Effective Multi-Modal Conversational Recommendation

Yaxiong Wu

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

January 2024

Abstract

Conversational recommender systems have recently received much attention for addressing the information asymmetry problem in information seeking, by eliciting the dynamic preferences of users and taking actions based on their current needs through multi-turn & closed-loop interactions. Despite recent advances in *uni-modal* conversational recommender systems that use only natural-language interfaces for recommendations, leveraging both visual and textual information effectively for *multi-modal* conversational recommender systems has not yet been fully researched. In particular, *multi-modal* conversational recommender systems are expected to leverage the multi-modal information (such as the natural-language feedback of users and textual/visual representations of recommendation items) during the communications between users and recommender systems.

In this thesis, we aim to effectively track and estimate the users' dynamic preferences from the multi-modal conversational recommendations (in particular with vision-and-language-based interactions), so as to develop realistic and effective multi-modal conversational recommender systems. In particular, we are motivated to answer the following questions: (1) how to better understand the users' natural-language feedback and the corresponding recommendations with the partial observability of the users' preferences over time; (2) how to better track the users' preferences over the sequences of the systems' visual recommendations and the users' natural-language feedback; (3) how to decouple the recommendation policy (i.e. model) optimisation and the multi-modal composition representation learning; (4) how to effectively incorporate the users' long-term and short-term interests for both cold-start and warm-start users; (5) how to ensure the realism of simulated conversations, such as positive/negative natural-language feedback. To address these five challenges, we propose to leverage recent advanced techniques (including multi-modal learning, deep learning, and reinforcement learning) for re-framing and developing more effective multi-modal conversational recommender systems. In particular, we introduce the framework of the multi-modal conversational recommendation task with cold-start or warm-start users, as well as how to measure the success of the tasks. Note that we also refer to multi-modal conversational recommendation as dialog-based interactive recommendation or multi-modal interactive recommendation throughout this thesis.

The first challenge refers to the partial observability in natural-language feedback. For example, the users' feedback, which takes the form of natural-language critiques about the displayed recommendation at each iteration, can only allow the recommender system to obtain a partial portrayal of the users' preferences. To alleviate such a partial observation issue, we propose a novel dialog-based recommendation model, the Estimator-Generator-Evaluator (EGE) model, which uses Q-learning for a partially observable Markov decision process (POMDP), to effectively incorporate the users' preferences over time. Specifically, we leverage an Estimator to track and estimate users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations, and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback.

The second challenge refers to multi-modal sequence dependency issue in multi-modal dialog state tracking. For instance, multi-modal dialog sequences (i.e. turns consisting of the system's visual recommendations and the user's natural-language feedback) make it challenging to correctly incorporate the users' preferences across multiple turns. Indeed, the existing formulations of interactive recommender systems suffer from their inability to capture the multi-modal sequential dependencies of textual feedback and visual recommendations because of their use of recurrent neural network-based (i.e., RNN-based) or transformer-based models. To alleviate the multi-modal sequence dependency issue, we propose a novel multi-modal recurrent attention network (MMRAN) model to effectively incorporate the users' preferences over the long visual dialog sequences of the users' natural-language feedback and the system's visual recommendations.

The third challenge refers to the coupling issue of policy (i.e. recommendation model) optimisation and representation learning. For example, it is typically challenging and unstable to optimise a recommendation agent to improve the recommendation quality associated with implicit learning of multi-modal representations in an end-to-end fashion in deep reinforcement learning (DRL). 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 (using goal-oriented reinforcement learning (GORL)) to explicitly learn the multi-modal representations using a multi-modal composition network when optimising the recommendation agent.

The fourth challenge refers to the personalisation for cold-start and warm-start users. For instance, it can be challenging to make satisfactory personalised recommendations across multiple 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. To perform the personalisation for cold-start and warm-start users, we propose a novel personalised multi-modal interactive recommendation model (PMMIR) using hierarchical reinforcement learning (HRL) to more effectively incorporate the users' preferences from both their past and real-time interactions.

The final challenge refers to the realism of simulated conversations. In a real-world shopping scenario, users can express their natural-language feedback when communicating with a shopping assistant by stating their satisfactions positively with “I like” or negatively with “I dislike” according to the quality of the recommended fashion products. A multi-modal conversational recommender system (using text and images in particular) aims to replicate this process by eliciting the dynamic preferences of users from their natural-language feedback and updating the visual recommendations so as to satisfy the users’ current needs through multi-turn interactions. However, the impact of positive and negative natural-language feedback on the effectiveness of multi-modal conversational recommendation has not yet been fully explored. To further explore the multi-modal conversational recommendation with positive and negative natural-language feedback, we investigate the effectiveness of the recent multi-modal conversational recommendation models for effectively incorporating the users’ preferences over time from both positively and negatively natural-language oriented feedback corresponding to the visual recommendations.

Overall, we contribute an effective multi-modal conversational recommendation framework that make accurate recommendations by leveraging visual and textual information. This framework includes models for tracking users’ preferences with partial observations, mitigating the multi-modal sequence dependency issue, decoupling the composition representation learning from policy optimisation, incorporating both the users’ long-term preferences and short-term needs for personalisation, and ensuring the realism of simulated conversations. These contributions make progress in the development of multi-modal conversational recommendation techniques and could inspire future directions of research in recommendation systems.

Acknowledgements

I would like to express my heartfelt gratitude to all those who have supported me throughout this journey, without whom this thesis would not have been possible. Their guidance, encouragement, and unwavering belief in my abilities have been instrumental in shaping this work.

I extend my sincere gratitude to my supervisors, Craig Macdonald and Iadh Ounis, for their expert guidance and valuable insights that shaped the direction of my research.

I am also grateful to my friends and colleagues at the Terrier team and the School of Computing Science, including Richard McCreadie, Bjørn Sand Jensen, Graham McDonald, Sean MacAvaney, Zaiqiao Meng, Debasis Ganguly, Roderick Murray-Smith, Ting Su, Xi Wang, Xiao Wang, Yanni Ji, Fuxiang Tao, Javier Sanz-Cruzado Puig, Erlend Frayling, Siwei Liu, Hitarth Narvala, Sarawoot Kongyoung, Zijun Long, Thomas Jänich, JingMin Huang, Maria Vlachou, Sasha Petrov, Zixuan Yi, Zeyuan Meng, Zeyan Liang, Lubingzhi Guo, Jinyuan Fang and many others. They have offered me much help and shared much of their knowledge. It has been a great pleasure to work with them.

To my wife, Yixuan Li, your unwavering belief in me provided the motivation I needed. Your love and understanding sustained me throughout. I would also like to thank my parents, my parents-in-law, my siblings, and other family members, whose support was my pillar of strength.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Recommender systems are widely used as an essential *information seeking* tool in a variety of online services, including e-commerce (e.g., Amazon, Alibaba, Shein), social networking (e.g., Facebook, Twitter, WeChat), and content sharing (e.g., TikTok, Instagram, YouTube). Typically, the main task of such recommender systems is to elicit the users' preferences from their historical interactions (e.g., clicks, ratings, purchases, and reviews) to make recommendations with the users' likely preferred items, thereby helping users find their desired information in situations of *information overload* (Ricci, Rokach, & Shapira, 2015). The existing recommender systems (e.g., Matrix Factorisation (Koren, Bell, & Volinsky, 2009) and Neural Collaborative Filtering (X. He et al., 2017)) typically adopt a *static* mode of user interaction modelling that trains the recommendation models *offline* on the users' historical behaviour data, while serving users *online* following a fixed strategy. These static recommender systems cannot lead to a satisfactory performance in that they limit the way in which user intention can be expressed and offer no opportunities to communicate with users about their preferences. Despite their wide usage, they inevitably suffer from a fundamental *information asymmetry* problem (Gao, Lei, He, de Rijke, & Chua, 2021) between users and systems: a recommender system will never know precisely *what a user likes* (especially when the user's preference drifts frequently) and *why a user likes an item* (especially when there are many different factors affecting a user's decision in real life, such as curiosity, mood and season).

Conversational recommender systems (CRSs) have recently received much attention for addressing the *information asymmetry* problem (Gao et al., 2021) in information seeking, owing to their flexible recommendation strategies and their natural multi-turn decision-making processes. In particular, Gao et al. (2021) defines a CRS as: "*A recommender system that can elicit the*

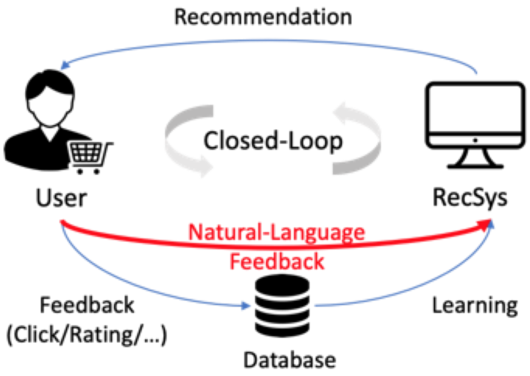


Figure 1.1: A diagram of the closed-loop interactions between users and recommender systems.

dynamic preferences of users and take actions based on their current needs through real-time multi-turn interactions". To this end, the conversational recommender systems can be generally considered to form a *closed loop system* (from *control theory* (Simrock, 2011)) in which the inputs (i.e. users' feedback) of the recommender systems are fully or partially determined by the outputs (i.e. recommended items). The conversational recommender systems can benefit from such closed-loop interactions by tracking and capturing users' current preferences across multiple interactions, thus mitigating the information asymmetry. Figure 1.1 presents the closed-loop interactions between users and recommender systems.

There are three typical forms of CRSs, namely *interactive recommender systems* (Zou et al., 2020), *critiquing-based recommender systems* (Antognini & Faltings, 2021) and *question-based recommender systems* (W. Lei et al., 2020; Y. Sun & Zhang, 2018). Specifically, the interactive recommender systems leverage an item-level feedback signal indicating whether and how much the user likes the corresponding recommendation (such as ratings or like/dislike), the critiquing-based recommender systems leverage the users' feedback on specific attributes of the recommended items so as to narrow down candidate items quickly, and the question-based recommender systems leverage a conversation strategy to determine when to ask and recommend. However, the existing formulations of CRSs have demonstrated their drawbacks. For instance, the interactive recommendations suffer from low efficiency by leveraging the item-level feedback when there are too many items (Gao et al., 2021), the critiquing-based recommendations are constrained by the limited attribute-based options for the attribute-level feedback (Gao et al., 2021), and the question-based recommendations suffer from inefficient interactions by requesting users to answer multiple questions (Iovine, Narducci, & Semeraro, 2020; Jannach, Manzoor, Cai, & Chen, 2021). To this end, we argue that developing the form of conversational recommender systems is still an open problem.

On the other hand, the recent research on conversational recommender systems primarily focuses on *uni-modal* interactions and information items, such as the question-based recommender systems using pure natural-language interfaces for recommendations (Gao et al., 2021).

However, it is widely known that human conversations are *multi-modal*, involving different actions from humans (e.g., word, speech, gesture, facial expression) and different representations of items from machines (e.g., audio, text, image, video) (Deldjoo, Trippas, & Zamani, 2021). In particular, vision-and-language-based interactions between users and recommender systems can be effective for the benefits of both visual information from the recommendations' images and textual information from the users' natural-language feedback (Guo et al., 2018; Uppal et al., 2021; H. Wu et al., 2021; Yuan & Lam, 2021). For instance, the users' natural-language critiques about the visual recommendations can allow the recommender systems to correctly track the users' preferences over time and adapt the systems' instant recommendations, thereby satisfying the users' information needs effectively (Guo et al., 2018). However, leveraging both visual and textual information effectively for multi-modal conversational recommender systems has not been fully researched. In this thesis, we aim to address the gap in effectively tracking and estimating the users' dynamic preferences from the multi-modal conversational recommendations, so as to develop realistic and effective multi-modal conversational recommender systems.

In the remainder of this chapter, we start with a discussion of the motivations of this thesis in Section 1.2. After that, we introduce the thesis statement in Section 1.3. In Section 1.4, we describe the contributions of this thesis, followed by acknowledging the origins of materials in Section 1.5.

1.2 Motivations

We aim to develop multi-modal conversational recommender systems by leveraging the multi-modal interactions (including both visual and textual information) between users and recommender systems. This is initially motivated by the recent well-cited dialog-based interactive image retrieval framework proposed by Guo et al. (2018) that allows for more natural and effective interactions by enabling users to provide feedback via natural language during the image search process. Furthermore, the interactive recommender systems (Yu, Shen, & Jin, 2019, 2020; Yu, Shen, Zhang, Zeng, & Jin, 2019; R. Zhang, Yu, Shen, Jin, & Chen, 2019) adapted the dialog-based interactive image retrieval framework for an interactive recommendation task by considering the images as the visual representations of fashion products. The recommender systems can better elicit the users' preferences by incorporating the users' natural-language feedback and visual recommendations over time. However, these primary works relating to the multi-modal conversational recommendation task have demonstrated their limitations with the current formulations in correctly understanding the users' natural-language feedback with the partial observations of the users' preferences over time, the multi-modal sequence dependency, the coupling issue of multi-modal representation learning and policy optimisation (i.e. optimising the recommendation models), the personalisation for both cold-start and warm-start users, and the realism of simulated conversations across the multi-turn recommendation process (such as positive and negative natural-language feedback). In addition, with the recent rapid development of advanced technologies in multi-modal learning (such as TIRG (Vo et al., 2019)), large foundation models (such as ViT (Dosovitskiy et al., 2020) for vision, BERT (Devlin, Chang, Lee, & Toutanova, 2019a) for language, and CLIP (Radford et al., 2021) for vision and language), dialog systems (such as ChatGPT (Y. Liu et al., 2023)), sequential recommender systems (such as SASRec (Kang & McAuley, 2018) and BERT4Rec (Petrov & Macdonald, 2022; F. Sun et al., 2019)) and reinforcement learning approaches (such as self-supervised reinforcement learning (SRL) (Levine, 2022; Xin, Karatzoglou, Arapakis, & Jose, 2020), goal-oriented reinforcement learning (GORL) (Colas, Karch, Sigaud, & Oudeyer, 2022; M. Liu, Zhu, & Zhang, 2022), and hierarchical reinforcement learning (HRL) (Hutsebaut-Buysse, Mets, & Latré, 2022; Pateria, Subagdja, Tan, & Quek, 2021)) provide more possibilities of reformulating and improving the multi-modal conversational recommendation framework.

