Personalised Multi-Modal Interactive Recommendation with Hierarchical State Representations

YAXIONG WU, University of Glasgow, UK
CRAIG MACDONALD, University of Glasgow, UK
IADH OUNIS, University of Glasgow, UK

Multi-modal interactive recommender systems (MMIRS) can effectively guide users towards their desired items through multi-turn interactions by leveraging the users' real-time feedback (in the form of natural-language critiques) on previously recommended items (such as images of fashion products). In this scenario, the users' preferences can be expressed by both the users' past interests from their historical interactions and their current needs from the real-time interactions. However, it is typically challenging to make satisfactory personalised recommendations across multi-turn interactions due to the difficulty in balancing the users' past interests and the current needs for generating the users' state (i.e. current preferences) representations over time. On the other hand, hierarchical reinforcement learning has been successfully applied in various fields by decomposing a complex task into a hierarchy of more easily addressed subtasks. In this journal article, we propose a novel personalised multi-modal interactive recommendation model (PMMIR) using hierarchical reinforcement learning to more effectively incorporate the users' preferences from both their past and real-time interactions. In particular, PMMIR decomposes the personalised interactive recommendation process into a sequence of two subtasks with hierarchical state representations: a first subtask where a history encoder learns the users' past interests with the hidden states of history for providing personalised initial recommendations, and a second subtask where a state tracker estimates the current needs with the real-time estimated states for updating the subsequent recommendations. The history encoder and the state tracker are jointly optimised with a single objective by maximising the users' future satisfaction with the recommendations. Following previous work, we train and evaluate our PMMIR model using a user simulator that can generate natural-language critiques about the recommendations as a surrogate for real human users. Experiments conducted on two derived fashion datasets from two well-known public datasets demonstrate that our proposed PMMIR model yields significant improvements in comparison to the existing state-of-the-art baseline models. The datasets and code are publicly available at: https://github.com/yashonwu/pmmir.

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

Additional Key Words and Phrases: interactive recommendation, multi-modal, personalisation, reinforcement learning

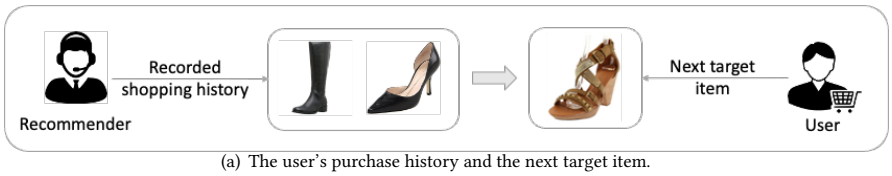
ACM Reference Format:
Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2024. Personalised Multi-Modal Interactive Recommendation with Hierarchical State Representations. 1, 1 (March 2024), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

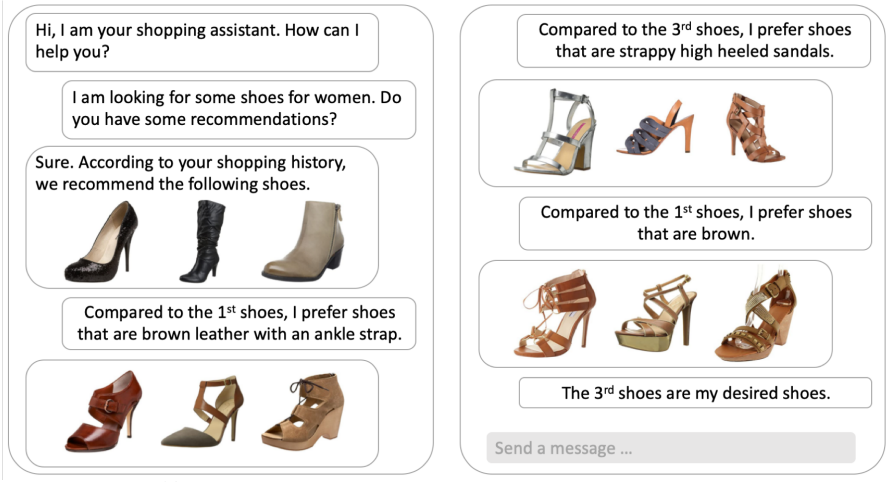
Recent advances in multi-modal interactive recommender systems (MMIRSs) enable the users to explore their desired items (such as images of fashion products) through multi-turn interactions by expressing their current needs with

Authors' addresses: Yaxiong Wu, University of Glasgow, Glasgow, UK; Craig Macdonald, University of Glasgow, Glasgow, UK; Iadh Ounis, University of Glasgow, Glasgow, UK.

Manuscript submitted to ACM



(a) The user's purchase history and the next target item.



(b) The real-time interactions between a recommender system and a user.

Fig. 1. An example of the personalised multi-modal interactive recommendation.

real-time feedback (often natural-language critiques) according to the quality of the recommendations [16, 28, 49, 51–53, 58, 59, 61, 62]. In this multi-modal interactive recommendation (MMIR) scenario, the users' preferences can be represented by both the users' past interests from their historical interactions and their current needs from their recent interactions. Figure 1 shows an example of the personalised multi-modal interactive recommendation with visual recommendations and the corresponding natural-language critiques. In particular, Figure 1 (a) demonstrates the users' past interests with the shopping history recorded by the recommender system and their current needs with the next item that they wish to purchase (the next target item). Next, Figure 1 (b) illustrates the real-time interactions between a recommender system and a user. The recommender system initiates the conversation by presenting a list of personalised initial recommendations to the user. Subsequently, during each interaction turn, the user provides natural-language critiques regarding the visual recommendation list in order to achieve items with more preferred features. An effective MMIRS will improve the users' experience substantially and will save users much efforts in finding their target items.

Despite the recent advances in incorporating the users' current needs (i.e. the target items) from the informative multi-modal information across the multi-turn interactions, we argue that it is typically challenging to make satisfactory personalised recommendations due to the difficulty in balancing the users' past interests and the current needs for generating the users' state (i.e. current preferences) representations over time. Indeed, the existing MMIRSs [16, 49, 51, 52] typically simplify the multi-modal interactive recommendation task by initiating conversations using randomly sampled recommendations irrespective of the users' interaction histories (i.e. the past interests), thereby only focusing on seeking the target item (i.e. the current needs) across real-time interactions. Although providing next-item recommendations

from sequential user-item interaction history is one of the most common use cases in the recommender system domain, the existing sequential and session-aware recommendation models [19, 20, 23, 41] currently only consider the explicit/implicit past user-item interactions (such as purchases and clicks) in the sequence modelling. In addition, these sequential/session-aware recommendation models have shown difficulties in learning sequential patterns over *cold-start* users (who have very limited historical interactions) compared to *warm-start* users (who have longer interaction sequences) [46, 64]. An obvious and simple solution for the personalised MMIR task is to conduct a pipeline, where a sequential/session-aware recommendation model (such as GRU4Rec [20]) generates the initial personalised recommendations and a multi-modal interactive recommendation model (such as EGE [51]) updates the subsequent recommendations across the multi-turn interactions. However, such pipeline-based recommender systems cannot effectively benefit from a proper cooperation between the sequential/session-aware recommendation models and the multi-modal interactive recommendation models when there is a shift between the users' past interests and their current needs (in particular with cold-start users), thereby possibly failing to provide satisfactory personalised recommendations over time.

Deep reinforcement learning (DRL) allows a recommender system (i.e. an agent) to actively interact with a user (i.e. the environment) while learning from the user's real-time feedback to infer the user's dynamic preferences. A variety of DRL algorithms has been successfully applied in various recommender system domains, such as e-commerce [55], video [7] and music recommendations [27]. In particular, recent research on multi-modal interactive recommendation (MMIR) has formulated the MMIR task with various DRL algorithms as MDPs [16], POMDPs [51], CMDPs [62] or multi-armed bandits [59]. However, all of these only consider a specific recommendation scenario where the users are all cold-start users, i.e. without using any interaction history. Indeed, the existing DRL-based recommender systems are not able to deal with the personalised multi-modal interactive recommendation task in an end-to-end fashion considering the computational complexity of learning users' the past interests from the interaction history and estimating the users' current needs from the real-time interactions. Hierarchical reinforcement learning (HRL) [21, 35] can decompose a complex task into a hierarchy of subtasks as semi-Markov decision processes (SMDPs), which reduces the computational complexity. Such a HRL formulation with a hierarchy of subtasks is particularly suitable for the multi-modal interactive task that requires to address different subtasks over time by either estimating the users' past interests or tracking the users' current needs. For instance, the "Options" framework of HRL provides a generic way for task decomposition where options represent closed-loop sub-behaviours that are carried out for multiple timesteps until the termination condition is triggered [21]. However, to the best of our knowledge, no prior work has investigated HRL in the multi-modal interactive recommendation task.

In this paper, we present our formulation of the personalised MMIR task as a semi-Markov decision process (SMDP) by simulating both the past and real-time interactions between a user (i.e. an environment) and a recommender system (i.e. an agent). To this end, we propose a novel personalised multi-modal interactive recommendation model (PMMIR) using hierarchical reinforcement learning to more effectively incorporate the users' preferences from both their past and real-time interactions. In particular, the proposed PMMIR model uses the Options framework of HRL to decompose the personalised interactive recommendation process into a sequence of two subtasks with hierarchical state representations: a first subtask where a *history encoder* learns the users' past interests with the *hidden states of history* for providing personalised initial recommendations, and a second subtask where a *state tracker* estimates the current needs with the *real-time estimated states* for updating the subsequent recommendations. The history encoder and the state tracker are jointly optimised using a typical policy gradient approach (i.e. REINFORCE [6]) with a single optimisation objective by maximising the users' future satisfaction with the recommendations (i.e. the cumulative future rewards). Following previous work [16, 49, 51], our PMMIR model is trained and evaluated by adopting a user simulator, which is capable of

producing natural-language critiques regarding the recommendations. This surrogate simulates the behaviour of real human users [16]. By conducting experiments on two fashion datasets derived from two well-known public datasets, we observe that our proposed PMMIR model outperforms existing state-of-the-art baseline models, leading to significant improvements. In short, we summarise the main contributions of this paper as follows:

- We propose a novel personalised multi-modal interactive recommendation model (PMMIR) that effectively integrates the users' preferences obtained from both past and real-time interactions by leveraging HRL with the Options framework.
- Our proposed PMMIR model decomposes the MMIR task into two subtasks: an initial personalised recommendation with the users' past interests and several subsequent recommendations with the users' current needs.
- We derive two fashion datasets (i.e. Amazon-Shoes and Amazon-Dresses) for providing the users' interaction histories from two well-known public datasets since there is no existing dataset suitable for the personalisation setting of the multi-modal interactive recommendation task.
- Through extensive empirical evaluations conducted on the personalised MMIR task, our proposed PMMIR model demonstrates significant improvements over existing state-of-the-art approaches. We also show that both cold-start and warm-start users can benefit from our proposed PMMIR model in terms of recommendation effectiveness.

