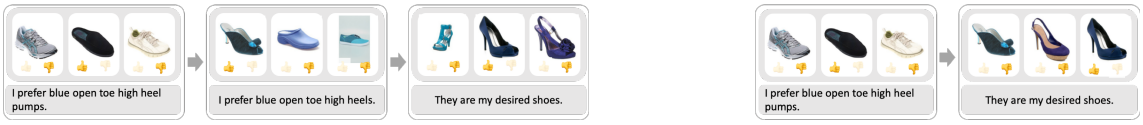
The same results/trends can be also observed for NDCG@3 and NDCG@10, we omit their reporting due to space constraints.

Effect of the reward discount factor (γ). Figure 7 (a) shows SR at the 10th turn in top-3 recommendation with various reward discount factors γ on the four datasets. In particular, the model can only consider the immediate goal-oriented reward with $\gamma = 0$ or weight all future rewards equally with $\gamma = 1$. We can observe that the performance of GOMMIR decreases when the reward discount factor γ is larger than 0.2. The better performance with a lower reward discount factor shows that the immediate reward is much more important compared to the future rewards.

Effect of the number of recommended items (K). Figure 7 (b) shows SR with different numbers of top-K recommendations at each turn (i.e. $K = 2, 3, 4, 5$). The K values indicate how deep the users can explore among a ranking list of all items at each interaction turn. Note that larger metrics indicate a better performance across top-K recommendations even though the number of exposed items at each turn is different. We observe that the performance of GOMMIR increases when the number of recommended items K increases from 2 to 5, as more items are exposed to the users and users provide more feedback. Overall, in response to RQ3, we find that a lower reward discount factor γ and more exposed top-K items can improve the effectiveness of our GOMMIR model.

5.4 Use Case

In this section, we present a use case of the multi-modal interactive recommendation on the *Shoes* dataset in Figure 8. In particular, the figures show the interaction process for the top-3 recommendations between the simulated users for the DEERS (i.e. the strongest baseline model) and GOMMIR models. For a fair comparison, the initial images are the same across the tested models given the target image from the testing set. When the target item is listed in the recommendation list, the user simulator will give a comment to end the interaction, such as "They are my desired shoes" in Figure 8 (b). Comparing the recommendations made by DEERS and GOMMIR on the *Shoes* dataset, we can observe that our proposed GOMMIR model can find the target items with fewer interaction turns compared to DEERS – this is expected, due to the increased effectiveness of GOMMIR shown in Section 5.1. In addition, our GOMMIR model is more effective at incorporating more relevant features of the critique in the following interaction turn. For instance, at the initial interaction turn in Figures 8 (a) and (b), the user



(a) DEERS (b) GOMMIR

Figure 8: Example use cases for the interactive recommendation with DEERS and GOMMIR on *Shoes*.

claimed that "I prefer blue open toe high heel pumps" in comparison to the 2nd image (i.e. black clogs). Our GOMMIR model suggests open-toe recommendations, while DEERS ignores the "open-toe" feature from the critique and instead recommends closed-toe blue clogs in the second place and closed-toe blue sneakers in the third place. We observed similar trends and results in use cases with the other baseline models on the *Shoes*, *Dresses*, *Shirts*, and *Tops & Tees* datasets. We omit their reporting in this paper because of space constraints.

6 CONCLUSIONS

In this paper, we proposed a novel goal-oriented multi-modal interactive recommendation (GOMMIR) model to effectively incorporate the users' preferences from both verbal and non-verbal relevance feedback over time, by addressing the coupling issue of policy optimisation and multi-modal composition representation learning. Specifically, we jointly leveraged both goal-oriented deep reinforcement learning and supervised learning objectives to explicitly learn the multi-modal representations with a multi-modal composition network (i.e. TIRG) during the recommendation policy optimisation process. We adopted a pre-trained CLIP model for image and text encoding, and a Transformer-based *state tracker* for estimating the users' preferences from the users' natural-language critiques and the previously combined representations from the composition network. Following previous work [19, 43, 45], we trained and evaluated our GOMMIR model by using a user simulator as a surrogate for real human users. Our experiments on the *Shoes*, *Dresses*, *Shirts* and *Tops & Tees* datasets demonstrated that our proposed GOMMIR model achieves better performances of 19-21%, 10-12%, 3-4%, and 7-8% compared to the best baseline models, respectively. Moreover, our reported results showed that our proposed GOMMIR model can benefit from explicit composition representation learning and goal-oriented policy optimisation with both verbal and non-verbal relevance feedback.