Overall, motivated by the limitations of the existing multi-modal conversational recommendation framework and the development of advanced technologies in multiple domains, we propose to leverage those advanced techniques for re-framing and developing more effective multi-modal conversational recommender systems. Hence, in this thesis, we are motivated to answer the following questions: (1) How to better understand the users' natural-language feedback and the corresponding recommendations with the partial observations of the users' preferences over time; (2) How to better track the users' preferences over the sequences of the systems' visual recommendations and the users' natural-language feedback; (3) How to decouple the recommendation policy (i.e. model) optimisation and the multi-modal composition representation learning; (4) How to effectively incorporate the users' long-term and short-term interests for both cold-start and warm-start users; (5) how to ensure the realism of simulated conversations, such as positive/negative natural-language feedback.

1.3 Thesis Statement

The statement of this thesis is that the tasks of modelling multi-modal conversational recommendations can be effectively achieved by tracking users' preferences with partial observations, mitigating the multi-modal sequence dependency issue, decoupling the composition representation learning from policy optimisation, incorporating both the users' long-term preferences and short-term needs for personalisation, and ensuring the realism of simulated conversations.

Research Topic 1: By modelling the multi-modal conversational recommendation process with (self-)supervised Q-learning in a partially observable environment, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time using the partial observations.

Research Topic 2: By mitigating the multi-modal sequence dependency issue in the multi-modal conversational recommendation process, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time with an RNN-enhanced Transformer structure for state tracking.

Research Topic 3: By decoupling the policy optimisation and the multi-modal composition representation learning with goal-oriented reinforcement learning, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time with a composition network and a multi-task learning approach.

Research Topic 4: By modelling the multi-modal conversational recommendation process with both the users' interaction history and the users' instant natural-language feedback, the multi-modal conversational recommender system can effectively incorporate both the users' long-term preferences and short-term needs into the personalised recommendations.

Research Topic 5: To make the multi-modal conversational recommendation task more realistic, we ensure the realism of simulated conversations by considering positive/negative natural-language feedback.

1.4 Contributions & Thesis Outline

The contributions of this thesis have five groups corresponding to the five aspects of the research topics in Section 1.3. Firstly, we propose a multi-modal conversational recommendation model, the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP, to effectively incorporate the users' preferences over time in the partial observable environment. Secondly, we propose a multi-modal recurrent attention network (MMRAN) model to effectively incorporate the users' preferences over the long visual dialog sequences of the users' natural-language feedback and the system's visual recommendations for addressing the multi-modal sequence dependency issue. Thirdly, we propose a goal-oriented multi-modal interactive recommendation model (GOMMIR) using goal-oriented reinforcement learning for decoupling the policy optimisation and representations learning by leveraging both verbal and non-verbal relevance feedback. Then, we propose a 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. Finally, we ensure the realism of simulated conversations by considering positive/negative natural-language feedback.

This thesis is organised as follows:

- Chapter 1 introduces the multi-modal conversational recommendation task, the motivations of this thesis, our research topics, the contributions of this thesis, and the supporting materials.

- Chapter 2 illustrates the background of developing multi-modal conversational recommender systems (MMCRSs). We first introduce the general overview of recommender systems, including a taxonomy of recommender systems, deep learning-based recommender systems, and evaluation methods. Then, we discuss closed-loop systems, including the definition of closed-loop systems, closed loops in recommender systems, and the properties of closed loops. We also introduce different types of conversational recommendation, including system-initiative vs user-initiative recommender systems, uni-modal vs multi-modal conversational recommender systems, and retrieval-based and generation-based conversational recommender systems. Finally, we describe the preliminaries of deep reinforcement learning algorithms.

- Chapter 3 describes the multi-modal conversational recommendation framework. We first discuss the related work in the literature, as well as five challenges, for addressing the multi-modal conversational recommendation task. Then, we illustrate our multi-modal conversational recommendation framework for leveraging the users' natural-language feedback. We describe the user simulators for training and testing the multi-modal conversational recommender systems. Finally, we discuss the opportunities within the framework from both recommender system and user sides.

- Chapter 4 introduces our proposed Estimator-Generator-Evaluator (EGE) model with Q-learning for POMDP that addresses the partial observations issue in the environment. We train our EGE model by using a user simulator, which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. We evaluate the effectiveness of our proposed EGE model by comparing to the existing state-of-the-art baseline models on two recommendation datasets – addressing images of fashion products (namely the Shoes and Fashion IQ Dresses datasets).

- Chapter 5 introduces our proposed multi-modal recurrent attention network (MMRAN) model for addressing the multi-modal sequence dependency issue so as to effectively incorporate the users' preferences over the long visual dialog sequences of the users' natural-language feedback and the system's visual recommendations. We conduct extensive experiments on the Fashion IQ Dresses, Shirts, and Tops & Tees datasets to assess the effectiveness of our proposed model by using a vision-language transformer-based user simulator as a surrogate for real human users.

- Chapter 6 introduces our proposed goal-oriented multi-modal interactive recommendation model (GOMMIR) that uses both verbal and non-verbal relevance feedback for decoupling the policy optimisation and representation learning issue, thereby effectively incorporating the users' preferences over time. We performed extensive experiments on four well-known fashion datasets (Shoes, Fashion IQ Dresses, Shirts, and Tops & Tees) to evaluate the effectivenss of our proposed GOMMIR model in comparison to the existing state-of-the-art baseline models.

- Chapter 7 introduces our proposed 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. We evaluate the effectiveness of our proposed PMMIR model on two derived fashion datasets (i.e. Amazon-Shoes and Amazon-Dresses) from two well-known public datasets (Amazon Review Data 2014 and 2018) in comparison to the existing state-of-the-art baseline models.
- Chapter 8 investigates the impact of positive and negative natural-language feedback in multi-modal conversational recommendation. We propose an approach to generate both positive and negative natural-language critiques about the recommendations within an existing user simulator. We evaluate the two existing conversational recommendation models by using the proposed user simulator on the Shoes dataset.
- Chapter 9 summarises our contributions of this thesis, describes the conclusions of each chapter, and discuss some possible future work.

1.5 Supporting Papers

The thesis builds on the following papers:

- **Yaxiong Wu**, Craig Macdonald, and Iadh Ounis. "Partially Observable Reinforcement Learning for Dialog-based Interactive Recommendation." Proceedings of the 15th ACM Conference on Recommender Systems. 2021. (**RecSys 2021**, Chapter 4)
- **Yaxiong Wu**, Craig Macdonald, and Iadh Ounis. "Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation." Proceedings of the 16th ACM Conference on Recommender Systems. 2022. (**RecSys 2022**, Chapter 5)
- **Yaxiong Wu**, Craig Macdonald, and Iadh Ounis. "Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback." Proceedings of the 17th ACM Conference on Recommender Systems. 2023. (**RecSys 2023**, Chapter 6)
- **Yaxiong Wu**, Craig Macdonald, and Iadh Ounis. "Personalised Multi-Modal Interactive Recommendation with Hierarchical State Representations." (**TORS, Accepted**, Chapter 7)
- **Yaxiong Wu**, Craig Macdonald, and Iadh Ounis. "Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback." Proceedings of the 4th Conference on Conversational User Interfaces. 2022. (**CUI 2022**, Chapter 8)

Chapter 2

Background

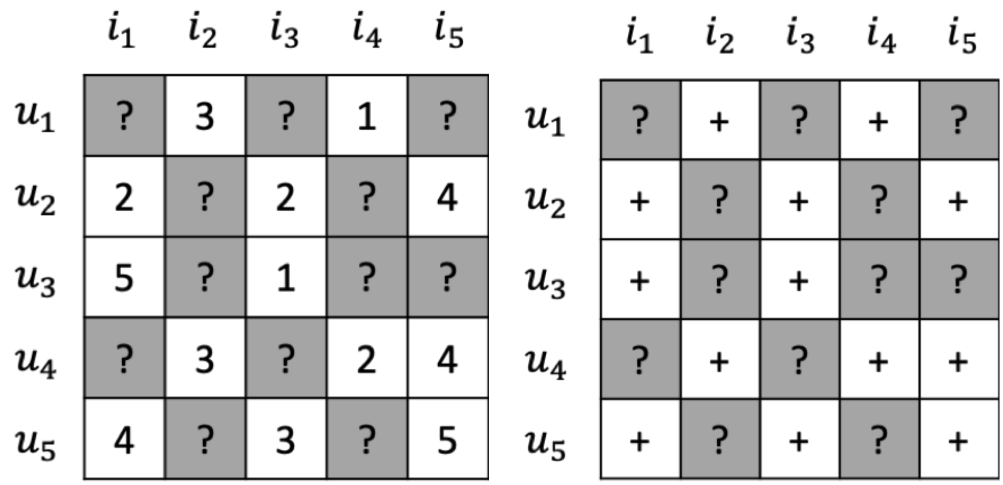
In this chapter, we provide background and preliminaries on multi-modal conversational recommendations. In Section 2.1, we first introduce the general overview of recommender systems, including a taxonomy of recommender systems, deep learning-based recommender systems, and evaluation methods. In Section 2.2, we discuss about closed-loop systems, including the definition of closed-loop systems, closed loops in recommender systems, and the properties of closed loops. In Section 2.3, we introduce different types of conversational recommendation, including system-initiative vs user-initiative conversational recommender systems, uni-modal vs multi-modal conversational recommender systems, and retrieval-based vs generation-based conversational recommender systems. Finally, in Section 2.4, we describe the preliminaries of deep reinforcement learning algorithms.

2.1 Overview of Recommender Systems

The formulations of recommender systems have been greatly enriched over the last decades, from the *static* preference estimation with historical interaction data to the *dynamic* preference elicitation with real-time multi-turn interactions (S. Zhang, Yao, Sun, & Tay, 2019). In this section, we describe different formulations of recommender systems, including a taxonomy of recommender systems and the formulations of recommender systems with various deep learning (DL) approaches. We also illustrate the evaluation methods.

2.1.1 Taxonomy of Recommender Systems

Recommender systems have been successfully applied in various online services (such as e-commerce, social media, and entertainment) by addressing the *information overload* issue and helping users find their preferred items (Ricci et al., 2015; S. Zhang et al., 2019). Recommender



	i_1	i_2	i_3	i_4	i_5
u_1	?	3	?	1	?
u_2	2	?	2	?	4
u_3	5	?	1	?	?
u_4	?	3	?	2	4
u_5	4	?	3	?	5

(a) Explicit Feedback

	i_1	i_2	i_3	i_4	i_5
u_1	?	+	?	+	?
u_2	+	?	+	?	+
u_3	+	?	+	?	?
u_4	?	+	?	+	+
u_5	+	?	+	?	+

(b) Implicit Feedback

Figure 2.1: Examples of (a) explicit feedback (such as ratings with a scale of 1-5) and (b) implicit feedback (such as clicks). *u* and *i* denote users and items, respectively. “+” and “?” denote positive implicit feedback (such as clicks) and unobserved feedback, respectively.

systems usually estimate the users’ preferences by modelling the users’ past behaviours offline, using either the users’ explicit feedback (e.g., purchases, ratings, reviews and critiques) or implicit feedback (e.g., clicks and skips). Figure 2.1 illustrates a representation of explicit feedback (such as ratings) and implicit feedback (such as clicks). In particular, explicit feedback provides direct and clear signals about the users’ preferences but can be sparse. On the other hand, implicit feedback is usually more abundant but noisy so it requires careful modelling to infer the users’ preferences.

Recommender systems have been defined as “a subclass of information filtering system that seeks to predict the ‘rating’ or ‘preference’ a user would give to an item” (Ricci et al., 2015). To this end, there are generally two types of recommendation tasks: (1) rating prediction and (2) item ranking. In particular, the rating prediction task aims to predict the missing ratings on the users’ unobserved items (i.e. “?” in Figure 2.1 (a)). Meanwhile, the item ranking task is expected to provide a ranking list of unobserved items (i.e. “?” in Figure 2.1 (b)) and correctly rank the users’ preferred items on top according to the estimated users’ preferences. In addition, compared to the rating prediction task, recommender systems for the ranking task can leverage richer information relating to the users’ preferences from different user-item interactions, using both implicit feedback (such as clicks and skips) and explicit feedback (such as purchases, reviews and critiques), thereby satisfying the users’ information needs with their more preferred items. Figure 2.2 illustrates an example ranking list of items for recommendation. This thesis focuses on developing recommendation methods for ranking items since the item ranking task is more realistic than the rating prediction task during the multi-turn information-seeking processes.