The paper is structured as follows: Section 2 provides a comprehensive review of the related work and highlights the contributions of our research in relation to the existing literature; In Section 3, we define the problem formulation and introduce our proposed PMMIR model; The experimental setup and results are presented in Sections 4 and 5, respectively; Finally, Section 6 summarises our findings.

2 RELATED WORK

Within this section, our primary focus is to introduce the concept of multi-modal interactive recommendation (MMIR). Then we discuss personalisation in interactive recommendation. Finally, we describe hierarchical reinforcement learning.

Multi-Modal Interactive Recommendation. Interactive recommender systems have been shown to be more effective in incorporating the users' dynamic preferences over time from their explicit and implicit real-time feedback (such as natural-language critiques and clicks) compared to static/traditional recommender systems that predict the users' preferences by mining the users' past behaviours offline (such as ratings, clicks, and purchases) [14]. In addition, multi-modal recommender systems can handle information with various modalities either from items (such as images and textual descriptions) or users (such as natural-language feedback) to effectively alleviate the problems of data sparsity and cold start [31, 65]. Therefore, multi-modal interactive recommender systems (MMIRSs) can effectively track/estimate the users' dynamic preferences from the informative information with different modalities across real-time interactions. As an example, Guo et al. [16] were among the first to tackle the MMIR task by introducing a Dialog Manager (DM) model that combined supervised pre-training and model-based policy improvement (MBPI). This approach aimed to effectively capture the users' preferences across multiple interaction turns by considering both visual recommendations and the corresponding natural-language critiques. Since then, research has focussed upon improving the recommendation performance by either formulating the MMIR task using various reinforcement learning approaches (such as CMDPs [62], multi-armed bandits [58] and POMDPs [51]) or adopting more advanced state tracking components (such as Transformer [49] and RNN-enhanced Transformer [52]). Unlike the uni-modal

(text-based) conversational recommendation task [27, 42], which usually leverages attribute-based clarification questions to elicit the users' preferences, the multi-modal interactive recommendation task addressed in this paper takes the critiquing-based task formulation by incorporating the users' preferences from their natural-language feedback.

Personalisation in Interactive Recommendation. The above-existing MMIR models only focus on incorporating the users' current needs across the multi-turn real-time interactions but omit their past behaviours, by initially presenting users with randomly selected items at the start of the interaction process. Meanwhile, a variety of interactive recommendation models have leveraged the users' past behaviours for personalised recommendations during the multi-turn interaction processes. For instance, the Estimation-Action-Reflection (EAR) model by Lei et al. [27] (a typical question-based interactive recommendation model [14]) leveraged the factorisation machine (FM) [39] to estimate the users' preferences with the users' past behaviours for predicting further preferred items and attributes. The users' online feedback is incorporated by feeding the accepted attributes back to FM to make a new prediction of items and attributes again or using the rejected items as negative signals for training FM again. However, such an FM-based method for the question-based interactive recommendation task is infeasible for our multi-modal interactive recommendation task, which leverages natural-language critiquing sentences freely expressed by the users rather than the brief terms of well-categorised attributes. In addition, a simple solution for the personalised multi-modal interactive recommendation task is to combine the sequential recommendation models (such as GRU4Rec [20]) with the multi-modal interactive recommendation models (such as EGE [51]) in a pipeline. For instance, GRU4Rec can be leveraged for generating the initial personalised recommendations, while EGE can be utilised for updating the subsequent recommendation across the multi-turn real-time interactions. However, we argue that such pipeline-based recommender systems are fragile at providing satisfactory personalised recommendations over time when there is a shift between the users' past interests and current needs since their components are optimised independently.

Furthermore, session-aware recommendation models [22, 26, 37, 47] decouple the users' long-term and short-term preferences for making better-personalised recommendations by exploiting the relationship between sessions for each user. For instance, Quadrana et al. [37] proposed a Hierarchical Recurrent Neural Network model (HRNN) for the personalised session-based recommendations. The HRNN model is structured with a hierarchy of two-level Gated Recurrent Units (GRUs): the session-level GRU that makes recommendations by tracking the user interactions within sessions; and the user-level GRU that tracks the evolution of the users' preferences across sessions. When a new session starts, the hidden state of the user-level GRU is used to initialise the session-level GRU, thereby providing personalisation capabilities to the session-level GRU. Such a hierarchy of two-level GRUs structure can also be leveraged in the multi-modal interactive recommendation task to make personalised recommendations over time. Therefore, we are inspired by the hierarchy of two-level GRUs structure to propose an effective end-to-end multi-modal interactive recommendation model with a dual GRUs/Transformers structure that can make personalised recommendations over time by incorporating both the users' past behaviours and the informative multi-modal information from real-time interactions. The HRNN model with two-level GRUs adopts a supervised learning approach for jointly optimising the user-level and session-level GRUs, which is less effective than the DRL approaches for maximising the future rewards [1, 8, 30].

Hierarchical Reinforcement Learning. Deep reinforcement learning (DRL) has been widely adopted in the recommendation field with various DRL algorithms, such as Deep Q-learning Network (DQN) [33], REINFORCE [48], and Actor-Critic [25], for coping with the users' dynamic preferences over time and maximising their long-term engagements [1, 8, 30]. In particular, the MMIR task has been formulated with various DRL algorithms as MDPs [16],

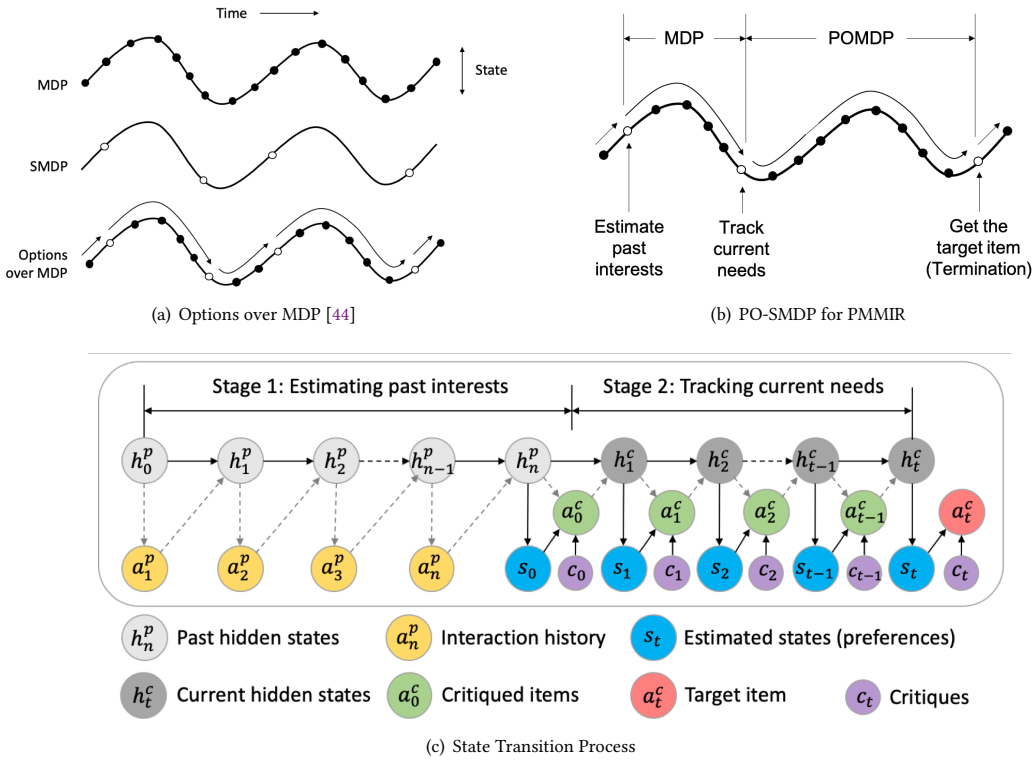
POMDPs [51], CMDPs [62] or multi-armed bandits [59] to simulate the multi-turn interactions between the recommender systems and the users. However, the existing MMIR models (e.g., MBPI [16], EGE [51], and RCR [62]) with DRL can only maximise the cumulative rewards when dealing with real-time requests within the conversational session, while simplifying the MMIR task by omitting the users' past interests. Indeed, making personalised recommendations across multi-turn interactions considering the users' past interests and current needs is a complex task. Hierarchical reinforcement learning provides a solution for decomposing a complex task into a hierarchy of easily addressed subtasks as semi-Markov decision processes (SMDPs) with various frameworks, such as Options [44], Hierarchical of Abstract Machines (HAMs) [34], and MAXQ value function decomposition [12]. The existing recommender systems with HRL [15, 29, 54, 63] typically formulate the recommendation task with two levels of hierarchies where a high-level agent (the so-called meta-controller) determines the subtasks and a low-level agent (the so-called controller) addresses the subtasks. For instance, CEI [15] formulates the conversational recommendation task with the Options framework using a meta-controller to select a type of subtasks (chitchat or recommendation) and a controller to provide subtask-specific actions (i.e. response for chitchat or candidate items for recommendation). In addition, recent research on question-based conversational recommendations (such as EAR [27] and FPAN [57]) follows a two-level architecture with a policy network as a meta-controller to decide either to ask for more information or to recommend items and a Factorisation Machine (FM) [39] as a controller to generate a set of recommendations [14]. Different from the standard HRL models, these question-based conversational recommendation models [14, 27, 57] only optimise the meta-controller with RL algorithms (such as REINFORCE [48]) to manage the conversational system, while the controller is separately optimised with supervised learning approaches (such as BPR [40]). However, to the best of our knowledge, no prior work has investigated HRL in the multi-modal interactive recommendation task. In this paper, we leverage HRL with the Options framework by proposing a personalised multi-modal interactive recommendation model (PMMIR) to effectively incorporate the users' past interests and their evolving current needs over time. In particular, the high-level agent for determining the subtasks is fully driven by the users' natural-language feedback (we will describe this in Section 3). Therefore, we mainly focus on modelling the cooperation of the low-level agents for estimating the users' past interests and tracking the users' current needs over time in our proposed PMMIR model.

3 THE PMMIR MODEL

In this section, we begin by formulating the problem of the multi-modal interactive recommendation task using hierarchical reinforcement learning within the framework of partially observable semi-Markov decision processes (PO-SMDP) and we introduce the notations used in our formulation (Section 3.1). Then, in Section 3.2, we propose a novel personalised multi-modal interactive recommendation model (PMMIR) using dual GRUs, as well as dual Transformers, to effectively incorporate the users' preferences from both past interests through the interaction history and the current needs via the real-time interactions. Finally, we define the rewards and describe the learning algorithm for the multi-modal interactive recommendation scenario (Section 3.3).