ACKNOWLEDGMENTS

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

REFERENCES

[1] M Mehdi Afsar, Trafford Crump, and Behrouz Far. 2022. Reinforcement learning based recommender systems: A survey. *Comput. Surveys* 55, 7 (2022), 1–38.
[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features. In *Proc. CVPR*. 4959–4968.
[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features. In *Proc. CVPR*. 21466–21474.
[4] 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.
[5] Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proc. ECCV*. 663–676.
[6] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. 2019. Conversational product search based on negative feedback. In *Proc. CIKM*. 359–368.
[7] 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.
[8] 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.
[9] Xiaocong Chen, Lina Yao, Julian McAuley, Guanglin Zhou, and Xianzhi Wang. 2021. A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. *arXiv preprint arXiv:2109.03540* (2021).
[10] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proc. CVPR*. 3001–3011.
[11] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. 2022. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research* 74 (2022), 1159–1199.
[12] 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).
[13] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. 2021. Towards multi-modal conversational information seeking. In *Proc. SIGIR*. 1577–1587.
[14] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proc. SIGIR*. 1431–1441.
[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*. 4171–4186.
[16] Benjamin Eysenbach, Tianjun Zhang, Ruslan Salakhutdinov, and Sergey Levine. 2022. Contrastive learning as goal-conditioned reinforcement learning. In *Proc. NeurIPS*.
[17] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *AI Open* 2 (2021), 100–126.
[18] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. 2021. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proc. MM*. 5185–5193.
[19] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Proc. NeurIPS*. 678–688.
[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. 770–778.
[21] Chenhao Hu, Shuhua Huang, Yansen Zhang, and Yubao Liu. 2022. Learning to Infer User Implicit Preference in Conversational Recommendation. In *Proc. SIGIR*. 256–266.
[22] Jin Huang, Harrie Oosterhuis, Bunyamin Cetinkaya, Thijs Rood, and Maarten de Rijke. 2022. State encoders in reinforcement learning for recommendation: A reproducibility study. In *Proc. SIGIR*. 2738–2748.
[23] Dietmar Jannach. 2022. Evaluating conversational recommender systems. *arXiv preprint arXiv:2208.12061* (2022).
[24] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.
[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. ICLR*.
[26] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. In *Proc. NeurIPS*. 19884–19895.
[27] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *Proc. ICML*. 5639–5650.
[28] Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *Proc. SIGIR*.

2425–2428.
[29] 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.
[30] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proc. KDD*. 2073–2083.
[31] 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.
[32] Yuanguo Lin, Yong Liu, Fan Lin, Pengcheng Wu, Wenhua Zeng, and Chunyan Miao. 2021. A survey on reinforcement learning for recommender systems. *arXiv preprint arXiv:2109.10665* (2021).
[33] Minghuan Liu, Menghui Zhu, and Weinan Zhang. 2022. Goal-Conditioned Reinforcement Learning: Problems and Solutions. In *Proc. IJCAI*. 5502–5511.
[34] Xufang Luo, Zheng Liu, Shitao Xiao, Xing Xie, and Dongsheng Li. 2022. MINDSim: User Simulator for News Recommenders. In *Proc. WWW*. 2067–2077.
[35] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP*. 1532–1543.
[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*. 8748–8763.
[38] Stefan Simrock. 2011. Tutorial on Control Theory. In *Proc. ICAELEPCS*. 10–14.
[39] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proc. SIGIR*. 235–244.
[40] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proc. NeurIPS*.
[41] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proc. CVPR*. 6439–6448.
[42] Kai Wang, Zhene Zou, Qilin Deng, Jianrong Tao, Runze Wu, Changjie Fan, Liang Chen, and Peng Cui. 2021. Reinforcement learning with a disentangled universal value function for item recommendation. In *Proc. AAAI*. 4427–4435.
[43] 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.
[44] Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022. State Graph Reasoning for Multimodal Conversational Recommendation. *IEEE Transactions on Multimedia* (2022).
[45] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2021. Partially Observable Reinforcement Learning for Dialog-Based Interactive Recommendation. In *Proc. RecSys*. 241–251.
[46] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation. In *Proc. RecSys*. 124–133.
[47] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback. In *Proc. CUI*. 1–10.
[48] Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. 2020. Self-Supervised Reinforcement Learning for Recommender Systems. In *Proc. SIGIR*. 931–940.
[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*. 2048–2057.
[50] 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.
[51] Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *Proc. ICLR*.
[52] Tong Yu, Yilin Shen, and Hongxia Jin. 2019. A visual dialog augmented interactive recommender system. In *Proc. KDD*. 157–165.
[53] Tong Yu, Yilin Shen, and Hongxia Jin. 2020. Towards Hands-Free Visual Dialog Interactive Recommendation. In *Proc. AAAI*, Vol. 34. 1137–1144.
[54] Tong Yu, Yilin Shen, Ruiyi Zhang, Xiangyu Zeng, and Hongxia Jin. 2019. Vision-language recommendation via attribute augmented multimodal reinforcement learning. In *Proc. MM*. 39–47.
[55] Yifei Yuan and Wai Lam. 2021. Conversational Fashion Image Retrieval via Multiturn Natural Language Feedback. In *Proc. SIGIR*. 839–848.
[56] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *arXiv preprint arXiv:2201.08808* (2022).
[57] 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.
[58] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proc. KDD*. 1512–1520.
[59] 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.
[60] Dongyang Zhao, Liang Zhang, Bo Zhang, Lizhou Zheng, Yongjun Bao, and Weipeng Yan. 2020. Mahrl: Multi-goals abstraction based deep hierarchical reinforcement learning for recommendations. In *Proc. SIGIR*. 871–880.
[61] 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.