Recommendation algorithms can be grouped into four categories: collaborative filtering

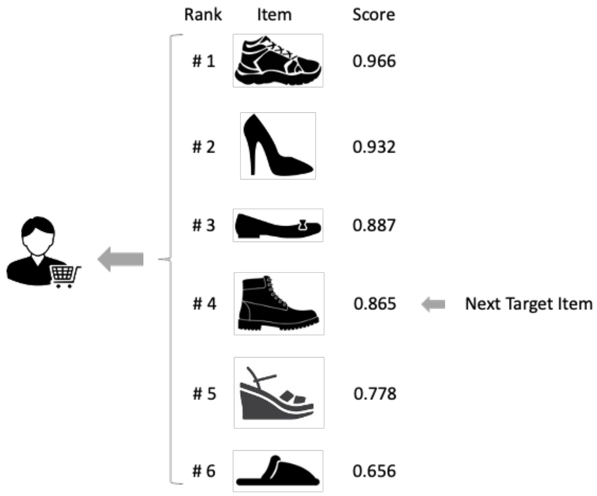


Figure 2.2: An example ranking list of items for recommendation.

(CF), content-based recommender systems, knowledge-based recommender systems and sequential recommender systems.

Collaborative Filtering (CF) Collaborative filtering is based on the idea that users with similar preferences in the past will have similar preferences in the future (Ricci et al., 2015). Collaborative filtering approaches (Linden, Smith, & York, 2003), such as Matrix Factorisation (Koren et al., 2009), make recommendations by leveraging the collective wisdom of users (C.-M. Chen, Wang, Tsai, & Yang, 2019) and learning from either explicit or implicit feedback with the user-item historical interactions. However, collaborative filtering can suffer from the "cold start" problem when a user has a limited history of interactions or a new item is added to the system.

Content-based Recommender Systems Content-based recommender systems (Mooney & Roy, 2000) learn to recommend items similar to the user liked items in the past. A content-based recommender system typically involves encoding items' representations with characteristics or attributes, creating profiles of the users' preferences based on their interactions, and computing the similarities between the users' profiles and the representations of items. By leveraging the properties (such as attributes) of items to make recommendations, the content-based recommender systems can effectively mitigate the "cold start" problem. However, if the items' attributes are not well-defined, incomplete, or not representative of the users' preferences, the content-based recommendations may not be satisfactory to meet the users' information needs.

Knowledge-based Recommender Systems Knowledge-based recommender systems (Akerkar & Sajja, 2009) recommend items on the basis of user-specified requirements rather than the historical interactions of the users. The knowledge-based methods rely on explicit domain knowledge or rules to generate personalized recommendations, which are well suited to complex domains (such as e-commerce) where users usually want to express their preferences explicitly (e.g., “I prefer red high heels"). In particular, knowledge-based recommender systems can handle situations where there is limited or no user interaction data, making them useful in "cold start" scenarios. There are two primary types of knowledge-based recommender systems: constraint-based recommender systems and case-based recommender systems. The constraint-based systems (Felfernig, Friedrich, Jannach, & Zanker, 2015) typically allow users to specify requirements or constraints (e.g., lower or upper limits) on the item attributes, while case-based recommender systems (Smyth, 2007) allow users to specify critiques (L. Chen & Pu, 2012) (e.g., “I would like to have a dress like this but with a lower price”) regarding the current recommended item. The interaction between a user and a recommender system may take the form of *conversational recommendations* (Gao et al., 2021) by eliciting the users’ requirements and preferences within the scope of a feedback loop (L. Chen et al., 2013), *search-based recommendations* (Burke, 2000) by leveraging the users’ queries or answers to questions to find the relevant items, or *navigation-based recommendations* (L. Chen & Pu, 2012) by using change requests (such as critiques) on the attributes of items.

Sequential Recommender Systems Sequential recommender systems are a class of recommender systems that make personalised recommendations by modelling the sequential dependencies over the user-item interactions, such as a sequence of purchased/clicked items (S. Wang et al., 2019). Different from the previously mentioned recommender systems (such as collaborative filtering, content-based recommender systems, and knowledge-based recommender systems) that mainly focus on user-item interactions (e.g., ratings or purchase history), sequential recommender systems take the temporal aspect of the users’ behaviours into account, such as the order in which items are interacted with over time. Such a sequence modelling of the sequential recommender systems is particularly useful in scenarios where user preferences and interests evolve over time, and the order of interactions matters. The task of sequential recommendation is usually formulated as a *next item prediction task*, where the recommender systems predict the next user-item interaction by taking the sequence of the user’s past interactions as their inputs.

Thus far, we have demonstrated the taxonomy of various recommender systems, such as collaborative filtering, content-based recommender systems, knowledge-based recommender systems, and sequential recommender systems. Recently, deep learning has shown to be effective in modelling complex patterns among data inputs, extracting high-level representations, and handling large-scale data. Extensive recent research has been focused on recommendation systems based on deep learning techniques. Therefore, in the next section, we describe various typical deep learning techniques that were used to develop effective recommendation models.

2.1.2 DL-based Recommender Systems

Deep learning (DL) is increasingly applied in the recommendation domain (Batmaz, Yurekli, Bilge, & Kaleli, 2019; S. Zhang et al., 2019) due to its expressive representation learning abilities

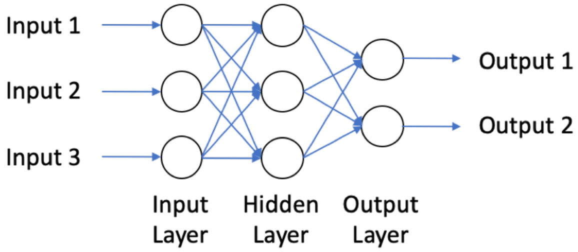


Figure 2.3: An MLP with a hidden layer of 3 hidden units.

by effectively capturing nonlinear and nontrivial user/item relationships and dealing with various types of data modalities (such as images and text). The deep learning techniques have been changing the recommendation architectures dramatically and bringing opportunities to improve the performances significantly. In the following, we discuss various representative deep neural networks, such as Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Attention Mechanisms (AMs) and Transformers.

Multilayer Perceptrons (MLPs) A Multilayer Perceptron (MLP) is a fully connected class of feed-forward neural networks and is also referred as the simplest deep network. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Figure 2.3 shows an MLP with a hidden layer of 3 hidden units. MLPs can easily model the nonlinear interactions between users and items by incorporating multiple hidden layers and nonlinear activation functions (e.g., Rectified Linear Unit (ReLU)). For instance, Neural Collaborative Filtering (NCF) (X. He et al., 2017) improves the recommendation performance by leveraging both the linearity of Matrix Factorisation (MF) and the nonlinearity of MLPs. Moreover, intuitively, the elementary compoment of MLPs is the single-layer perceptron network (Taud & Mas, 2018). In particular, the single-layer perceptron network can be formulated as follows:

$$y = \sigma(Wx + b) \tag{2.1}$$

where the single-layer perceptron network computes the output y according to the summation of the input data x, W is a randomly initialised vector and b is a bias term. In particular, $\sigma(\cdot)$ is an activation function, such as ReLU (rectified linear unit), Sigmoid (logistic activation function), Tanh (hyperbolic tangent activation function), that can introduce non-linearity into the neural networks (Sharma, Sharma, & Athaiya, 2017). In this thesis, we adopt MLPs with a non-linear transformation of the input data (such as ReLU, Sigmoid, or Tanh) to extract/generate abstractive representations for the recommendation task.

Convolutional Neural Networks (CNNs) A Convolutional Neural Network (CNN) is a class of neural networks used primarily for processing structured arrays of data such as images (T. Liu, Fang, Zhao, Wang, & Zhang, 2015). A typical CNN consists of five parts: an input layer, con-

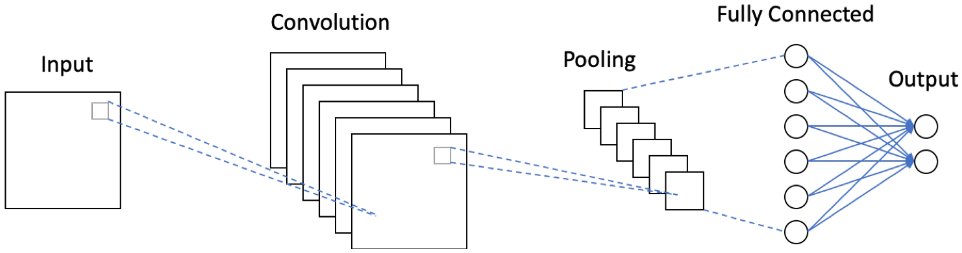
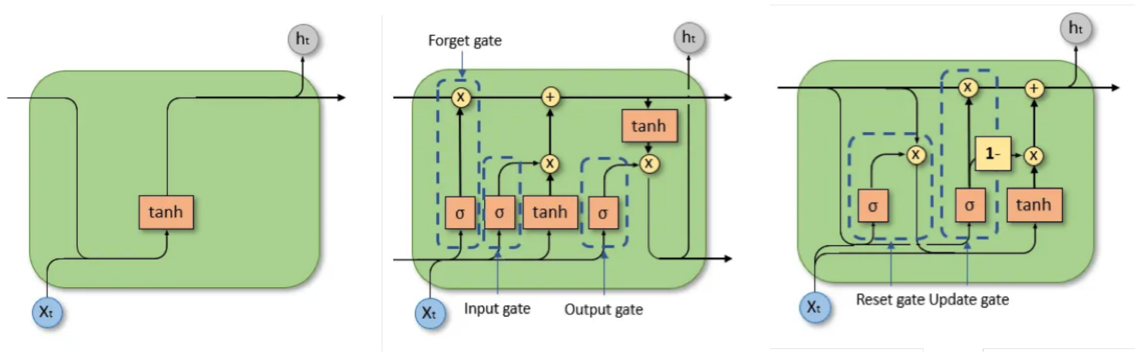


Figure 2.4: An example of the CNN architecture diagram.

volutional layers, pooling layers, fully-connected layers and an output layer. Figure 2.4 shows an example of the CNN architecture diagram. The convolutional layers extract features from the input images with multiple learnable filters (also called kernels) that convolve or slide over the input images to capture local patterns and relationships between pixels. The pooling layers reduce the spatial dimensions while retaining the essential information by down-sampling the feature maps. Then, the outputs are generated by flattening the feature maps into a one-dimensional vector and passing the vector to one or more fully connected layers. A CNN is capable of extracting local and global features from data sources with different modalities, such as text and images. For instance, Deep Cooperative Neural Networks (DeepCoNN) (L. Zheng, Noroozi, & Yu, 2017) extracts rich semantic representations from review texts with CNNs, thereby alleviating the sparsity problem and enhancing the model's interpretability. In this thesis, we adopt CNNs for processing the natural-language sentences and generating the abstractive representations for the textual modality.

Recurrent Neural Networks (RNNs) A Recurrent Neural Network (RNN) a class of neural networks where connections between nodes can create a cycle to capture the dynamics of sequences by maintaining a hidden state or memory of the past information. There are two RNN-variants, namely Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014), for alleviating the so-called gradient vanishing problem. LSTM introduces memory cells and gating mechanisms to selectively remember or forget information over time. The memory cell is responsible for storing and updating the memory state, and the gates (input gate, forget gate, and output gate) control the flow of information into and out of the cell. LSTM's ability to preserve information over long sequences makes it effective in tasks involving longer-term dependencies, such as language modeling and speech recognition. GRU combines the memory cell and hidden state into a single vector and uses two gates (reset gate and update gate) to control the flow of information. The reset gate determines how much of the past information should be forgotten, while the update gate decides how much of the new information should be retained. GRU has been shown to be computationally efficient and effective in various sequence-related tasks (Hidasi, Karatzoglou, Baltrunas, & Tikk, 2016; Manotumruksa, Macdonald, & Ounis, 2018; Quadrana, Cremonesi, & Jannach,



(a) RNN (b) LSTM (c) GRU

Figure 2.5: The architectures of RNNs from Dancker (2022).

2018). Figure 2.5 shows the architectures of typical RNNs from Dancker (2022). RNNs enable the recommender systems to model the temporal dynamics and sequential evolution of the users' preferences. For instance, a GRU-based model (GRU4Rec) (Hidasi et al., 2016) is proposed for sequential recommendation to model the behaviour sequences and predict the next item in a sequence given the last items the user has interacted with.