3.1 Preliminaries

Our research focuses on investigating the personalised multi-modal interactive recommendation (MMIR) task within a hierarchical reinforcement learning (HRL) formulation, specifically utilising the Options framework [44] in a partially observable environment. In such an environment, the users' preferences can only be partially expressed with the natural-language critiques at each turn [51]. Figure 2 (b) & (c) illustrate the state transition process with hierarchical state representations for the personalised MMIR task.



(a) Options over MDP [44]

(b) PO-SMDP for PMMIR

(c) State Transition Process

Fig. 2. State transition process with hierarchical state representations for the personalised MMIR task.

3.1.1 PO-SMDP for Personalised MMIR. Figure 2 (a) shows the extension of a Markov decision process (MDP) with *options* (i.e. closed-loop policies for taking action over a period of time [44]) into a semi-Markov decision process (SMDP). In particular, the state trajectory of an MDP is made up of discrete-time transitions. Meanwhile, SMDP is a type of MDP suitable for modelling continuous-time discrete-event systems, therefore its state trajectory consists of continuous-time transitions. Sutton et al. [44] defined a set of options over an MDP as a semi-Markov decision process (SMDP), which enables an MDP trajectory to be analysed in either discrete-time transitions or continuous-time transitions. In this paper, we adopt a partially observable semi-Markov decision process (PO-SMDP, as shown in Figure 2 (b)) for the personalised MMIR task with two *low-level agents* for addressing the subtasks: (1) estimating the users' past interests from their interaction history using a *history encoder* as a Markov decision process (MDP), and (2) tracking the users' current needs from the real-time interactions using a *state tracker* as a partially observable Markov decision process (POMDP). In the initial stage, the users' preferences are *fully observed* (i.e. as an MDP), since each item the user has interacted with (e.g., purchased fashion products) can be seen as their preferences at that time. However, in the subsequent stage of tracking current needs, the users' preferences are only *partially observed* (i.e. as a POMDP) since they can only be expressed partially through their natural-language feedback in relation to the critiqued items. The subtasks for taking actions can be selected in sequence with a fixed *high-level agent* according to the users' requests in natural language following the example of the interaction process in Figure 1. The history encoder is initiated as a

one-step option for the initial personalised recommendations corresponding to the request for recommending "some shoes for women" in Figure 1. The history encoder is then terminated and the state tracker is initiated when the user requests "shoes that are brown leather with an ankle strap". Since the high-level agent for determining the subtasks is fully driven by the users' natural-language feedback, we mainly focus on modelling the cooperation of the low-level agents for addressing the MMIR task.

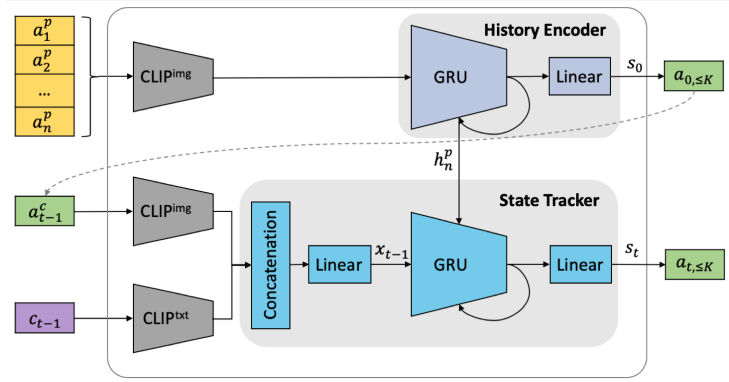
3.1.2 Notations. We specifically approach the multi-modal interactive recommendation (MMIR) process as a partially observable semi-Markov decision process (PO-SMDP) with a tuple consisting of eight elements $(\mathcal{S}, \mathcal{A}, C, O, \mathcal{R}, \mathcal{T}, \mathcal{P}, \gamma)$, where:

- $\mathcal{S}$ is a set of *states* (i.e. the users' preferences),
- $\mathcal{A}$ is a set of *actions* (i.e. the items for recommendations),
- C is a set of *observations* (i.e. the users' natural-language critiques),
- O is a set of *options* (i.e. options for selecting subtasks, either estimating past interests or tracking current needs),
- $\mathcal{R}$ is the *reward function*,
- $\mathcal{T}$ is a set of transition probabilities between states,
- $\mathcal{P}$ is a set of transition probabilities between options, and
- $\gamma \in [0, 1]$ is the *discount factor* for future rewards.

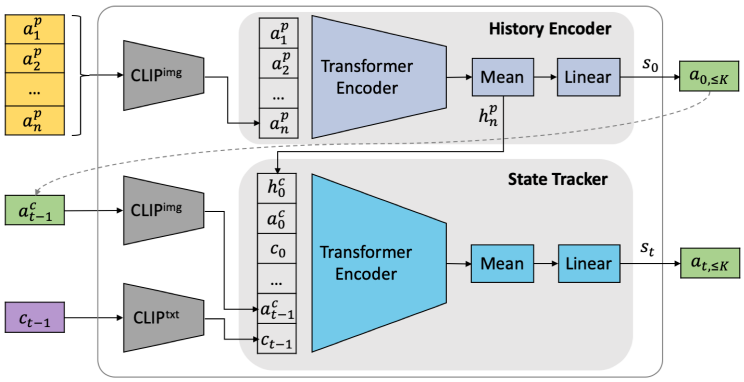
The estimated users' preferences at turn t are denoted by $s_t \in \mathcal{S}$. When the recommender system (i.e. the agent) provides a ranking of K items, $a_t \in \mathcal{A}$ $(a_{t,\leq K} = (a_{t,1}, ..., a_{t,K}))$ and receives a natural-language critique $c_t \in C$ and a reward $r_t \sim \mathcal{R}(s_t, a_t)$, the estimated preferences s_t change in accordance with the transition distribution, $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t, c_t)$. A recommender system acts according to its policy $\pi(a_{t+1}|a_{\leq t}, c_{\leq t})$ by returning the probability of selecting action a_t at turn t, where $a_{\leq t} = (a_0, ..., a_t)$ and $c_{\leq t} = (c_0, ..., c_t)$ are the action and critique histories, respectively. Figure 2 (b) shows that the personalised multi-modal interactive recommendation process starts with the past interests s_0 estimated from the users' interaction history $(a_1^p, ..., a_n^p)$ with the past hidden states $(h_0^p, ..., h_n^p)$ while following with the current needs $s_t (t \neq 0)$ tracked from the users' real-time interactions (i.e. the sequence of the critiqued items $(a_0^c, ..., a_t^c)$ and the sequence of the corresponding critiques $(c_0, ..., c_t)$) with the current hidden states $(h_0^c, ..., h_t^c)$. Generally, for a partially observable semi-Markov decision process (PO-SMDP), the recommender system's goal is to learn policies π_ϕ (i.e. the history encoder) and π_θ (i.e. the state tracker) that maximise the expected future return over trajectories $\tau = ((a_{0,\leq K}, c_0), ..., (a_{T,\leq K}, c_T))$ induced by the policies. Note that we assume that the users seek a single target item based on its visual features, have a single history session for estimating the past interests, and interact with the recommender system within a single interaction session. We leave the handling of more complex situations (such as multiple target items based on both visual & non-visual features (such as brands, prices and sizes) across multiple interaction sessions) in the multi-modal interactive recommendation task as interesting future work.

3.2 The Model Architecture

We propose a personalised multi-modal interactive recommendation model (PMMIR) comprising multi-modal encoders, a history encoder, and a state tracker. In particular, both GRU and Transformer encoders are two popular neural networks for sequence modelling and state tracking. Therefore, our proposed PMMIR model can adopt either GRU or Transformer as the history encoder and/or state tracker. Here, we consider two versions of PMMIR: PMMIR$_{GRU}$ with GRUs only and PMMIR$_{Transformer}$ with Transformers only. Figure 3 shows our proposed end-to-end personalised multi-modal interactive recommendation model (PMMIR) with hierarchical state representations based on GRUs



(a) PMMIR$_{GRU}$



(b) PMMIR$_{Transformer}$

Fig. 3. The proposed personalised multi-modal interactive recommendation (PMMIR) model with hierarchical state representations.

(Figure 3 (a) with PMMIR$_{GRU}$) and Transformers (Figure 3 (b) with PMMIR$_{Transformer}$). In the following, we describe the major components of our PMMIR models.

The Multi-Modal Encoders. To properly represent the system' recommendations and the users' feedback, we leverage visual and textual encoders for encoding the images of the recommendations and the natural-language critiques into embedded vector representations, respectively. In particular, both images of recommendations and natural-language critiques made by users can be encoded with a pre-trained vision-language model, called CLIP [38], as the unified visual and textual representations. There are also other alternatives for the multi-modal encoders [16, 49, 51], for instance the pre-trained language models (such as GloVe [36] and BERT [11]) for text and the pre-trained vision models (such as ResNet [18] and ViT [13]) for images. Compared to these alternative encoders, CLIP has the capability of providing a single representation vector for each modality with the same dimensionality. CLIP has been shown to be effective in capturing the fine-grained features of fashion products, such as shirts and dresses, in the conditioned and combined image retrieval tasks [2, 3]. We denote the multi-modal encoders for encoding a visual item a as $a^{'} = CLIP^{img}(a)$ and a

textual critique c as $c^{'} = CLIP^{txt}(c)$. Note that we directly use a and c to denote their encoded representations (i.e. $a^{'}$ and $c^{'}$), respectively.

The History Encoder. The users' interaction history (i.e. a sequence of the interacted items $a^p_{1:n} = (a^p_1, ..., a^p_n)$) can be first encoded with the above visual encoder $CLIP^{img}(\cdot)$. To estimate the users' past interests, we adopt a gated recurrent unit (GRU) [9] as the history encoder (similar to the GRU4Rec [20] model for sequential recommendations) for encoding the past hidden states as follows:

$$h^p_n = GRU^{past}(h^p_{n-1}, a^p_n) \tag{1}$$

The last hidden state h^p_n of $GRU^{past}(\cdot)$ is further mapped with a linear layer as the overall-representation of the users' past interests (i.e. the initial state $s_0 = Linear(tanh(h^p_n))$ for the MMIR task).