Attention Mechanisms (AMs) & Transformers The attention mechanism (AM) is a fundamental component in many deep learning models for sequence modeling, particularly in Transformer and the Transformer-based models (such as BERT (Devlin et al., 2019a), GPT-1 (Radford, Narasimhan, Salimans, Sutskever, et al., 2018), GPT-2 (Radford et al., 2019)). The attention mechanisms allows a model to selectively process and weigh different parts of the input sequence during the sequence modelling. There are several popular attention mechanisms, such as additive attention (Bahdanau, Cho, & Bengio, 2014), dot-product attention (Luong, Pham, & Manning, 2015), and scaled dot-product attention (or also called self-attention) (Vaswani et al., 2017). In particular, a Transformer is entirely rely on the self-attention mechanism to draw global dependencies between input and output without using sequence-aligned recurrent architecture. The major component in the Transformer is the unit of multi-head attention mechanism that adopts the scaled dot-product attention. Figure 2.6 shows a diagram of the Transformer architecture with the multi-head attention mechanism and the scaled dot-product attention (Weng, 2018). The input sequence is first transformed into three vectors (known as queries Q, keys K, and values V) that are typically obtained by linearly projecting the input embeddings into a higher-dimensional space. The output of the scaled dot-product attention is a weighted sum of the values, where the weight assigned to each value is determined by the scaled dot-product of the query with all the keys:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{n}})V \tag{2.2}$$

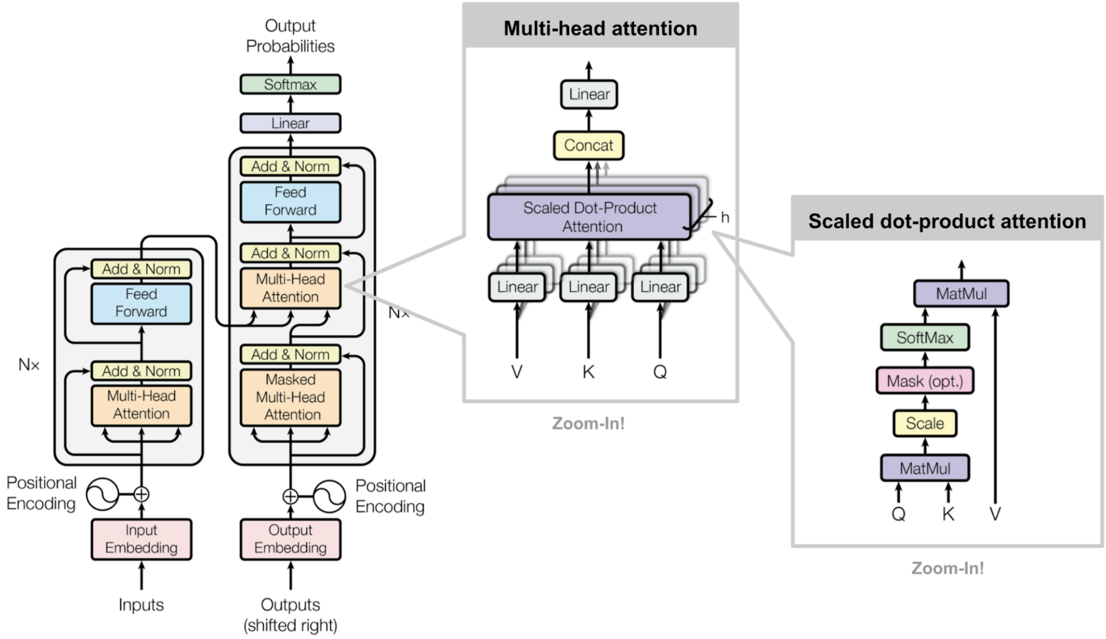


Figure 2.6: A diagram of the Transformer architecture with the multi-head attention mechanism and the scaled dot-product attention from Weng (2018).

where n is the input sequence length. In addition, the multi-head attention mechanism performs the scaled dot-product attention multiple times in parallel and ensembles the multiple outputs with a concatenation to improve the model's performance (Vaswani et al., 2017). The multi-head attention mechanism can be formulated as the following equations:

$$\begin{gathered} MultiHead(Q,K,V) = [head_1;...;head_h]W^O \\ \text{where } head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{gathered} \tag{2.3}$$

and W_i^Q, W_i^K, W_i^V, and W^O are the learned parameter matrices. h is the number of the parallel attention layers (or called heads).

Transformers have been successfully applied to sequential recommendations. For instance, SASRec (Kang & McAuley, 2018) is the first sequential recommender system based on self-attentive mechanism. SASRec models the item sequences as a sequence of embeddings, which are then processed by the multiple self-attention layers to capture long-term dependencies. By attending to different positions within the sequence, SASRec can learn contextual representations that effectively capture user preferences and item relationships. In addition, BERT4Rec (Petrov & Macdonald, 2022; F. Sun et al., 2019) adapts the BERT (Devlin, Chang, Lee, & Toutanova, 2019b) (Bidirectional Encoder Representations from Transformers) model for sequential recommendations. BERT4Rec formulates the sequential recommendation task as a masked language modelling problem, where the goal is to predict the masked items in a sequence based on the surrounding context.

2.1.3 Evaluation of Recommender Systems

The general goals in evaluating recommender systems include factors such as *accuracy*, *diversity*, *serendipity*, *novelty*, *robustness*, and *scalability* (Aggarwal et al., 2016). In particular, the accuracy metrics are used to evaluate either the prediction accuracy of estimating the ratings of specific user-item combinations or the accuracy of the top-k ranking predicted by a recommender system. For the ranking-based recommendation task, the recommender systems aim to ranked the users' preferred items on top according to the estimated preferences from the users' past user-item interactions (Cremonesi, Koren, & Turrin, 2010). Figure 2.2 illustrates an example ranking list of items for recommendation. The goal of a ranking-based recommender system is to prioritise the users' preferred items by assigning them higher scores, thereby placing these items at the top of the list of recommended items. These are various evaluation metrics that can be used for examining the performance of a ranking-based recommender system, such as Precision, Recall, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) (Shi et al., 2012) and Normalised Discounted Cumulative Gain (NDCG) (Järvelin & Kekäläinen, 2002). Assessing a ranking-based recommendation model becomes crucial when considering the performance of the suggested items at various ranking positions. To this end, we mainly evaluate the ranking-based recommender systems with MRR and NDCG:

- **MRR**: The MRR metric measures how well the recommender system ranks relevant (i.e., the user preferred/target) items in response to the users' estimated preferences. The reciprocal rank is the reciprocal of the position (i.e., $rank_u$ for a user $u \in U$) of the first relevant item in the ranked list. MRR takes the average of the reciprocal ranks over a set of users U, as follows:

$$MRR = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{rank_u} \tag{2.4}$$

- **NDCG**: The NDCG metric takes into account the relevance of the ranked items and their positions in the list, providing a more comprehensive evaluation of the ranking quality than simpler metrics like Precision or Recall. NDCG is formulated as:

$$NDCG@N = \frac{DCG@N}{IDCG@N} \tag{2.5}$$

$$DCG@N = \sum_{i=1}^{N} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.6}$$

$$IDCG@N = \sum_{i=1}^{|REL_N|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.7}$$

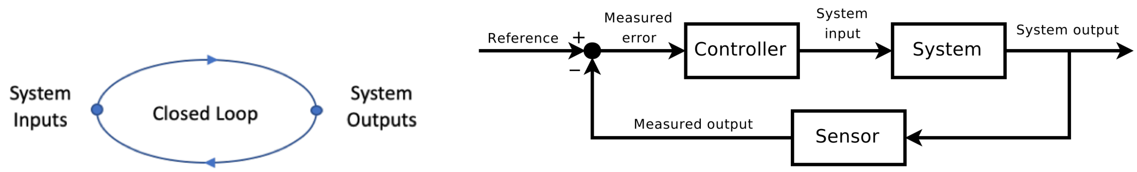
where rel_i is the graded relevance of the item at position i. $DCG@N$ is discounted cumulative gain and it penalises highly relevant documents that appear lower in the search result list. *IDCG@N* is ideal discounted cumulative gain, and REL_N represents the list of relevant items (ordered by their relevance) in the candidate pool up to position N.

Recommender systems can be evaluated either *online* or *offline* (Aggarwal et al., 2016). In the online evaluation of a recommender system, the users' feedback is usually measured with respect to the presented recommendations, such as the *conversion rate* of users clicking on recommended items. For instance, the evaluation of recommender systems based on deep reinforcement learning, such as a news recommendation model named DRN (G. Zheng et al., 2018), are generally conducted in an online production environment of a commercial recommendation application. However, since online evaluations require active user participation, it is often not feasible to use them in research. On the other hand, testing over multiple datasets from multiple domains (e.g., music, movies, news) is particularly important for assuring greater generalisation power of the recommender systems (Aggarwal et al., 2016). In such cases, offline evaluations with historical datasets are used. In recent years, online user-item interaction simulators, which are trained on users' feedback logs, have been deployed to simulate the online environment and evaluate the recommender systems offline in the RL community, such as DEERS (X. Zhao, Zhang, et al., 2018), DeepPage (X. Zhao, Xia, et al., 2018), and LIRD (X. Zhao et al., 2017).

Thus far, we have shown the overview of recommender systems, including the taxonomy of recommender systems, deep learning based recommender systems, and the evaluation of recommender systems. The existing formulations of the recommendation task usually consider the recommendation process as a static process by estimating the users' preferences from their past user-item interactions and predict their next preferred items. However, the users' information seeking process is actually dynamic by involving closed loops with multiple interaction turns between users and recommender systems. In the following section, we will illustrate how/why a recommender system is a closed-loop system, as well as the challenges of a closed-loop recommender system.

2.2 Overview of Closed-Loop Systems

The recommendation process is usually a dynamic process with continuous interactions between users and recommender systems. As we mentioned in Section 1.1, conversational recommender systems can be generally considered to form a *closed loop system* (from *control theory* (Simrock, 2011)) in which the inputs (i.e. users' feedback) of the recommender systems are fully or partially determined by the outputs (i.e. recommended items). In particular, conversational recommender systems can benefit from such closed-loop interactions by tracking and capturing users' current preferences across multiple interactions, thus mitigating the information asymmetry. In this part, we will illustrate the concept of *closed loops* in the recommendation processes.



(a) A closed loop

(b) A block diagram of a closed-loop control system

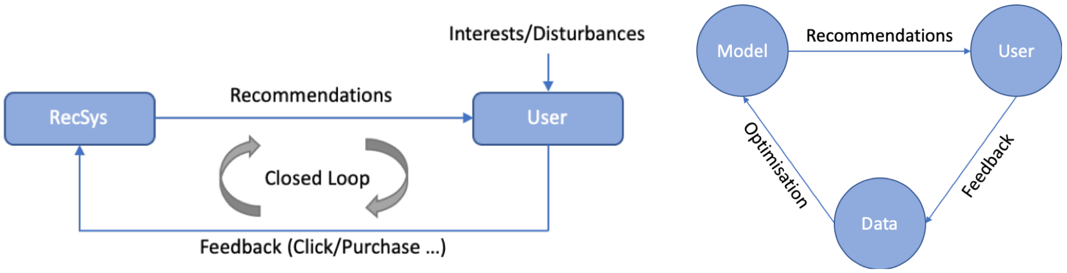
Figure 2.7: The architecture of a closed loop system.

2.2.1 Closed Loops

A closed-loop system, also known as a feedback control system, is a system in which the output of the system is used to self-regulate and adapt to changes or disturbances in its environment. Figure 2.7 (a) shows the formulation of a closed loop where the inputs of a closed loop system are determined, at least in part, by the outputs of the system (Simrock, 2011). Figure 2.7 (b) shows a block diagram of a *closed-loop control system* (Stefani, Shahian, Savant, & Hostetter, 2002) where the system continuously receives feedback about its output with a sensor and uses that information to make adjustments or corrections to its input. The output of the system is compared to a desired reference, and the difference is used to generate an error signal. This error signal is then applied in a controller to initiate a corrective action, adjusting the system's input to reduce or eliminate the error. By continuously monitoring the output and making adjustments, a closed-loop system can maintain stability, accuracy, and desired performance even in the presence of uncertainties or disturbances (Doyle, Francis, & Tannenbaum, 2013). For this reason, closed-loop systems have been widely used in various fields, including engineering, electronics, control systems, automation, and robotics.

2.2.2 Closed Loops in Recommender Systems

A recommender system is a typical closed-loop system in which the inputs (i.e., users' feedback) of the recommendation models are fully or partially determined by the outputs (i.e., reommendations). Figure 2.8 (a) presents a diagram of a closed loop involving a user's feedback and a recommender system's recommendations. In such a closed loop, a user interacts with the recommender systems by clicking or purchasing the recommended items, while a recommender system receives the user's feedback to adjust the recommendation strategies and update the recommendation list of items (Jadidinejad, Macdonald, & Ounis, 2020). Indeed, the users' feedback can be affected by the evolution of the users' internal interests and the disturbances from the environment (such as weathers, seasons, and temperatures). The users' interaction logs can be usually stored offline where the logged behaviour data can be further used for optimising the recommender systems. To this end, there are three major components involved in the closed loops: *models*, *users*, and *data*. Figure 2.8 (b) shows the relationship between the three components.



(a) A diagram of closed loops in recommendations (b) Components in closed loops

Figure 2.8: The closed loops between users and recommender systems.

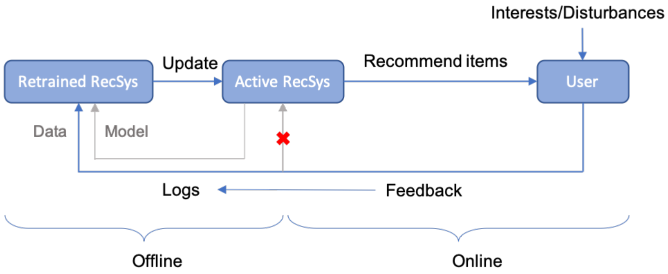


Figure 2.9: Decoupling of data generation and model optimisation.

2.2.3 Properties of Closed Loops

As a widely adopted type of real-world application, recommender systems usually suffer from *delays* and *uncertainties* in the real-world closed loops, which can greatly hinder the recommender systems from understanding and satisfying the users' information needs.

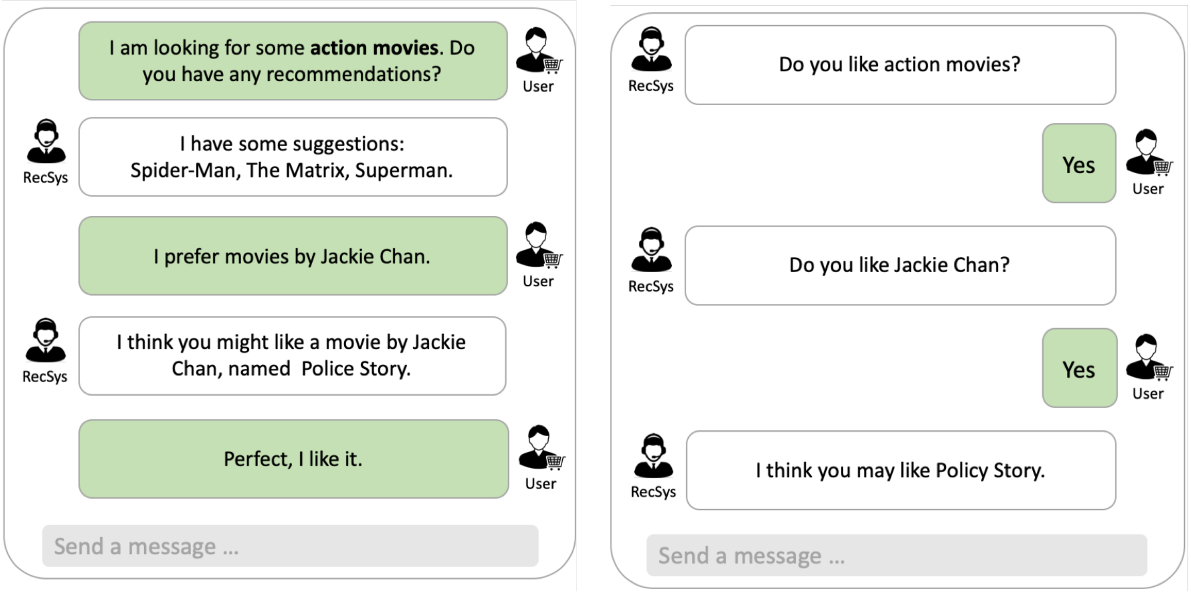
Delays According to the architectures of recommender systems as described in Figure 2.8 (a), delays can appear in data and model updating. Data is usually collected from the users' feedback, which can indicate the users' past preferences estimated from the interaction history or the real-time information needs expressed/indicated by the users' explicit/implicit feedback. In particular, a delay in data updating describes the effect that the instant users' feedback is not immediately considered in the next prediction. In real-world scenarios, users can be unwilling/unable to provide intermediate feedback when they receive recommendations, such as recommendations for restaurants. In this situation, the systems have to wait for the users' feedback corresponding to the current recommendations until the users have checked the recommendations (such as restaurants). In addition, the delay in updating the recommendation model is mainly caused by the decoupling of data generation and model optimisation in the closed loops. Figure 2.9 presents the decoupling of data generation and model optimisation in a closed loop. Concurrently, various recommendation models have different capabilities of modelling the users' sequential behaviours across multi-turn interactions. For instance, non-sequential recommendation models, such as Matrix Factorisation (Section 2.1), make recommendations for the next item regardless of the users' latest feedback. Meanwhile, sequential recommendation models, such as GRU4Rec, SASRec, and BERT4Rec (Section 2.1), can continuously predict the next item using the last user-item interactions as inputs, in order to make up-to-date recommendations.

Uncertainties Uncertainties of the users' feedback can be caused by changes of the users' temporal interests in the closed loops of recommender systems. Uncertainties in the closed loops include the evolution of users' internal preferences and the environmental disturbances on the users' interests, e.g., contextual information. During the recommendation process, the users' information needs or preferences are either expressed by explicit feedback (such as ratings, critiques, or reviews) or implicit feedback (such as clicks, skips, or add-to-cart). Although the recommender systems estimate users' preferences on items and produce recommendations to assist the users in the decision-making process, the users' temporal interests can be continuously changing during the recommendation process. For instance, a user may click different types of products when browsing an e-commerce website. Meanwhile, environmental disturbances, such as weathers, seasons, temperatures, can also greatly influence the user's choices on the items. For instance, the users are more likely to buy a coat than a t-shirt in winter from an e-commerce website.

Therefore, the tasks of modelling and evaluating closed loop interactions in recommender systems mainly focus on mitigating the delays and uncertainties in the closed loops to improve the quality of recommendations. To this end, a *conversational recommender system* is a typical type of systems that can address the *delays* by continuously updating the recommendation strategy with the latest users' feedback/request and mitigate *uncertainties* by allowing the users to express their preferences or needs explicitly, such as natural-language feedback. In this thesis, we mainly focus on modelling and evaluating various conversational recommendation models for effectively mitigating the delay and uncertainty issues in the closed loops.

2.3 Overview of Conversational Recommender Systems

The existing traditional/DL-based recommender systems, such as those described in Section 2.1, usually suffer from a fundamental *information asymmetry* problem (Gao et al., 2021), where a recommender system will never know precisely *what a user likes* and *why a user likes an item*. On one hand, the users' sparse and noisy historical interaction data make it difficult to precisely model the users' preferences, especially when the users' preferences drift frequently. On the other hand, the users' decisions can also be affected by various reason, such as curiosity, mood, and season. Conversational recommendation enables the recommender systems to directly communicate with the engaged users by either asking clarification questions or re-



I am looking for some **action movies**. Do you have any recommendations?

I have some suggestions: Spider-Man, The Matrix, Superman.

I prefer movies by Jackie Chan.

I think you might like a movie by Jackie Chan, named Police Story.

Perfect, I like it.

(a) User-initiative conversational recommendation

Do you like action movies?

Yes

Do you like Jackie Chan?

Yes

I think you may like Policy Story.

(b) System-initiative conversational recommendation

Figure 2.10: An example of user-initiative and system-initiative conversational recommendation.

ceiving natural-language feedback from users, thereby addressing the *information asymmetry* problem (Gao et al., 2021) in information seeking. Developing the form of conversational recommender systems is still an open problem. There are different categories of conversational recommender systems (CRSs), such as system-initiative vs. user-initiative CRSs(Zamani, Trippas, Dalton, & Radlinski, 2022), uni-modal vs. multi-modal CRSs (Deldjoo et al., 2021), and retrieval-based vs. generation-based CRSs (Manzoor & Jannach, 2021).

2.3.1 System-Initiative vs. User-Initiative CRSs

The conversations can be roughly divided into three types based on the initiatives: *system-initiative*, *user-initiative*, and *mixed-initiative*. Here, we mainly focus on system-initiative and user-initiative conversations since the mixed-initiative conversation can be considered as a mixture of system-initiative and user-initiative conversations. Figure 2.10 shows an example of user-initiative and system-initiative conversational recommendation. To this end, user-initiative conversational recommendation typically allows users to actively initiate a conversation with a textual query and provide natural-language feedback with more preferred attributes considering the current recommendations through multi-turn interactions, such as *interactive recommender systems* (Zou et al., 2020) and *critiquing-based recommender systems* (Antognini & Faltings, 2021). Specifically, the interactive recommender system leverages a feedback signal indicating whether and how much the user likes the corresponding recommendation (such as ratings or like/dislike), the *critiquing-based recommender system* leverages the users' feedback on specific attributes of the recommended items so as to narrow down candidate items quickly. Meanwhile, system-initiative conversational recommendation generally adopts a question-based preference elicitation process by deciding when and what to ask, such as *question-based recommender systems* (W. Lei et al., 2020; Y. Sun & Zhang, 2018). Specifically, a question-based recommender system leverages a conversation strategy to determine when to ask and recommend.

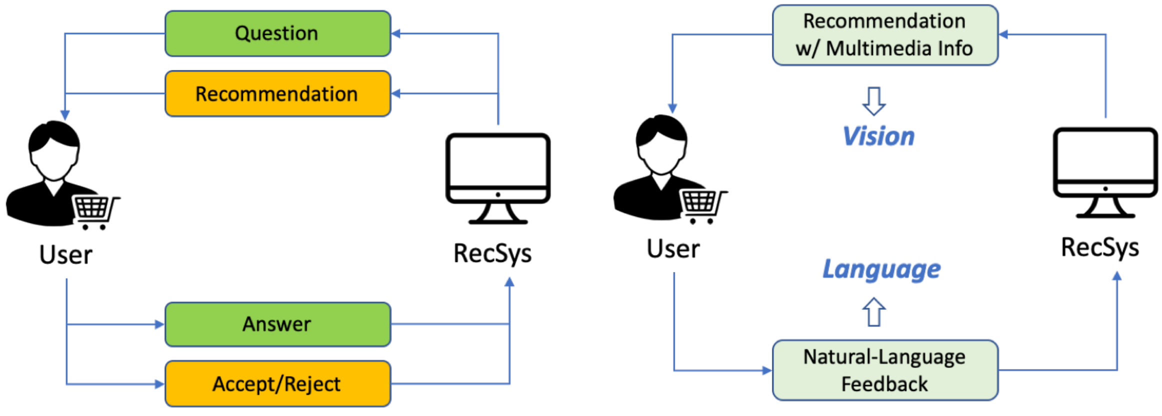
However, the existing formulations of CRSs have demonstrated their drawbacks. For instance, the interactive recommendations suffer from low efficiency when there are too many items (Gao et al., 2021), the critiquing-based recommendations are constrained by the limited attribute-based options for feedback (Gao et al., 2021), and the question-based recommendations (i.e. using pure natural-language interfaces) suffer from less efficient interactions (Iovine et al., 2020; Jannach et al., 2021). To this end, we argue that developing the form of conversational recommender systems is still an open problem. Next, we will describe conversational recommender systems considering different modalities.

2.3.2 Uni-modal vs. Multi-Modal CRSs

The recent research on conversational recommender systems primarily focuses on *uni-modal* interactions and information items, such as the question-based recommender systems using pure natural-language interfaces for recommendations. However, it is widely known that human conversations are *multi-modal*, involving different actions from humans (e.g., word, speech, gesture, facial expression) and different representations of items from machines (e.g., audio, text, image, video) (Deldjoo et al., 2021). In particular, vision-and-language-based interactions between users and recommender systems can be effective for the benefits of both visual information from the recommendations' images and textual information from the users' natural-language feedback (Guo et al., 2018; Uppal et al., 2021; H. Wu et al., 2021; Yuan & Lam, 2021). For instance, the users' natural-language critiques about the visual recommendations can allow the recommender systems to correctly track the users' preferences over time and adapt the systems' instant recommendations, thereby satisfying the users' information needs effectively. In this thesis, we mainly focus on effectively tracking and estimating the users' dynamic preferences from the multi-modal conversational recommendations, so as to develop realistic and effective multi-modal conversational recommender systems.

2.3.3 Retrieval-based vs. Generation-based CRSs

Conversational recommender systems, as a type of conversational system, can be formulated as either retrieval-based and generation-based conversational recommender systems (CRSs). In particular, the retrieval-based systems create responses by either matching items from a candidate pool as a recommendation list or selecting attributes from an attribute pool to formulate clarification questions according to similarity metrics or rules. Retrieval-based CRSs provide



(a) Uni-modal conversational recommendation (b) Multi-modal conversational recommendation

Figure 2.11: An example of uni-modal and multi-modal conversational recommendation.

accurate and attribute-based responses, but they are limited to the information available in their predefined candidate and attribute sets. Meanwhile, generation-based conversational recommender systems generate responses from scratch instead of selecting pre-defined responses. These systems employ techniques such as sequence-to-sequence (Seq2Seq) models or Transformers to generate coherent and contextually relevant responses. Generation-based CRSs have the advantage of being able to generate novel responses, which allows them to handle a wider range of inputs. However, generating responses from scratch can be challenging, and the generated output may sometimes be less accurate or coherent than in retrieval-based CRSs. In this thesis, we mainly focus on retrieval-based conversational recommender systems by ranking the recommendations based on the estimated users' preferences across multiple interaction turns.

Moreover, in the following, we formulate a conversational recommendation process as a decision making process using deep reinforcement learning.

2.4 Preliminaries of Deep Reinforcement Learning

Deep reinforcement learning (DRL) aims to train an agent that can learn from the interaction trajectories provided by the environment by combining the power of deep learning and reinforcement learning. DRL is especially suitable for learning from the closed-loop interactions, such as in CRSs, since it actively learns from the users' real-time feedback. In this section, we describe the essential formulations of decision processes in closed-loop systems, DRL training approaches and recent deep reinforcement learning approaches.