Alternatively, we can adopt a Transformer encoder [45] as the history encoder (similar to the SASRec [23] model for sequential recommendations) by directly processing the sequence of the interacted items $a^p_{1:n}$ as the input, while averaging the output embeddings with $Mean(\cdot)$. Note that we also use h^p_n to denote the estimated historical preferences using a Transformer encoder as follows:

$$h^p_n = Mean(Transformer^{past}(a^p_{1:n})) \tag{2}$$

The State Tracker. To incorporate the users' current needs over time from the visual recommendations and the corresponding natural-language feedback, we leverage a simple concatenation operation for the multi-modal feature fusion, as in [16, 49] and then a state tracker (either based on a GRU [16, 51] or a Transformer encoder [49, 53]) for estimating the users' interaction states. In particular, both the visual and textual representations are concatenated and then mapped into a low dimensional space as input to a subsequent GRU-based state tracker to model the user's current needs at each turn t.

$$x_{t-1} = Linear([a^c_{t-1}, c_{t-1}]) \tag{3}$$

$$h^c_t = GRU^{current}(h^c_{t-1}, x_{t-1}) \tag{4}$$

We argue that the users usually hold a certain preference state (such as the estimated past preference state h^p_n) when they start seeking their current needs in a real-time interaction session. To this end, the initial hidden state h^c_0 of the state tracker $GRU^{current}(\cdot)$ can be initialised by the last hidden state h^p_n of the history encoder $GRU^{past}(\cdot)$, that is $h^c_0 = h^p_n$. In addition, the hidden state h^c_t at each turn t ($t \neq 0$) is further mapped with a linear layer into the estimated users' current needs (i.e. $s_t = Linear(h^c_t)$).

Similarly, a Transformer-based state tracker concatenates and encodes all previous visual and textual representations:

$$h^c_t = Mean(Transformer^{current}([h^c_0, a^c_0, c_0, ..., a^c_{t-1}, c_{t-1}])) \tag{5}$$

The last hidden state h^p_n of the history encoder $Transformer^{past}(\cdot)$ is concatenated as the input of $Transformer^{current}(\cdot)$, that is $h^c_0 = h^p_n$. In addition, the hidden state h^c_t at each turn t ($t \neq 0$) is further mapped with a linear layer into the estimated users' current needs (i.e. $s_t = Linear(tanh(h^c_t))$).

Considering the estimated state s_t representing the user's preferences, we adopt a greedy policy [16, 51] by recommending the top-K candidate items $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$ for the next action. More specifically, we choose the top-K items that are closest to s_t in the multi-modal (i.e. visual and textual) feature space using the Euclidean distance: $a_{t,\leq K} \sim KNNs(s_t)$, where $KNNs(\cdot)$ represents a softmax distribution over the top-K nearest neighbours of s_t and

$a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$. Furthermore, we incorporate a post-filtering step to eliminate any candidate item from the ranking list that has already been shown to the user based on the real-time interaction history $a_{\leq t}$ as [51].

3.3 The Learning Algorithm

To optimise PMMIR, we leverage a two-stage optimisation method following [16, 51] with a supervised learning (SL) loss for initialising the policies and then a reinforcement learning (RL) loss for further improving the performances.

3.3.1 Supervised Learning. We initialise PMMIR with a supervised pre-training process to improve the sample efficiency during the RL training process. In particular, we leverage a triplet loss objective $L(\pi_\phi, \pi_\theta)$ as in [16, 51] to jointly pre-train the recommendation policies π_ϕ (for estimating the past interests) and π_θ (for tracking the current needs):

$$\max L(\pi_\phi, \pi_\theta) = \sum_{t=0}^{T} \max\left(0, l_2(s_t, a^+) - l_2(s_t, a^-) + \epsilon\right) \tag{6}$$

where $\phi \in \mathbb{R}$ and $\theta \in \mathbb{R}$ denote policy parameters. $l_2(\cdot)$ denotes the l_2 distance. a^+ is the target item and a^- is a randomly sampled item from the candidate pool. ϵ is a constant for the margin to keep the negative samples a^- far apart.

3.3.2 Reinforcement Learning. The objective of policy optimisation with RL is to find the target item via the policies π_ϕ and π_θ that maximise the expectation of the cumulative return:

$$\max J(\pi_\phi, \pi_\theta) = \max \mathop{\mathbb{E}}_{\tau \sim \pi_\phi, \pi_\theta}[R(\tau)], \ \textit{where } R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_{t,\leq K}) \tag{7}$$

where $R(\tau)$ is the discounted cumulative reward, and T is the maximum turn in the interaction trajectory. The expectation is taken over trajectories $\tau = ((a_{0,\leq K}, c_0), ..., (a_{T,\leq K}, c_T))$.

We adopt a policy gradient method (e.g., REINFORCE [48]) for PO-SMDP to further optimise our PMMIR model. Indeed, the policy gradient methods have been shown to be more stable with a small learning rate [6] compared to the value-based methods (such as DQN [33]). Specifically, the gradient of Equation (7) can be computed as follows:

$$\nabla J(\pi_\phi, \pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\phi, \pi_\theta}[\sum_{t=0}^{T} \nabla \log \pi(a_{t,\leq K}|s_t) R(\tau)] \tag{8}$$

We define $\log \pi(a_{t,\leq K}|s_t)$ as a softmax cross-entropy objective to identify the positive sample (i.e. the target item a^+) amongst a set of hard negative samples (i.e. the rejected items a_j^- ($j \in [1, J]$)):

$$\log \pi(a_{t,\leq K}|s_t) = \log\left(\frac{e^{sim(s_t, a^+)}}{e^{sim(s_t, a^+)} + \sum_{j=1}^{J} e^{sim(s_t, a_j^-)}}\right) \tag{9}$$

where $sim(\cdot)$ is a similarity kernel that can be the dot product or the negative l_2 distance in our experiments.

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

$$r(s_t, a_{t,\leq K}) = \sum_{i=1}^{K} sim(a_{t,i}, a^+) \tag{10}$$

3.3.3 Training Procedure. We also present the training procedure of our PMMIR model for PO-SMDP with REINFORCE in Algorithm 1. To facilitate the training processes, a user simulator [16, 49] is adopted as a substitute for real human users. Further information regarding the specific user simulator employed is discussed in Section 4.2. As shown in

Algorithm 1, the recommender policies π_ϕ and π_θ aim to maximise the expected rewards by properly cooperating with each other.

Algorithm 1 Training procedure of PMMIR

Input: User-item interaction sequence set $\mathcal{X}$, history encoder π_ϕ, and state tracker π_θ, discount factor γ, learning rates $\eta_{sl} > \eta_{rl}$
Output: All learned parameters ϕ, and θ

Initialise all trainable parameters
Pre-train π_ϕ & π_θ with Eq. (6)
Load all parameters with weights from pre-training
repeat
    Draw a batch of $(a^p_{1:n}, a^{target})$ from $\mathcal{X}$
    Start with π_ϕ for estimating the past interests
    Generate h^p_n from $a^p_{1:n}$ with Eq. (1)/Eq. (2)
    Map h^p_n into s_0
    Switch into π_θ for tracking the current needs
    Initialise $h^c_0 = h^p_n$
    for t = 0, 1, ... T **do**
        Sample $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$ with s_t
        Receive a critique c_t with a user simulator
        Calculate a reward $r(s_t, a_{t,\leq K})$ with Eq. (10)
        if t==0 **then**
            Calculate $\log \pi(s_t, a_t; \phi)$ with Eq. (9)
        else
            Calculate $\log \pi(s_t, a_t; \theta)$ with Eq. (9)
        end if
        Estimate and update next state s_{t+1}
    end for
    Calculate $R(\tau)$ with Eq. (7)
    Perform updates by $\nabla J(\pi_\phi, \pi_\theta)$ with Eq. (8)
until converge
return all parameters of policies ϕ, and θ

4 EXPERIMENTAL SETUP

We proceed to evaluate the effectiveness of our proposed PMMIR model, along with its two variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$), in comparison to existing approaches from the literature. In particular, we aim to address the following three research questions:

• RQ1: Is there a significant improvement in the performance of our proposed PMMIR model compared to the existing state-of-the-art baseline models in the multi-modal interactive recommendation task?
• RQ2: Can both cold-start and warm-start users benefit from our proposed PMMIR model?
• RQ3: What are the impacts of the components of the PMMIR model (such as $h^c_0 = h^p_n$ and CLIP backbones) and the introduced hyper-parameters (such as γ & K) on the overall performance?

Table 1. Datasets' statistics.

Dataset	Total Items	Train Users	Test Users	Lengths
Amazon-Shoes	31,940	14,892	3,722	3-9
Amazon-Dresses	18,501	13,657	3,414	4-9

4.1 Datasets & Setup

Datasets. Since there is no existing dataset suitable for the personalisation setting of the multi-modal interactive recommendation task, we derive two datasets (i.e. Amazon-Shoes and Amazon-Dresses) for providing the user-item interaction sequences from two well-known public fashion datasets, i.e. Amazon Review Data (2014)[1] and Amazon Review Data (2018)[2] with the "Clothing, Shoes and Jewelry" category. In particular, we derive the Amazon-Shoes dataset by including various types of shoes for women (such as "Athletic", "Boot", "Clog", "Flat", "Heel", "Pump", "Sneaker", "Stiletto", and "Wedding") from the "Clothing, Shoes and Jewelry" category of Amazon Review Data (2014). Meanwhile, we also derive the Amazon-Dresses dataset by including the fashion products with the "dress" label for women from the "Clothing, Shoes and Jewelry" category of Amazon Review Data (2018). On both derived datasets, we construct the user-item interaction sequences by concatenating the IDs of a user's purchased items according to their interaction timestamps. Table 1 summarises the statistics of the Amazon-Shoes and Amazon-Dresses datasets. Our both derived datasets are open to the public via the anonymised link in the abstract. Both datasets provide an image for each fashion product. In addition, for training/testing the user simulators, we use two well-known fashion datasets, namely the *Shoes* [4, 16] and *Fashion IQ Dresses* [49] datasets (discussed further in Section 4.2) for relative captioning with the provided triplets (i.e. $\langle a_{target}, a_{candidate}, c_{caption} \rangle$). The relative captions ($c_{caption}$) of the image pairs (a_{target} and $a_{candidate}$) describe the attributes of the target item a_{target} that is missing in candidate item $a_{candidate}$ in natural language, and have been written by real users via crowd-sourcing. The *Shoes* dataset contains 10,751 triplets in total, while the *Fashion IQ Dresses* dataset provides 11,970 and 4,034 triplets for training and testing, respectively. Note that the triplets in the *Shoes* and *Fashion IQ Dresses* datasets for training the user simulators do not include any data from our derived *Amazon-Shoes* and *Amazon-Dresses* datasets, which are used for training the recommendation models.