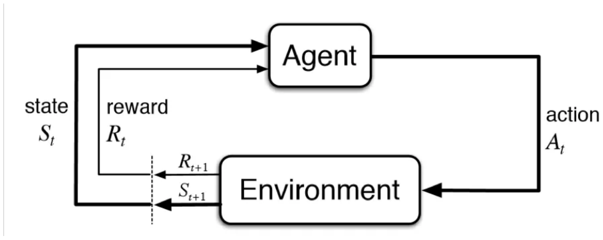


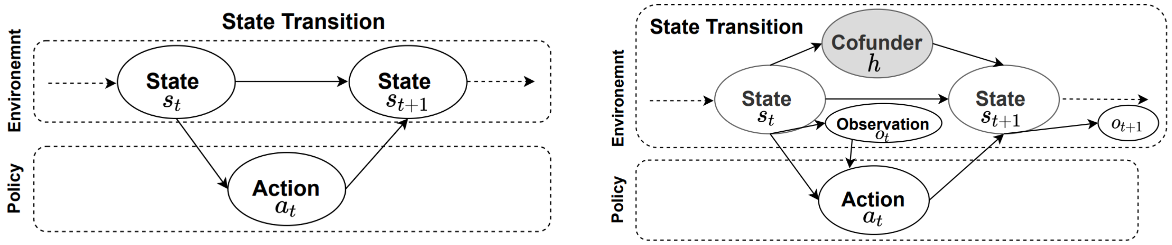
Figure 2.12: An example of reinforcement learning.

2.4.1 Formulations of Decision Making Processes

Deep reinforcement learning provides various formualtion approches for the closed loop systems, such as Markov decision prcesses (MDPs) when the environment is fully observable and partially observable Markov decision processes (POMDPs) when the environment is partially observable. In particular, observability refers to the extent to which an agent (i.e. a recommender system) can perceive and obtain information about its environment (i.e. the users). In a fully observable environment, the agent has access to complete and accurate information about the current state of the environment (i.e. the current preferences of a user). Meanwhile, in a partially observable environment, the agent does not have direct access to the true state of the environment. Instead, it receives partial and often noisy observations (such as the users' explicit or implicit feedback), making it challenging to determine the exact state (i.e. the estimation of the users' preferences). Figure 2.12 shows an example of deep reinforcement learning with the closed-loop interactions between an agent and an environment. Deep reinforcement learning (DRL) enables the recommender systems to continuously update recommendation strategies according to users' latest feedback, and maximise the expected cumulative long-term reward from users (Li, Chu, Langford, & Schapire, 2010; X. Zhao, Xia, et al., 2018; X. Zhao, Zhang, et al., 2018; X. Zhao et al., 2017; G. Zheng et al., 2018).

Markov Decision Process (MDP)

Reinforcement learning (RL) deals with how *agents* ought to take *actions* in an *environment* with certain *states* in order to maximise the notion of cumulative *rewards*. Basic reinforcement learning is modelled as a Markov decision process (MDP) with an assumption that the complete information of the environment is *fully observable*. Figure 2.13 (a) shows an example of an Markov decision process (MDP). There are many variants of RL algorithms, such as Q-learning, SARSA, Policy Gradient, Actor-Critic, and many others (Hasselt, 2010; Konda & Tsitsiklis, 2000; Silver et al., 2014; Sutton & Barto, 2018). A Markov decision process, a classical formalisation of sequential decision making, is used by models, such as DRN (G. Zheng et al., 2018), DEERS (X. Zhao, Zhang, et al., 2018), DeepPage (X. Zhao, Xia, et al., 2018), and LIRD (X. Zhao et al., 2017), for capturing users' dynamic preferences. Deep reinforcement learning with both value-based approaches (such as DQN (Mnih et al., 2013; X. Zhao, Zhang,



(a) MDP

(b) POMDP

Figure 2.13: An example of decision making processes from X. Chen et al. (2021).

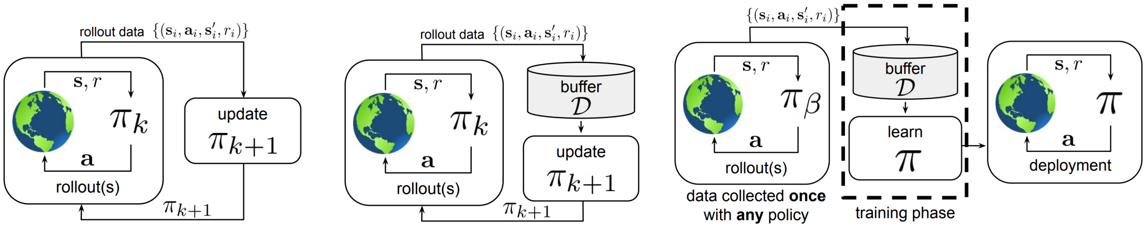
et al., 2018)) and policy-based approaches (such as REINFORCE (M. Chen et al., 2019)) has been widely applied in various recommendation tasks, such as interactive recommedation (Zou et al., 2020), sequential recommendation (Xin et al., 2020), and conversational recommendation (Y. Sun & Zhang, 2018).

Partially Observable Markov Decision Process (POMDP)

When complete information about the environment is not available, a deep reinforcement learning problem can be modelled as a partially observable Markov decision process (POMDP) (Sutton & Barto, 2018). Figure 2.13 (b) shows an example of a partially observable Markov decision process (POMDP). For instance, given only a single game screen, the game of Pong is a POMDP because a single observation does not reveal the velocity of the ball while it only reveals the location of the paddles and the ball (Hausknecht & Stone, 2015). The estimated states, which characterise the distribution over the latent states in a POMDP, are typically modelled using recurrent neural networks (RNNs), and have been shown to be effective for reinforcement learning in POMDP scenarios (Hausknecht & Stone, 2015; Igl, Zintgraf, Le, Wood, & Whiteson, 2018). For example, a Deep Recurrent Q-Network (DRQN) (Hausknecht & Stone, 2015) was proposed to successfully integrate information (i.e. the location of the paddles and the ball) through time in Pong to detect the ball's velocity, although it was capable of seeing only a single screen at each timestep. The POMDP formulation approach has been shown to be suitable for sequential recommendations (Lu & Yang, 2016) and conversational recommendations (Y. Sun & Zhang, 2018) where it is not possible to fully observe a user's actions on all items in a recommender system, as well as all the desired features expressed by the users' natural-language feedback. For the same reason, the POMDP formulation approach is typically also applicable for multi-modal conversational recommendations.

2.4.2 DRL Training Approaches

There are usually three types of DRL training approaches (Levine et al., 2020): on-policy, off-policy, and offline. Figure 2.14 shows an example of DRL training approaches. Figure 2.14 (a)



(a) On-Policy DRL (b) Off-Policy DRL (c) Offline DRL

Figure 2.14: An example of DRL training approaches from Levine et al. (2020).

shows a training framework of on-policy (or online) DRL, where the policy π_k is updated with streaming data collected by π_k itself. Examples of on-policy algorithms include Sarsa, REINFORCE, and Proximal Policy Optimization (PPO). Figure 2.14 (b) shows a training framework of off-policy DRL, where the new policy π_{k+1} learns from data generated by a different policy or behavior policy π_k. In particular, the agent's experience with all previous policies (π_0, π_1, ..., π_k) is appended to a data buffer (or called replay buffer) $\mathscr{D}$ for updating a new policy π_{k+1}. Examples of off-policy algorithms include Q-learning, Deep Q-Networks (DQN), and Twin Delayed DDPG (TD3). Different from on-policy and off-policy algorithms that are able to explore during the training processes, offline DRL (Figure 2.14 (c)) employs a dataset $\mathscr{D}$ collected by some (potentially unknown) behaviour policy π_β. The training process does not interact with the MDP at all, and the policy is only deployed after being full trained. The offline DRL algorithms are data-driven approaches and can use large previously collected datasets. Examples of offline DRL algorithms include Batch Q-learning, Batch-Constrained Q-learning (BCQ) and Soft Actor-Critic Offline (SAC Offline).

2.4.3 Recent Reinforcement Learning Approaches

There are various reinforcement learning approaches, such as self-supervised reinforcement learning (SSRL), goal-oriented reinforcement learning (GORL), and hierarchical reinforcement learning (HRL).

Self-Supervised Reinforcement Learning (SSRL)

Self-supervised reinforcement learning (SSRL) is a combination of self-supervised learning and reinforcement learning (RL) techniques. In self-supervised reinforcement learning, the agent leverages self-supervised learning methods to learn useful representations or features from raw or partially observed sensory input. These learned representations serve as a knowledge base for the RL agent to make informed decisions and solve tasks more efficiently. Self-supervised reinforcement learning approaches, such as the Supervised Q-learning (SQN) framework (Xin et al., 2020), have demonstrated a generally better performance compared to neural recommendation models using supervised learning, such as GRU4Rec (Hidasi et al., 2016), Caser (Tang

& Wang, 2018) and SASRec (Kang & McAuley, 2018). In particular, the SQN framework (Xin et al., 2020) extends existing sequential recommendation models (Hidasi et al., 2016; Kang & McAuley, 2018) with a Q-learning layer to introduce *reward*-driven properties to the recommendation process.

Goal-Oriented Reinforcement Learning (GORL)

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 (M. Liu et al., 2022). A goal is defined as “a cognitive representation of a future object" (Colas et al., 2022), which the agent is committed to achieve or maintain. The goal-oriented reinforcement learning approaches have been shown to improve sample efficiency by learning from self-generated rewards (i.e. intrinsic rewards) when the external rewards are sparse. For example, GoalRec (K. Wang et al., 2021), a novel model-based model based on a Dueling Deep Q-Network (DDQN), designed a disentangled universal value function with the users' desired future trajectory (i.e. goal). In addition, a novel multi-goals abstraction-based deep hierarchical reinforcement learning algorithm (MaHRL) (D. Zhao et al., 2020) generated 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.

Hierarchical Reinforcement Learning (HRL)

Common approaches to reinforcement learning are seriously challenged by large-scale applications involving huge state/action spaces and sparse delayed reward feedback (Rafati & Noelle, 2019). 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 (Sutton, Precup, & Singh, 1999), Hierarchical of Abstract Machines (HAMs) (Parr & Russell, 1997), and MAXQ value function decomposition (Dietterich, 2000). The existing recommender systems with HRL (Greco, Suglia, Basile, & Semeraro, 2017; Y. Lin et al., 2022; Xie, Zhang, Wang, Xia, & Lin, 2021; D. Zhao et al., 2020) 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 (Greco et al., 2017) 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).

2.5 Conclusions

In this chapter, we have provided a comprehensive overview of recommender systems in the literature. We illustrated various types of recommender systems, including traditional recommender systems and deep learning-based recommender systems. We also introduced how a recommender system is evaluated. Then, we discussed about the closed-loop systems, including the definition and properties of closed-loop systems, closed loops in recommender systems, delay and uncertainty issues in closed loops. Then, we introduced different types of conversational recommendation, including system-initiative vs user-initiative recommender systems, unimodal vs multi-modal conversational recommender systems, and retrieval-based vs generation-based conversational recommender systems. Finally, we also describe the preliminaries of deep reinforcement learning algorithms, including formuations of decision making processes, DRL training approaches, and recent reinforcement learning approaches (such as self-supervised reinforcement learning, goal-oriented reinforcement learning and hierarchical reinforcement learning). In the next chapter, we describe our proposed multi-modal conversational recommendation framework and discuss the challenges to address within each framework component. In particular, we leverages the recent reinforcement learning approaches (Section 2.4), such as SSRL, GORL, and HRL, to effectively formulate the conversational recommendation task.

Chapter 3

A Multi-Modal Conversational Recommendation Framework

As discussed in Section 1.1, this thesis aims to leverage multi-modal interactions between users and recommender systems (including both visual and textual information) effectively for tracking and estimating the users' dynamic preferences, so as to develop realistic and effective multi-modal conversational recommender systems. Indeed, as we mentioned in Section 2.3, the formulation of conversational recommender systems is still an open problem. In this chapter, in Section 3.1, we first explore the conversational recommendation techniques in the literature to distinguish our main contributions from existing work. Next, in Section 3.2, we introduce our framework for multi-modal conversational recommendations. In particular, we consider two different scenarios about recommendation tasks: cold-start users and warm-start users. We also introduce the methodology for evaluating the success of recommender systems for these tasks. Then, in Section 3.3, we illustrate the users simulator for training and evaluating the multi-modal conversational recommender systems, including the formulation of user simulators, datasets, and evaluation metrics. Finally, in Section 3.4, we discuss the opportunities within the framework for enhancing the performance of multi-modal conversational recommender systems.