Setup. As described in Algorithm 1, we leverage a two-stage training procedure for optimising the PMMIR model following [16, 51]. In particular, we first pre-train and initialise the PMMIR model with the supervised learning (SL) setting using a learning rate $\eta_{sl} = 10^{-3}$ [16] and then further optimise the PMMIR model in the reinforcement learning (RL) setting using a learning rate $\eta_{rl} = 10^{-5}$ [16]. We use Adam [24] with Eq. (6) and Eq. (8) for optimising the PMMIR model's parameters, respectively. The pre-trained CLIP image and text encoders are loaded with the "ViT-B/32" checkpoint[3], and the visual and textual embedding dimensionalities of the multi-modal feature space are both set to 512. The initial hidden state h_0^p of the history encoder $GRU^{past}(\cdot)$ in PMMIR$_{GRU}$ is initialised with zeros. Meanwhile, PMMIR$_{Transformer}$ does not have such an explicit initial hidden state h_0^p of the history encoder $Transformer^{past}(\cdot)$. Indeed, PMMIR$_{Transformer}$ directly takes the sequence of the interacted items as the input. The batch size is set to 128 following the setting in [16]. The maximum number of epochs for SL & RL training is set to 20 with early stopping as in [52], while the maximum number of interaction turns is set to 10 as in [51, 52]. At each interaction turn for both training and testing, the recommender system provides the top-K (i.e. $K = 3$) items as a recommendation. For the RL stage, the number of hard negative samples (i.e. J) is set to 5, following [51]. The similarity kernel $sim(\cdot)$ in Equation (9)

[1] http://jmcauley.ucsd.edu/data/amazon/index_2014.html [2] https://nijianmo.github.io/amazon/ [3] https://github.com/openai/CLIP

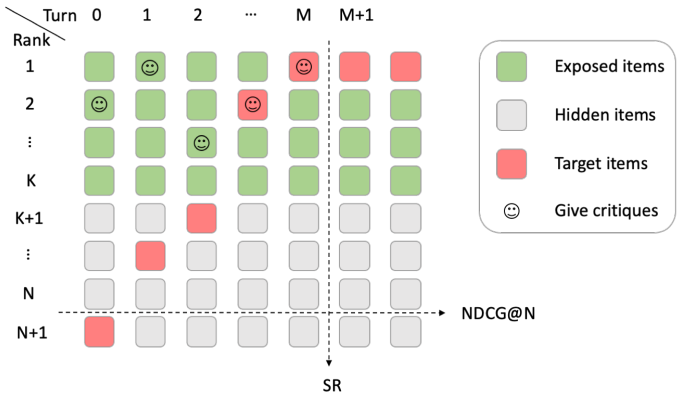


Fig. 4. Online evaluation with top-K recommendations across multi-turn interactions in the personalised MMIR scenario.

is set to be the dot product by default with the normalised visual and textual representations. If not mentioned otherwise, the discount factor γ is set to 0.2. We consider users with the least interactions (3 interactions on Amazon-Shoes and 4 interactions on Amazon-Dresses) as cold-start users, while the other users with longer interaction sequences are considered as warm-start users. For each user-item interaction sequence, we leave the last interaction as the user's target item (i.e. the current needs) and the previous sequence of interactions as the users' interaction history (i.e. the past interests).

4.2 Online Evaluation & Metrics

Online Evaluation. The success of the personalised MMIR task is measured by the number of interaction turns to obtain the target item(s) and the rank of the target item(s) in each interaction turn. We evaluate the effectiveness of our proposed PMMIR model for personalised multi-modal interactive recommendation in comparison to the existing approaches from the literature based on an online evaluation approach. Figure 4 shows an example of online evaluation with top-K (e.g., $K = 3$) recommendation across multi-turn interactions in the personalised MMIR scenario. In this scenario, the recommender system ranks all the items and shows the top-K items as the recommendations at each turn. Meanwhile, a user browses the exposed top-K items, gives a natural-language critique on the most preferred item and rejects the others at each turn. In particular, the figure illustrates how a user can find the desired item through multi-turn interactions. Following the methodology in [51, 52, 62], we measure the effectiveness of the interactive recommendation models at interaction turn M. On the other hand, the user may check more items in the ranking list at each turn, down to rank N.

User Simulators. In both the optimisation and evaluation processes, user simulators have been employed as substitutes for real human users in the context of relative captioning tasks [16, 49, 53, 62]. A user simulator based on relative captioning can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images as users' natural-language feedback. This natural-language feedback generation process within the user simulator closely resembles a shopping conversation between a customer and a shopping assistant. The rewards returned by the user simulators are calculated using Eq. (10). The user simulator can actively interact with the recommender system to provide various real-time natural-language feedback, thereby allowing to learn satisfactory

multi-modal interactive recommender systems with enough training data. In particular, we adopt a user simulator with the Show, Attend, & Tell [56] model trained with triplets from *Shoes* by using the checkpoint[4] [4, 16] provided by Guo et al. [16]. In addition, we adopt the VL-Transformer model introduced in [49, 52] as a user simulator, specifically trained on triplets extracted from the *Fashion IQ Dresses* dataset, following the setting[5] in [49, 52] and using the checkpoint provided by Wu et al. [52]. Both user simulators are deployed by using an image captioning tool (called ImageCaptioning.pytorch[6] [32]). The user simulators are intensively trained using crowdsourced relative expressions to describe the visual distinctions between pairs of images [16, 49, 52]. Moreover, the pre-trained user simulators have previously been thoroughly assessed through both quantitative evaluations and user studies, making them a reliable substitute for real users in conducting evaluation experiments [16, 49, 51]. Following [16, 49, 53], we assume that the user simulator only gives a natural-language critique on a single recommended item (the most similar to the target item) at each turn by describing the desired attributes in the target item that are missing in the recommended item. Such simplification is necessitated by the existing available datasets and the availability of accurate user simulators.

Metrics. We measure the effectiveness of the multi-modal interactive recommendations at interaction turn M in terms of Normalised Discounted Cumulative Gain (NDCG@N truncated at rank $N = 3$) and Success Rate (SR). To assess the quality of the ranking lists, the Normalized Discounted Cumulative Gain (NDCG) metric emphasises the importance of higher ranks compared to lower ones. On the other hand, the Success Rate (SR) metric measures the percentage of users for whom the target image was successfully retrieved within a specific number of interactions, denoted as M within the range of 1 to 10. For significance testing, we employ both evaluation metrics, namely NDCG@3 and SR, at the 5th and 10th interaction turns.

4.3 Baselines

We conduct a comparative analysis between our proposed PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and existing state-of-the-art baseline models, including their extensions, for the multi-modal interactive recommendation (MMIR) task.

The first group of baseline models are all based on GRUs in order to compare with PMMIR$_{GRU}$:

• **GRU**$_{hist}$: The GRU$_{hist}$ model is adapted from the GRU4Rec [20] model for sequential recommendations. Unlike the GRU4Rec model, which takes a sequence of item IDs as its input, GRU$_{hist}$ adopts a GRU to model the user-item interaction history with images.

• **GRU**$_{img+txt}$: The GRU$_{img+txt}$ model (or called Dialog Manager (DM) [16]) leverages a single *GRU* as a state tracker with images of items and natural-language critiques as its inputs for addressing the multi-modal interactive recommendation task.

• **EGE** [51]: Estimator-Generator-Evaluator (EGE) is also a GRU-based model for MMIR. It uses a multi-task learning approach to optimise the model, combining a cross-entropy classification loss for supervised learning and a Q-learning prediction loss for reinforcement learning.

• **GRU-EGE**: To provide strong baseline models for the personalised MMIR task considering both the users' past interests and the current needs, we integrate the existing sequential recommendation model (i.e. GRU$_{hist}$) and the RL-based MMIR model (i.e. EGE) within a pipeline. In particular, the sequential recommendation model estimates the users' past interests from the interaction history and provides the initial recommendations, while the RL-based MMIR model tracks the users' current needs from the real-time interactions and updates the subsequent recommendations.

[4] https://github.com/XiaoxiaoGuo/fashion-retrieval [5] https://github.com/XiaoxiaoGuo/fashion-iq [6] https://github.com/ruotianluo/ImageCaptioning.pytorch

• **GRU**$_{all}$: We extend a single GRU for both estimating the users' past interests and tracking the users' current needs. We optimise the GRU$_{all}$ model with a triplet loss (i.e. GRU$_{all}$-SL) and then extend it with REINFORCE [43] (i.e. GRU$_{all}$-RL) to further improve the performance by maximising the long-term rewards.

The next group of baseline models are based on Transformers in order to compare with PMMIR$_{Transformers}$:

• **Transformer**$_{hist}$: The Transformer$_{hist}$ model is adapted from the SASRec [23] model for sequential recommendations. Unlike the SASRec model, which takes a sequence of item IDs as its input, Transformer$_{hist}$ adopts a Transformer encoder to model the user-item interaction history with images and predict the target item.

• **Transformer**$_{img+txt}$ **& MMT**: The Transformer$_{img+txt}$ model, also called Multi-Modal Interactive Transformer [49, 53], is a state-of-the-art multi-modal interactive recommendation model. It incorporates a Transformer encoder to directly attend to the entire multi-modal real-time interaction sequences, encompassing the users' textual feedback and the system's visual recommendations. We optimise the Transformer$_{img+txt}$ model with a triplet loss and then extend it with REINFORCE (denoted by MMT) to further improve the performance by maximising the long-term rewards.

• **Transformer-MMT**: Similar to GRU-EGE, we also make both well-trained Transformer$_{hist}$ and MMT models into a pipeline for making personalised initial recommendations with Transformer$_{hist}$ and updating the subsequent recommendation during the real-time interactions with Transformer.

• **Transformer**$_{all}$: We also extend a single Transformer encoder for both estimating the users' past interests and tracking the users' current needs. We optimise Transformer$_{all}$ with a triplet loss (Transformer$_{all}$-SL) and then extend it with REINFORCE (Transformer$_{all}$-RL) to further improve the performance by maximising the long-term rewards.

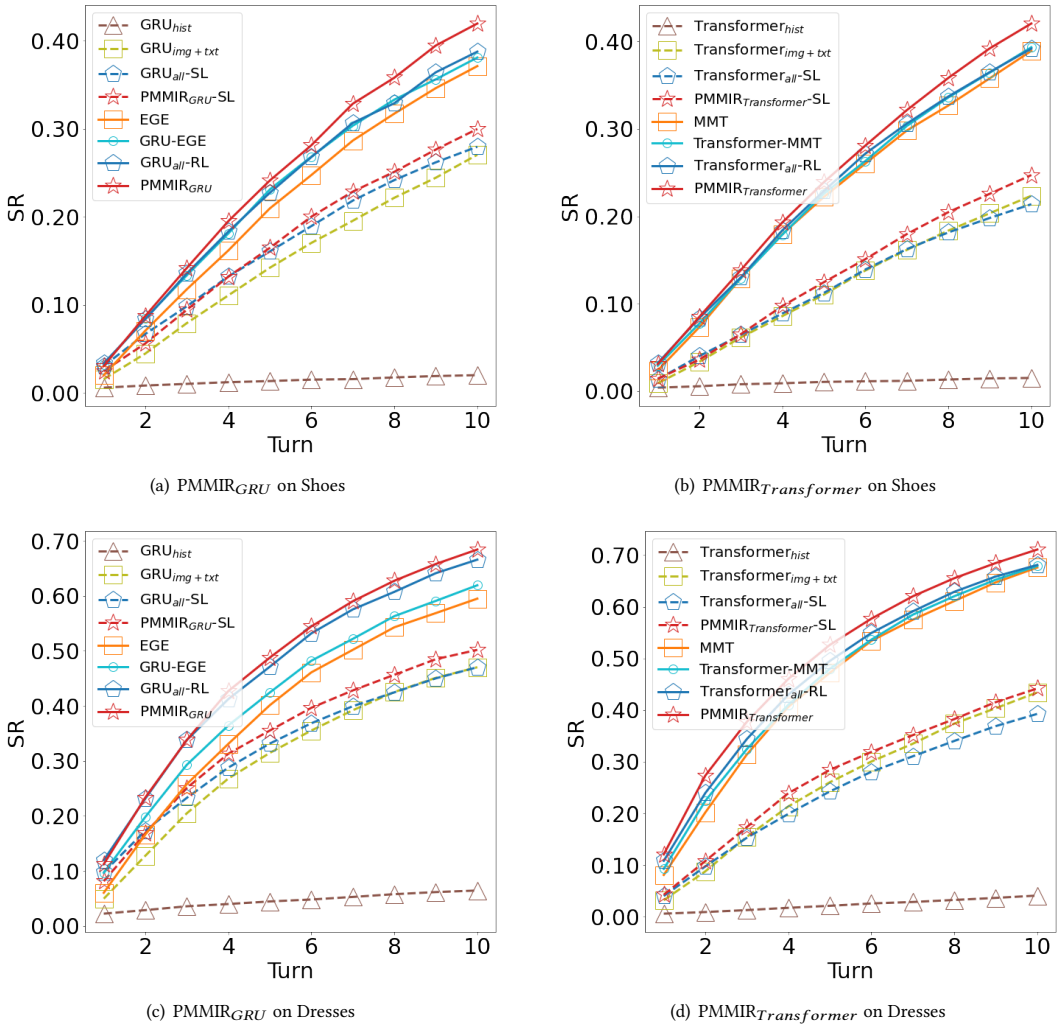
Although there are a few more attention-based/Transformer-based sequential recommendation models (such as BERT4Rec [41] and Transformers4Rec [10]) and multi-modal interactive recommendation models (such as MMRAN [52] with a RNN-enhanced Transformer structure), they can make the PMMIR model overly complex compared to using a simple GRU-based/Transformer-based history encoder. We leave the integration of these more advanced sequential recommendation models for estimating past interests and multi-modal interactive models for tracking the current needs as future work. In addition to the above baseline models for the MMIR task, we also investigate variants of PMMIR for ablation studies. Such variants can also act as solid baselines:

• **PMMIR w/o** $\boldsymbol{h_0^c = h_n^p}$: The "PMMIR w/o $h_0^c = h_n^p$" variant initialises the initial hidden state h_0^c of the state tracker randomly instead of using $h_0^c = h_n^p$.

• **PMMIR w/** $\boldsymbol{Linear^{img/txt}}$: The "PMMIR w/ $Linear^{img/txt}$" variant adds both a $Linear^{img}$ layer in the image encoder and a $Linear^{txt}$ layer in the textual encoder for fine-tuning the CLIP visual and textual representations. The parameters of both the $Linear^{img}$ and $Linear^{txt}$ layers are frozen during the RL training procedure following [16, 51].

• **PMMIR w/ "RN101"**: The "PMMIR w/ RN101" variant replace the ViT-based CLIP checkpoint (i.e. "ViT-B/32") with a ResNet101-based [18] CLIP checkpoint (i.e. "RN101").

For fair comparisons, all of the tested baseline models and variants use CLIP to encode the text and image as the backbone representations (as described in Section 3.2). On the other hand, GRU$_{hist}$ and Transformer$_{hist}$ can be considered as sequential recommendation models, since they only take the users' interaction history into consideration, allowing to compare with models that do not consider text or image representations. Although there are a few more other models with different formulations for the interactive recommendation task (e.g., RCR [62], EAR [27], CRM [42], and SGR [50]), these models are not comparable with our scenario due to requiring additional attributes of items for learning [17, 60–62], requiring a multi-modal knowledge graph for reasoning [50], or their inability to incorporate both the textual and visual modalities during the recommendation process [27, 42].



(a) PMMIR$_{GRU}$ on Shoes

(b) PMMIR$_{Transformer}$ on Shoes

(c) PMMIR$_{GRU}$ on Dresses

(d) PMMIR$_{Transformer}$ on Dresses

Fig. 5. Comparison of the recommendation effectiveness in terms of SR between our proposed PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and the baseline models at various interaction turns with top-3 recommendations on both datasets.

5 EXPERIMENTAL RESULTS

In this section, we present an analysis of the experimental results in relation to the three research questions outlined in Section 4, in order to demonstrate the effectiveness of our proposed PMMIR model. Specifically, we address the overall effectiveness of the PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) for multi-modal interactive recommendations (RQ1, discussed in Section 5.1), its performance on both cold-start and warm-start users (RQ2, detailed in Section 5.2), and the impact of various components and hyperparameters (RQ3, covered in Section 5.3). To further consolidate our findings, we provide a use case based on the logged experimental results in Section 5.4.

Table 2. The recommendation effectiveness of our proposed PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and the baseline models at the 5th and 10th turns on the *Amazon-Shoes* and *Amazon-Dresses* datasets.

	Input Type			Learning	Amazon-Shoes				Amazon-Dresses			
					Turn 5		Turn 10		Turn 5		Turn 10	
Models	hist	img	txt	Type	NDCG@3	SR	NDCG@3	SR	NDCG@3	SR	NDCG@3	SR
GRU												
GRU$_{hist}$	✓	✗	✗	SL	0.0131*	0.0134*	0.0198*	0.0201*	0.0435*	0.0445*	0.0638*	0.0644*
GRU$_{img+txt}$	✗	✓	✓	SL	0.1342*	0.1421*	0.2635*	0.2705*	0.3015*	0.3145*	0.4658*	0.4703*
GRU$_{all}$-SL	✓	✓	✓	SL	0.1520*	0.1606*	0.2740*	0.2796*	0.3204*	0.3315*	0.4653*	0.4703*
PMMIR$_{GRU}$-SL	✓	✓	✓	SL	0.1564*	0.1647*	0.2925*	0.2998*	0.3441*	0.3552*	0.4966*	0.5019*
EGE	✗	✓	✓	RL	0.1970*	0.2095*	0.3644*	0.3712*	0.3825*	0.4012*	0.5885*	0.5950*
GRU-EGE	✓	✓	✓	SL/RL	0.2160*	0.2310*	0.3746*	0.3809*	0.4102*	0.4243*	0.6114*	0.6193*
GRU$_{all}$-RL	✓	✓	✓	RL	0.2160*	0.2272*	0.3821*	0.3876*	0.4573*	0.4712*	0.6587*	0.6659*
PMMIR$_{GRU}$	✓	✓	✓	RL	**0.2299**	**0.2412**	**0.4120**	**0.4196**	**0.4748**	**0.4878**	**0.6766**	**0.6843**
% Improvement	-	-	-	-	6.44	4.42	7.83	8.26	3.83	3.52	2.72	2.76
Transformer												
Transformer$_{hist}$	✓	✗	✗	SL	0.0104*	0.0107*	0.0149*	0.0150*	0.0213*	0.0228*	0.0411*	0.0422*
Transformer$_{img+txt}$	✗	✓	✓	SL	0.1102*	0.1176*	0.2235*	0.2286*	0.2603*	0.2735*	0.4343*	0.4436*
Transformer$_{all}$-SL	✓	✓	✓	SL	0.1122*	0.1179*	0.2138*	0.2192*	0.2425*	0.2553*	0.3927*	0.3994*
PMMIR$_{Transformer}$-SL	✓	✓	✓	SL	0.1245*	0.1311*	0.2472*	0.2536*	0.2842*	0.2937*	0.4419*	0.4498*
MMT	✗	✓	✓	RL	0.2220*	0.2302*	0.3894*	0.3973*	0.4721*	0.4867*	0.6759*	0.6826*
Transformer-MMT	✓	✓	✓	SL/RL	0.2258*	0.2340*	0.3935*	0.4013*	0.4798*	0.4958*	0.6789*	0.6858*
Transformer$_{all}$-RL	✓	✓	✓	RL	0.2289*	0.2412*	0.3919*	0.3989*	0.4950*	0.5086*	0.6809*	0.6876*
PMMIR$_{Transformer}$	✓	✓	✓	RL	**0.2390**	**0.2517**	**0.4207**	**0.4276**	**0.5261**	**0.5394**	**0.7107**	**0.7171**
% Improvement	-	-	-	-	4.41	4.35	6.91	6.55	6.28	6.06	4.38	4.29

5.1 PMMIR vs. Baselines (RQ1)

To address RQ1, we investigate the performance of our proposed PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and the baseline models. Figure 5 depicts the recommendation effectiveness of our proposed PMMIR model variants, along with the corresponding baseline models, for top-3 recommendations in terms of Success Rate (SR) on the *Amazon-Shoes* and *Amazon-Dresses* datasets. Specifically, Figure 5 (a) and (c) represent the results using PMMIR$_{GRU}$, while Figure 5 (b) and (d) correspond to PMMIR$_{Transformer}$. The x-axis indicates the number of interaction turns. Comparing the results presented in Figure 5, we can observe that our proposed PMMIR model variants consistently outperform the baseline models in terms of Success Rate (SR) across different interaction turns (in particular from 4th to 10th turns). This indicates the superior overall performance of our PMMIR models. As the number of interaction turns increases, the differences in effectiveness between our PMMIR models and the baseline models become more pronounced, as observed from the increasing gaps in Success Rate (SR). This suggests that our PMMIR models demonstrate a stronger performance advantage over the baseline models as the interaction process unfolds. We can also observe the same trends on NDCG@3. We omit their reporting in a figure to reduce redundancy. The better overall performance of PMMIR suggests that our PMMIR model can better incorporate the users' preferences from both the interaction history and the real-time interactions compared to the baseline models.