3.1 Related Work

Multi-modal conversational recommendation is an emerging topic in recent years and 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) (Chakraborty et al., 2021; Deldjoo et al., 2022; Guo et al., 2018; Liao, Long, Zhang, Huang, & Chua, 2021; Yu et al., 2020). Figure 3.1 il-

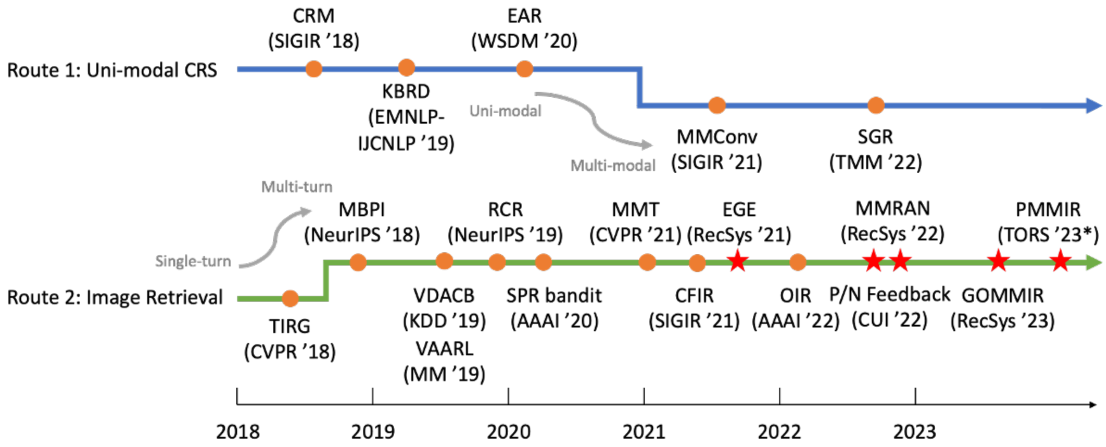


Figure 3.1: A developing process of multi-modal conversational recommender systems in two different routes: (1) enhancing uni-modal (i.e. text-based) conversational recommender systems with multi-modal (i.e. both textual and visual) information, (2) enhancing image retrieval models with multi-turn interactions. * denotes under-review. Red denotes our work.

lustrates a timeline of development of multi-modal conversational recommender systems. In particular, there are two primitive technical routes for developing multi-modal conversational recommender systems: (1) enhancing uni-modal (i.e. text-based) conversational recommender systems with multi-modal (i.e. both textual and visual) information, (2) enhancing image retrieval models with multi-turn interactions. The first route usually follows the generation-based conversational recommender systems (mentioned in Section 2.3.3) by considering multi-modal information in the conversations and generating natural-language responses, while the second route usually follows the retrieval-based conversational recommender systems (mentioned in Section 2.3.3) by matching items from a candidate pool as a recommendation list.

Both routes start from a simplified task to reduce the complexity of the models. In particular, the first route starts with combining recommender systems and conversational systems together. For instance, Y. Sun and Zhang (2018) proposed a unified framework (called CRM) to integrate recommender systems (Factorization Machine (FM)) and dialogue system technologies (a LSTM-based Belief Tracker to understand the user's intention correctly and a MLP-based Policy Network to make sequential decisions and take appropriate actions in each turn) together for building an intelligent conversational recommender system. KBRD (Q. Chen et al., 2019) bridged the gap between recommender system and dialog system via knowledge propagation. Dialog information (such as the users' responses to the system's clarification questions) is effective for the recommender system especially in the setting of cold start, and the introduction of knowledge can strengthen the recommendation performance significantly. In addition, information from the recommender system that contains the user preference and the relevant knowledge can enhance the consistency and diversity of the generated dialogs. EAR (W. Lei et al., 2020) formulated the conversational recommendation task into three stages, i.e. Estimation–Action–Reflection, to better converse with users. Specifically, Estimation builds predictive models to estimate user preference on both items and item attributes; Action learns a dialogue policy to determine whether to ask attributes or recommend items, based on Estimation stage and conversation history; and Reflection updates the recommender model when a user rejects the recommendations made by the Action stage. There are also many other conversational recommendation models (Gao et al., 2021; Jannach et al., 2021; Xu et al., 2021) to further improve the recommendation performance based on the previously mentioned formulations (i.e. CRM (Y. Sun & Zhang, 2018), KBRD (Q. Chen et al., 2019), and EAR (W. Lei et al., 2020)). However, all of them are uni-modal conversational models that leverage text only. The real breakthrough of conversational recommender systems has largely been blocked by a comprehensive multi-modal conversational search environment for facilitating the corresponding research tasks. To this end, a Multimodal Multi-domain Conversational dataset (MMConv) (Liao et al., 2021), specifically a fully annotated collection of human-to-human role-playing dialogues spanning over multiple domains and tasks (including multi-modal conversational recommendations), has been introduced recently. By leveraging the MMCov environment, a State Graph-based Reasoning model (SGR) (Y. Wu, Liao, et al., 2022) was proposed to explicitly model the users' dynamic preferences and integrate with a multi-modal knowledge graph for better state representation. Although MMConv provides a good foundation for the research in multi-modal conversational recommendations across different domains (such as food, hotel, nightlife, mall, and sightseeing), MMConv is constrained with limited dialogues and turns. In addition, it is also time-consuming and unrealistic to extend to other domains (such as books, music, movies, and e-commerce) by collecting conversations from real human directly. To this end, the development of the route beginning with uni-modal conversational recommender systems is mainly hindered by the limited availablility of multi-modal conversational recommendation datasets (Liao et al., 2021). We are still optimistic in such a technical route by expecting more research on MMConv and more multi-modal conversational recommendation datasets in the future.

Meanwhile, the other technical route from image retrieval is more realistic and easier to progress than the previous technical route. In particular, the task of image retrieval can be formulated as a critiquing-based search process, where the input query is specified in the form of an image plus some text that describes desired modifications to the input image. The target of the critiquing-based search process is to find the target item(s) with less effort by users, such as fewer interaction turns. The recommender systems guide the users toward the users' target items by recommending items with users' preferred attributes according to both the users' critiques and the systems' previously recommended items. In particular, TIRG (called Text Image Residual Gating) (Vo et al., 2019) was proposed to address the composition of image and text in the context of image retrieval. Parallel to this work, MBPI (model-based policy improvement) (Guo et al., 2018) firstly extended the critiquing-based image retrieval task with multi-turn interactions and focused on modeling the interactions between users and the agent. MBPI formulated the task of dialog-based interactive image retrieval as a reinforcement learning problem, where

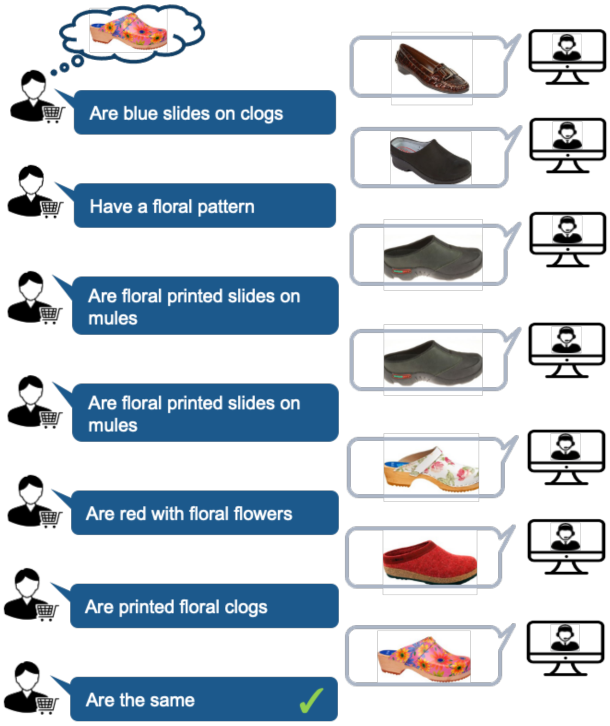

Are blue slides on clogs
Have a floral pattern
Are floral printed slides on mules
Are floral printed slides on mules
Are red with floral flowers
Are printed floral clogs
Are the same ✓

Figure 3.2: An example of multi-modal conversational recommendations with MBPI.

the dialog system is rewarded for improving the rank of the target image during each dialog turn. Figure 3.2 illustrates an example of multi-modal conversational recommendations with MBPI (Guo et al., 2018). A relative captioner was also proposed to act as a surrogate for real human users by automatically generating sentences that can describe the prominent visual differences between any pair of target and candidate images. Such a natural-language feedback generation process with the user simulator is very similar to a scenario of a shopping conversation session between a shopping assistant and a customer. The relative captioner was well-trained with crowdsourced datasets and carefully evaluated with real human users. Such a user simulator enables to train and test the interactive system actively across multi-turn interactions. Following the MBPI (Guo et al., 2018) work, various vision-language interactive recommender systems based on GRU/LSTM were proposed to further improve the performance by leveraging cascading bandit (VDACB (visual dialog augmented cascading bandit) (Yu, Shen, & Jin, 2019)), actor-critic with visual attributes (VAARL (visual attribute augmented reinforcement learning) (Yu, Shen, Zhang, et al., 2019)), constrained reinforcement learning (RCR (reward-constrained recommendation) (R. Zhang et al., 2019)), pairwise ranking bandit (SPR bandit) (Yu et al., 2020), offline reinforcement learning (OIR (offline interactive recommendation (R. Zhang, Yu, Shen, & Jin, 2022))). Furthermore, MMT (multi-modal Transformer) (H. Wu et al., 2021) leverages a Transformer encoder for fusing the visual and textual sequences (including images, natural-language feedback, and attributes) to better elicit the users' preferences cross multi-turn interactions. CFIR (conversational fashion image retrieval) (Yuan & Lam, 2021) leveraged three modules for compositing visual and textual feature representations, comparing the difference between the reference image and the candidate image, and exploiting the attribute information of the candidate image for calculating the mutual attention between candidate image and feedback texts.

However, the existing multi-modal conversational recommender systems have demonstrated their limitations, as follows:

- *Partial observations*: Despite the expressiveness of natural-language feedback in multi-modal conversational recommendations, the users' feedback can only allow the recommender system to obtain a partial portrayal of the users' preferences. Such partial observations of the users' preferences from their natural-language feedback can drive the recommender system towards a degenerate preference estimation that ignores certain features in the historical observations (Gangwani, Lehman, Liu, & Peng, 2020), i.e. historical natural-language feedback and historical recommendations. For instance, recommendations by the MBPI model can violate the users' preferences from previous natural language feedback and can also be repeated. We provide more detailed analysis in Chapter 4.

- *Multi-modal sequence dependency*: The multi-modal conversational recommendation task has been previously modelled using recurrent neural networks (RNNs, using a gated recurrent unit (GRU) (Guo et al., 2018; Yu et al., 2020) or a long short-term memory (LSTM) (R. Zhang et al., 2019)) or using a transformer (H. Wu et al., 2021) as *a state tracker* for both *multi-modal sequence combination* (Beard et al., 2018; Gkoumas, Li, Lioma, Yu, & Song, 2021) (i.e. combining the users' natural-language feedback sequence and the systems' visual recommendation sequence) and *dialog state tracking* (Fu, Xian, Zhang, & Zhang, 2020; Liao et al., 2021; Y. Sun & Zhang, 2018) (i.e. eliciting the users' preferences over time). However, the actual neural networks adopted as the state trackers (such as GRUs (Chung, Gulcehre, Cho, & Bengio, 2014), LSTMs (Hochreiter & Schmidhuber, 1997) or transformers (Vaswani et al., 2017)) are all originally designed for *single-modal* sequence modelling tasks (such as natural language processing (Otter, Medina, & Kalita, 2020)). These GRU/LSTM-based and transformer-based models suffer from an inability to capture multi-modal sequence dependencies, because of their limitations in either *combining multi-modal sequences* with a concatenation operation or *tracking dialog states* by inferring directly from all the concatenated textual and visual representations at all turns instead of the multi-modal abstract representations of the past interactions. We give more detailed analysis in Chapter 5.

- *Coupling of representation learning and policy optimisation*: The representations of visual candidate items and natural-language feedback are initially generated with pre-trained models (such as ResNet (K. He, Zhang, Ren, & Sun, 2016) for image encoding and BERT (Devlin et al., 2019a)/GloVe (Pennington, Socher, & Manning, 2014) for text encoding), and are then implicitly further tuned along with the recommendation policy optimisation. Most existing multi-modal conversational recommendation models adopt a simple concatenation operation for feature composition. However, learning representations in an end-to-end fashion in DRL is usually unstable (Eysenbach, Zhang, Salakhutdinov, & Levine, 2022; Laskin, Lee, et al., 2020; Yarats, Kostrikov, & Fergus, 2020) 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) (Eysenbach et al., 2022). 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. We provide more detailed analysis in Chapter 6.

- *Personalisation*: 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. their current preferences) representations over time. Indeed, the existing multi-modal conversational recommendation models typically simplify the multi-modal conversational 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 (Hidasi & Karatzoglou, 2018; Hidasi et al., 2016; Kang & McAuley, 2018; F. Sun et al., 2019) currently only consider the explicit/implicit past user-item interactions (such as purchases and clicks) in the sequence modelling. We give more detailed analysis in Chapter 7.

- *The realism of simulated conversations with positive/negative feedback*: Despite the generally good performances in the multi-modal conversational recommendation task, these research only focus on positive natural-language feedback with the users' preferred attributes in the top-K (in particular $K = 1$) recommendation task. However, the users in the real-world shopping scenario can freely express their satisfaction over the top-K ($K \geq 1$) recommendations *positively* or *negatively*. Therefore, both positive and negative natural-language feedback should be directly incorporated into the multi-modal conversational recommendation models to ensure the realism of simulated conversations. We will give a more detailed analysis in Chapter 8.