In order to quantify the improvements achieved by our proposed PMMIR model in comparison to the baseline models, we measure their performances in terms of Success Rate (SR) and Normalized Discounted Cumulative Gain at rank 3 (NDCG@3) at the 5th and 10th interaction turns. This enables us to assess the progress and effectiveness of our PMMIR model at different stages of the interaction process. Table 2 presents the obtained recommendation performances of the PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and their corresponding baseline models. These baseline models include the GRU-based models (GRU$_{hist}$, GRU$_{img+txt}$, GRU$_{all}$-SL, EGE, GRU-EGE, GRU$_{all}$-RL) as

Table 3. Personalised multi-modal interactive recommendation effectiveness of our proposed PMMIR model variants (PMMIR$_{GRU}$ and PMMIR$_{Transformer}$) and the baseline models on the cold-start and warm-start users at the 10th turn on the *Amazon-Shoes* and *Amazon-Dresses* datasets. * indicates a significant difference (p<0.05, paired t-test with Holm-Bonferroni correction) wrt. PMMIR for each group.

	Amazon-Shoes						Amazon-Dresses					
	NDCG@3			SR			NDCG@3			SR		
Models	Cold	Warm	Overall	Cold	Warm	Overall	Cold	Warm	Overall	Cold	Warm	Overall
GRU												
EGE	0.3726*	0.3546*	0.3644*	0.3807	0.3600*	0.3712*	0.5876*	0.5892*	0.5885*	0.5935*	0.5963*	0.5950*
GRU-EGE	0.3764	0.3724*	0.3746*	0.3827	0.3787*	0.3809*	0.6120*	0.6109*	0.6114*	0.6210*	0.6179*	0.6193*
GRU$_{all}$-RL	0.3827	0.3814*	0.3821*	0.3886	0.3864*	0.3876*	**0.6575**	0.6597*	0.6587*	0.6639	0.6676*	0.6659*
PMMIR$_{GRU}$	**0.4007**	**0.4253**	**0.4120**	**0.4089**	**0.4322**	**0.4196**	0.6569	**0.6933**	**0.6766**	**0.6665**	**0.6994**	**0.6843**
% Improvement	4.70	11.51	7.83	5.22	11.85	8.26	-0.09	5.09	2.72	0.39	4.76	2.76
Transformer												
MMT	0.3902*	0.3885	0.3894*	0.3980*	0.3964	0.3973*	0.6691*	0.6817	0.6759*	0.6754*	0.6886	0.6826*
Transformer-MMT	0.3973*	0.3889	0.3935*	0.4059*	0.3958	0.4013*	0.6894	0.6701*	0.6789*	0.6959	0.6773*	0.6858*
Transformer$_{all}$-RL	0.3900*	0.3941	0.3919*	0.3970*	0.4011	0.3989*	0.6797*	0.6819	0.6809*	0.6869*	0.6881	0.6876*
PMMIR$_{Transformer}$	**0.4352**	**0.4035**	**0.4207**	**0.4406**	**0.4122**	**0.4276**	**0.7168**	**0.7055**	**0.7107**	**0.7228**	**0.7124**	**0.7171**
% Improvement	9.54	2.39	6.91	8.55	2.77	6.55	3.97	3.46	4.38	3.87	3.46	4.29

well as the Transformer-based models (Transformer$_{hist}$, Transformer$_{img+txt}$, Transformer$_{all}$-SL, MMT, Transformer-MMT, Transformer$_{all}$-RL). The performances are evaluated using the same test datasets from the *Amazon-Shoes* and *Amazon-Dresses* datasets at the 5th and 10th interaction turns. The table provides a comprehensive overview of the recommendation performances, allowing for a direct comparison between the PMMIR model and the various baseline models. In Table 2, the best overall performing results across the four groups of columns are highlighted in bold. * indicates a significant difference, determined by a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), when compared to the PMMIR model within each group. Comparing the results in the table, we observe that our proposed PMMIR$_{GRU}$ model consistently achieves significantly better performances, with improvements on both metrics ranging from 4%-8% and 2%-4% on the Amazon-Shoes and Amazon-Dresses datasets, respectively, compared to the best GRU-based baseline model. Similarly, the PMMIR$_{Transformer}$ model also demonstrates similar improvements, with performance gains ranging from 4%-7% and 4%-6% compared to the best Transformer-based baseline model. These findings highlight the effectiveness of our proposed PMMIR models in outperforming the baseline models across both datasets. Furthermore, it is worth noting that the PMMIR$_{Transformer}$ model, which is based on Transformers, generally outperforms the PMMIR$_{GRU}$ model, which is based on GRUs, in terms of both metrics on both the *Amazon-Shoes* and *Amazon-Dresses* datasets. This observation highlights the superiority of the Transformer-based approach in achieving improved recommendation performances.

In response to RQ1, the results obtained clearly demonstrate that our proposed PMMIR model variants exhibit a significant performance advantage over the state-of-the-art baseline models. Therefore, our proposed PMMIR model with hierarchical state representations in PO-SMDP can effectively incorporate the users' preferences from both the interaction history and the real-time interactions.

5.2 Cold-Start vs. Warm-Start Users (RQ2)

To address RQ2, we investigate the performance of our proposed PMMIR model on cold-start and warm-start users. We classify users with the minimum interactions (3 interactions on Amazon-Shoes and 4 interactions on Amazon-Dresses)

Table 4. Ablation study at the 10th turn. w/o and w/ denote that a component is removed or replaced in PMMIR, respectively. Notation as per Table 3.

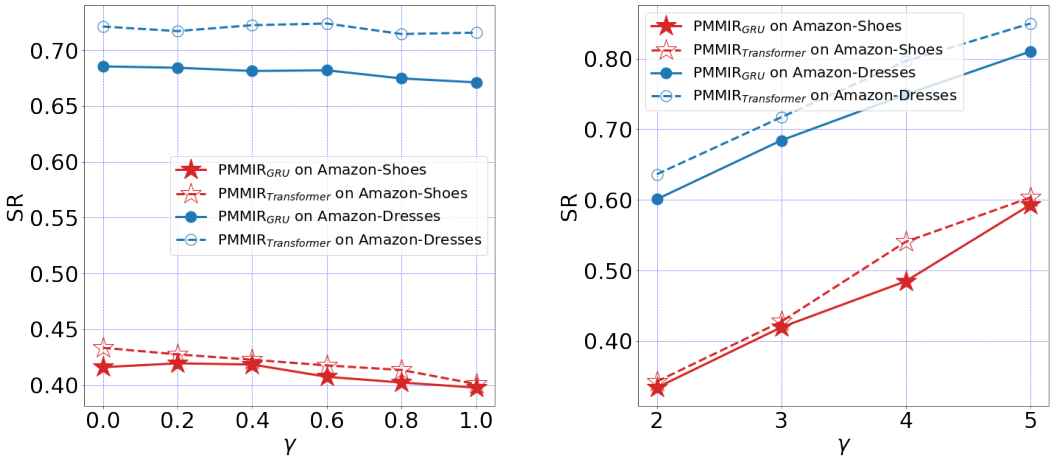
	Amazon-Shoes				Amazon-Dresses			
	GRU		Transformer		GRU		Transformer	
Models	NDCG@3	SR	NDCG@3	SR	NDCG@3	SR	NDCG@3	SR
PMMIR	**0.4120**	**0.4196**	**0.4207**	**0.4276**	**0.6766**	**0.6843**	**0.7107**	**0.7171**
1. w/o $h_0^c = h_n^p$	0.4013	0.4102	0.4074	0.4155	0.6658	0.6714	0.6835*	0.6899*
2. w/ Linear$^{img/txt}$	0.3966	0.4048	0.3510*	0.3575*	0.6462*	0.6530*	0.6252*	0.6322*
3. w/ "RN101"	0.3891	0.3954*	0.3914*	0.4024*	0.6338*	0.6392*	0.6913*	0.6969*

as cold-start users, while those with longer interaction sequences are categorised as warm-start users. This investigation aims to understand how effectively our model adapts to different user scenarios and assess its performance in each case. Table 3 presents the performances of our PMMIR model variants, as well as the RL-based and pipeline-based baseline models, in terms of NDCG@3 and SR. The table is divided into two parts: the top part focuses on the GRU-based models, while the second part pertains to the Transformer-based models. This division facilitates a comprehensive comparison of the performances across different model types. Comparing the results in Table 3, we observe that our proposed PMMIR$_{GRU}$ and PMMIR$_{Transformer}$ models can achieve better performances than the corresponding baseline models in terms of both metrics on both cold-start and warm-start users on the two used datasets, except for the cold-start users with PMMIR$_{GRU}$ in terms of NDCG@3 on Amazon-Dresses. The reported results in Table 3 show that both the cold-start and warm-start users can generally benefit from our proposed PMMIR model variants with hierarchical state representations. In addition, we also observe that the warm-start users can generally benefit more from the GRU-based variant compared to the cold-start users. In particular, PMMIR$_{GRU}$ achieves improvements of 11-12% (warm-start) vs. 4-5% (cold-start) on Amazon-Shoes and 4-5% (warm-start) vs. 0-1% (cold-start) on Amazon-Dresses in terms of both metrics. Conversely, we observe that cold-start users can generally benefit more from the Transformer-based variant compared to warm-start users. In particular, PMMIR$_{Transformer}$ achieves improvements of 8-9% (cold-start) vs. 2-3% (warm-start) on Amazon-Shoes and 3.8-4.0% (cold-start) vs. 3.4-3.5% (warm-start) on Amazon-Dresses in terms of both metrics. We postulate that this difference in performance on cold-start and warm-start users between PMMIR$_{GRU}$ and PMMIR$_{Transformer}$ can be attributed to the features of the interaction history sequences and the different sequence modelling abilities of GRUs and Transformers. The long sequences of purchases (warm-start users) can have a greater timespan and can be noisy due to the users' preferences drifting over time, while short sequences of purchases (cold-start) can have a relatively smaller timespan but can be less informative in relating to the users' preferences. Meanwhile, GRUs (adopted by PMMIR$_{GRU}$) can effectively denoise the sequences with their internal forgetting mechanism with a forget gate, while the Transformer encoders (adopted by PMMIR$_{Transformer}$) have stronger sequence modelling abilities due to the complex neural structures but have been shown to be insufficient to address noisy items within sequences [5].

In response to RQ2, we find that both cold-start and warm-start users can benefit from our proposed PMMIR model. The warm-start users can generally benefit more with PMMIR$_{GRU}$, while the cold-start users can generally benefit more with PMMIR$_{Transformer}$.

5.3 Impact of Components & Hyper-Parameters (RQ3)

To address RQ3, we investigate the impact of the components and the hyper-parameters of our proposed PMMIR model.

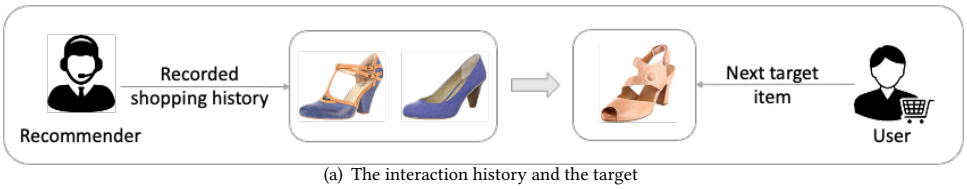


(a) γ for *SR* (b) K for *SR*

Fig. 6. Comparison of the recommendation effectiveness at 10th turn with different γ and K values.

Impact of Components. Table 4 reports the performances of our PMMIR model with different applied ablations in terms of NDCG@3 and SR. The original setting is shown in the top part of the table. The PMMIR ablation variants (i.e. PMMIR w/o $h_0^c = h_n^p$, PMMIR w/ $Linear^{img/txt}$, and PMMIR w/ "RN101") are shown in the second part of the table. All the examined PMMIR ablation variants perform generally worse than the corresponding original PMMIR model. The results of PMMIR w/o $h_0^c = h_n^p$ suggest that our PMMIR model can benefit from the initialisation of the state tracker with the final hidden state of the history encoder. The results of PMMIR w/ $Linear^{img/txt}$ and PMMIR w/ "RN101" indicate that the CLIP model with the "ViT-B/32" checkpoint can provide better visual and textual representations than the "RN101" checkpoint, and further fine-tuning the CLIP embeddings is not necessary for our personalised MMIR task.

Impact of Hyper-Parameters. Figure 6 depicts the effects of the reward discount factor ($\gamma \in [0, 1]$) when training the PMMIR model on both datasets and the number of exposed top-K items ($K \in [2, 5]$) in each ranking list in terms of SR at 10th turn, respectively. In our analysis, we primarily compare the performances of our PMMIR model with different values of discount factors (i.e. $\gamma \in [0, 1]$) at the 10th interaction turn. Specifically, when the discount factor γ is set to 0, it indicates that the model exclusively considers immediate rewards and does not take future rewards into account. On the other hand, when γ is set to 1, the model assigns equal importance to all future rewards and considers them on an equal footing. From Figure 6 (a), we observe that there is a decreasing trend in the performance of PMMIR$_{GRU}$ on both datasets and a decrease in the effectiveness of PMMIR$_{Transformer}$ on Amazon-Shoes when the discount factor γ increases from 0.2 to 1.0. We observe the same trend for PMMIR$_{Transformer}$ on Amazon-Dresses with $\gamma \in [0.6, 1.0]$. This trend shows that both the history encoder and the state tracker in PMMIR are more influenced by the immediate rewards than by future rewards. Additionally, Figure 6 (b) highlights that the PMMIR model exhibits better performance when more items are exposed to users at each interaction turn. This suggests that increasing the number of items presented to users during the interaction process leads to improved recommendation performance for PMMIR.



(a) The interaction history and the target

(b) EGE

(c) PMMIR$_{GRU}$

Fig. 7. Example use cases for the multi-modal interactive recommendation task with EGE (without personalisation) and PMMIR$_{GRU}$ (with personalisation) on *Amazon-Shoes*.

Overall, in response to RQ3, we find that the PMMIR model can generally benefit more in terms of effectiveness from the hierarchical state representations, adequate multi-modal CLIP encoders, using low values for the discount factor γ, and from more exposed top-K items.

5.4 Use Case

In this section, we present use cases of the multi-modal interactive recommendation task with/without personalisation on the *Amazon-Shoes* dataset in Figure 7. In particular, the figure shows a user's interaction history and the next target item, as well as the interaction process for the top-3 recommendations between the simulated users for the EGE and PMMIR$_{GRU}$ models that are both based on GRUs. When the target item is listed in the recommendation list, the user simulator will give a comment to end the interaction, such as "The 3^{rd} shoes are my desired shoes" in Figure 7 (c). Comparing the recommendations made by EGE and PMMIR$_{GRU}$ on the *Amazon-Shoes* dataset, we can observe that our proposed PMMIR$_{GRU}$ model is able to find the target items with fewer interaction turns compared to EGE – this is expected, due to the increased effectiveness of PMMIR$_{GRU}$ shown in Section 5.1. In addition, our PMMIR$_{GRU}$ model is more effective at incorporating the users' preferences from both the users' interaction history and the real-time interactions. For instance, our PMMIR$_{GRU}$ model suggests personalised recommendations with different "high-heeled sandals" at the initial interaction turn, then easily finds the target items with a critique "tan with a higher heel" at the next turn. Meanwhile, the EGE model can only randomly sample items as the initial recommendations, but the "high heel" feature is missing in the initial recommendation, which leads to the EGE model's failure in finding the target item at the next turn. We observed similar trends and results in other use cases involving other baseline models compared to the PMMIR variants on the *Amazon-Shoes* and *Amazon-Dresses* datasets. We omit their reporting in this paper to reduce redundancy.

6 CONCLUSIONS

In this paper, we proposed a novel personalised multi-modal interactive recommendation model (PMMIR) using hierarchical reinforcement learning with the Options framework to more effectively incorporate the users' preferences from both their past and real-time interactions. Specifically, PMMIR decomposes the personalised interactive recommendation process into a sequence of two subtasks with hierarchical state representations: a first subtask where a *history encoder* learns the users' past interests with the *hidden states of history* for providing personalised initial recommendations, and a second subtask where a *state tracker* estimates the current needs with the *real-time estimated states* for updating the subsequent recommendations. The history encoder and the state tracker are jointly optimised with a single optimisation objective by maximising the users' future satisfaction. Following previous work [16, 49, 51], we trained and evaluated our PMMIR model using a user simulator that can generate natural-language critiques about the recommendations as a surrogate for real human users. Our experiments on the *Amazon-Shoes* and *Amazon-Dresses* datasets demonstrate that our proposed PMMIR model variants achieve significantly better performances compared to the best baseline models – for instance, improvements of 4-8% and 2-4% with PMMIR$_{GRU}$ and 4-7% and 4-6% with PMMIR$_{Transformer}$ at the 5th and 10th turns. The reported results show that our proposed PMMIR model benefits from the dual GRUs/Transformers structure and the initialisation of the state tracker with the final hidden state of the history encoder. In addition, the results show that both cold-start and warm start users can benefit from our proposed PMMIR model.

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

[5] Huiyuan Chen, Yusan Lin, Menghai Pan, Lan Wang, Chin-Chia Michael Yeh, Xiaoting Li, Yan Zheng, Fei Wang, and Hao Yang. 2022. Denoising self-attentive sequential recommendation. In *Proc. RecSys*. 92–101.

[6] 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.

[7] 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.

[8] 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:2109.03540* (2021).

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

[10] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proc. RecSys*. 143–153.

[11] 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.

[12] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *JAIR* 13 (2000), 227–303.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*

(2020).

[14] 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.

[15] Claudio Greco, Alessandro Suglia, Pierpaolo Basile, and Giovanni Semeraro. 2017. Converse-et-impera: Exploiting deep learning and hierarchical reinforcement learning for conversational recommender systems. In *Proc. AI*IA*. 372–386.

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

[17] ASM Haque and Hongning Wang. 2022. Rethinking Conversational Recommendations: Is Decision Tree All You Need? *arXiv:2208.14614* (2022).

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

[19] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proc. CIKM*. 843–852.

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

[21] Matthias Hutsebaut-Buysse, Kevin Mets, and Steven Latré. 2022. Hierarchical reinforcement learning: A survey and open research challenges. *Machine Learning and Knowledge Extraction* 4, 1 (2022), 172–221.

[22] Dietmar Jannach, Massimo Quadrana, and Paolo Cremonesi. 2022. Session-based recommender systems. In *Recommender Systems Handbook*. Springer, 301–334.

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

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

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

[26] Sara Latifi, Noemi Mauro, and Dietmar Jannach. 2021. Session-aware recommendation: A surprising quest for the state-of-the-art. *Information Sciences* 573 (2021), 291–315.

[27] 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.

[28] 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.

[29] Yuanguo Lin, Fan Lin, Wenhua Zeng, Jianbing Xiahou, Li Li, Pengcheng Wu, Yong Liu, and Chunyan Miao. 2022. Hierarchical reinforcement learning with dynamic recurrent mechanism for course recommendation. *Knowledge-Based Systems* 244 (2022), 108546.

[30] Yuanguo Lin, Yong Liu, Fan Lin, Lixin Zou, Pengcheng Wu, Wenhua Zeng, Huanhuan Chen, and Chunyan Miao. 2021. A survey on reinforcement learning for recommender systems. *arXiv:2109.10665* (2021).

[31] Qidong Liu, Jiaxi Hu, Yutian Xiao, Jingtong Gao, and Xiangyu Zhao. 2023. Multimodal Recommender Systems: A Survey. *arXiv:2302.03883* (2023).

[32] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *arXiv:1803.04376* (2018).

[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv:1312.5602* (2013).

[34] Ronald Parr and Stuart Russell. 1997. Reinforcement learning with hierarchies of machines. In *Proc. NeurIPS*.

[35] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–35.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP*. 1532–1543.

[37] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proc. RecSys*. 130–137.

[38] 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.

[39] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[40] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv:1205.2618* (2012).

[41] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proc. CIKM*. 1441–1450.

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

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

[44] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*, Vol. 30.

[46] Jianling Wang, Kaize Ding, and James Caverlee. 2021. Sequential recommendation for cold-start users with meta transitional learning. In *Proc. SIGIR*. 1783–1787.

[47] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–38.

[48] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[49] 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.

[50] 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).

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

[52] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation. In *Proc. RecSys*. 124–133.

[53] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback. In *CUI*. 1–10.

[54] Ruobing Xie, Shaoliang Zhang, Rui Wang, Feng Xia, and Leyu Lin. 2021. Hierarchical reinforcement learning for integrated recommendation. In *Proc. AAAI*, Vol. 35. 4521–4528.

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

[56] 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.

[57] 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.

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

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

[60] 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.

[61] Yifei Yuan and Wai Lam. 2021. Conversational Fashion Image Retrieval via Multiturn Natural Language Feedback. In *Proc. SIGIR*. 839–848.

[62] 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.

[63] 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.

[64] Yujia Zheng, Siyi Liu, Zekun Li, and Shu Wu. 2021. Cold-start sequential recommendation via meta learner. In *Proc. AAAI*. 4706–4713.

[65] Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A Comprehensive Survey on Multimodal Recommender Systems: Taxonomy, Evaluation, and Future Directions. *arXiv:2302.04473* (2023).