In summary, we have categorised multi-modal conversational recommendation approaches into two technical routes: either enhancing uni-modal conversational recommender systems with multi-modal information or enhancing image retrieval models with multi-turn interactions. The first route with uni-modal CRSs can be hindered by the limited available multi-modal recommendation datasets, while the second route is more realistic and easier to proceed based on the existing formulations. However, we argue that there are several limitations in the existing multi-modal conversational recommender systems in terms of the partial observations of the users' preferences over time, the multi-modal sequence dependency, a coupling of multi-modal composition representation learning and recommendation policy optimisation, the personalisation with the users' long-term and short-term interests, and the realism of simulated conversations with positive/negative natural-language feedback. Therefore, in this thesis, we aim to address and mitigate the above-mentioned limitations in the literature. In the next section, we illustrate the overall framework of the multi-modal conversational recommendation task, including the task formulations with cold-start and warm-start users, as well as how we measure the success of the multi-modal conversational recommendation task.

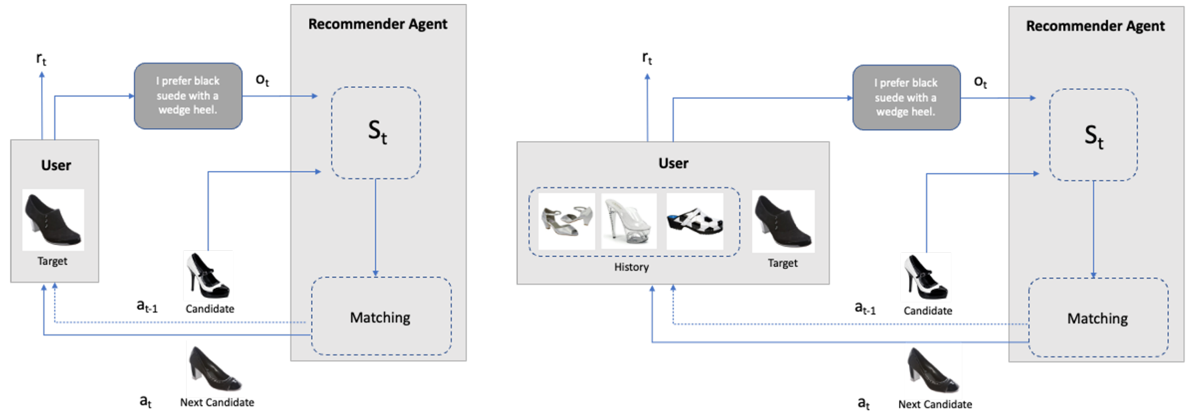
3.2 Framework Overview

The multi-modal conversational recommendation task has been usually formulated as a conversation process between a user and a recommender system (as we have shown in Section 3.1 with MBPI (Guo et al., 2018)). In this thesis, we follow such a formulation of the multi-modal conversational recommendation task, while we assume that the users can be cold-start users and/or warm-start users and we measure the success of multi-modal conversational recommendations in terms of both higher ranking performance (such as NDCG and MRR in Section 2.1) and less interaction turns (called Success Rate (SR)).

3.2.1 Task Formulations

A multi-modal conversational recommendation task is where users can express natural-language feedback as critiques about the visual recommendations when interacting with the recommender system according to the task formulation in MBPI (Guo et al., 2018). In the real-world scenario, a user can be a cold-start user without any interaction history or a warm-start user with many past interactions. To this end, we adapt the existing formulation of the multi-modal conversational recommendation task into such two different scenarios (i.e. cold-start and warm-start users). Figure 3.3 shows two different multi-modal conversational recommendation tasks with two different assumptions:

- Task 1 (as shown in Figure 3.3 (a)): These users are *cold-start users*, **without** *interaction history*, while keeping a target item in their minds. The interactive recommendation pro-



(a) Task 1 with cold-start users

(b) Task 2 with warm-start users

Figure 3.3: Multi-modal conversational recommendation task with cold-start users and warm-start users.

cess starts with an initial random recommendation. The users give natural-language critiques about the visual recommendation at each interaction turn, while the recommender systems track and capture the users' preferences based on their natural-language feedback and give next recommendations.

- Task 2 (as shown in Figure 3.3 (b)): These users are *warm-start users* **with** *interaction history* and keep a target item in their minds. The recommender systems estimate the users' general preferences based on their interaction history and start the recommendation process with a personalised initial recommendation. During the recommendation process, the recommender systems need to consider both the users' general preferences (estimated with their interaction history) and their current needs (i.e. the target item) in the next recommendations.

3.2.2 Measurements

The success of the above tasks (i.e. Task 1 with cold-start users and Task 2 with warm-start users) 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. Figure 3.4 shows an example of a recommendation scenario to illustrate how the users can obtain their target items through interactions with the recommender systems in the top-1 recommendation scenario. In particular, the users are cold-start users without interaction history and receive an randomly sampled item as a recommendation at the initial turn (i.e. turn=0). In addition, Figure 3.5 shows an example of top-K recommendations when more items are exposed to the users at each turn. In particular, a cold-start user receives an randomly sampled item as a recommendation at the initial turn (i.e. turn=0), while a warm-start user perceives a list of personalised recommendations at the initial turn.

The recommender system ranks items based on the ranking scores (i.e. the similarities be-

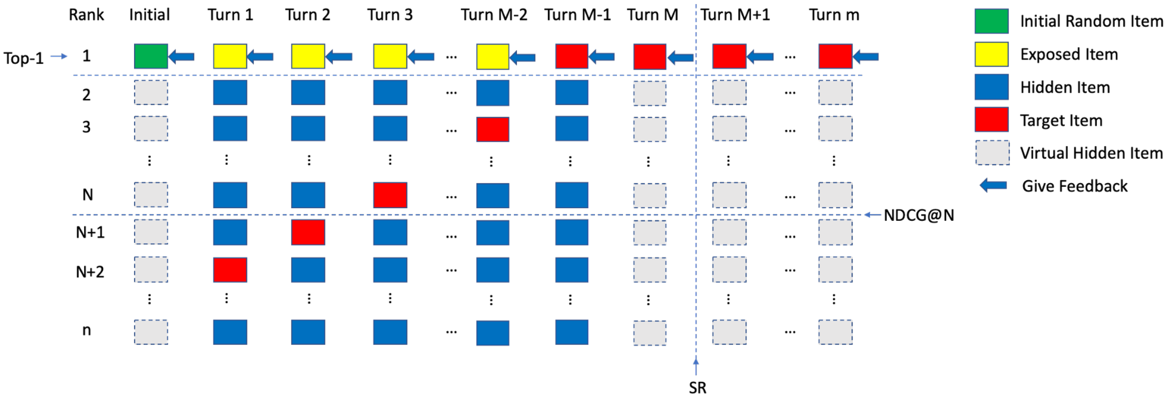


Figure 3.4: A successful top-1 multi-modal conversational recommendation procedure for Task 1.

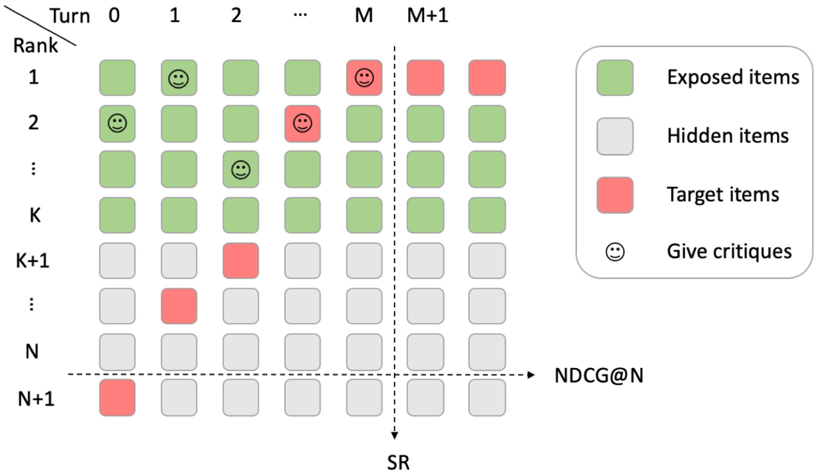


Figure 3.5: An example of top-*K* multi-modal conversational recommendations for both Task 1 and Task 2.

tween the estimated preferences and all items), while the users give feedback on the single top-ranked recommendation presented to them. Hence, the effectiveness can be measured by the percentage of user sessions for which the target item is presented at the top rank by interaction turn M. Furthermore, it is possible that the user may view more of the ranking of items at each interaction turn, down to rank N. Therefore, we define success in the multi-modal conversational recommendation task as higher values in top-heavy metrics such as NDCG@N with a truncation at rank N calculated at the M-th turn, or Success Rate (SR) at the M-th turn.

So far, we have formulated the multi-modal conversational recommendation task with both cold-start and warm-start users. We also measure the success of the multi-modal conversational recommendation task considering both the ranking performance at each turn, as well as the efforts (i.e. turns) needed to get the target items. Next, we will illustrate how we train and evaluate the multi-modal conversational recommendation models by leveraging user simulators.

3.3 User Simulators

Training and evaluating the multi-modal conversational recommender systems with real human users are usually expensive, time-consuming, and do not scale (Gao et al., 2021; S. Zhang & Balog, 2020). User simulators for relative captioning can act as a reasonable proxy of real users for training and evaluation by simulating the users' natural-language feedback for multi-modal conversational recommendations. In the following, we illustrate the detailed implementations of the user simulators, as well as use cases with different user simulators for relative captioning. Then, we demonstrate the datasets available for enabling our research in multi-modal conversational recommendations. The advantages of the user simulators for relative captioning are demonstrated to support our research, while their limitations are illustrated to provide directions for future work. While real human evaluation can provide valuable insights and feedback, we have chosen to leave it as a part of our future work due to considerations such as cost, time constraints, and scalability.

3.3.1 Architecture of User Simulators

The multi-modal conversational recommendation task (in Section 3.2) is specifically concerned with a goal-oriented sequence of interactions between users and recommender systems, where users can continuously receive visual recommendations (i.e. the items' images) and express fine-grained natural-language critiques about the recommendations in terms of their preferences. In particular, such users' natural-language feedback corresponding to the visual recommendations allows a multi-modal conversational recommender system to obtain richer information relating to users' current preferences, thereby leading to a more suitable recommendation compared to clickthrough data and ratings (R. Zhang et al., 2019). However, it is challenging to train and evaluate the multi-modal conversational recommender systems by either getting real human users involved in the interaction processes or collecting and annotating entire multi-modal conversations which are expensive, time-consuming, and do not scale (Gao et al., 2021; S. Zhang & Balog, 2020). Indeed, multi-modal conversational recommender systems can be generally considered to form *closed loop systems* (in Section 2.2), in which the inputs (i.e. the users' natural-language feedback) of the recommender systems are fully or partially determined by the outputs (i.e. the visual recommendations). Nevertheless, the sequences of the recommended items in the collected conversations are usually not aligned with the sequences of recommendations generated by the multi-modal conversational recommender systems, which results in less usefulness of the annotated users' natural-language feedback from the collected conversations.

To learn satisfactory multi-modal conversational recommender systems, user simulators based on vision and language have been considered as a surrogate for real human users in the optimisation and evaluation processes (Guo et al., 2018; H. Wu et al., 2021; Yu, Shen, & Jin, 2019;

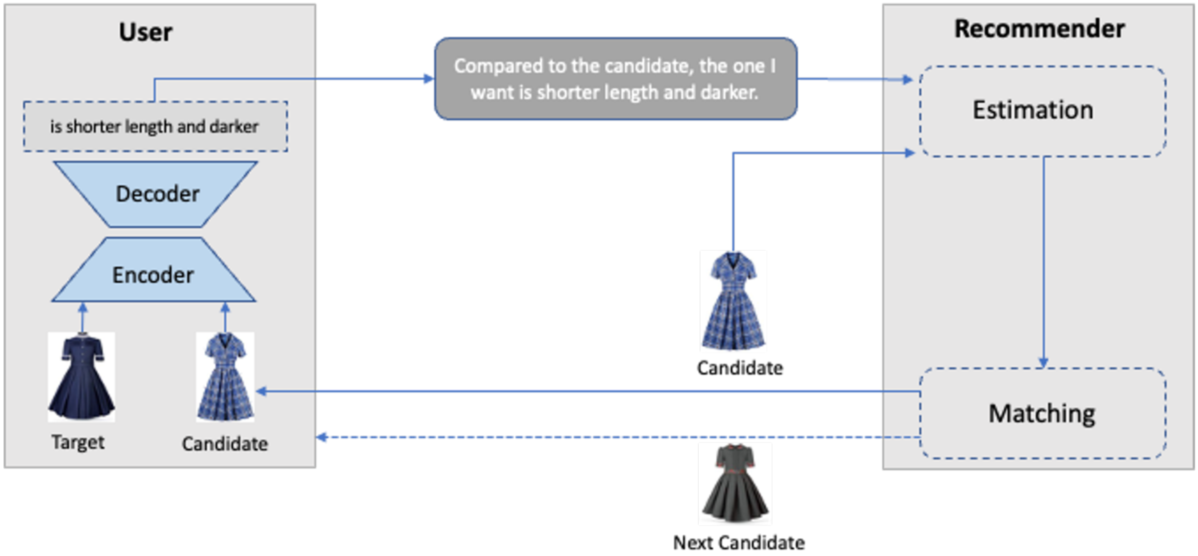

User
is shorter length and darker
Decoder
Encoder
Target
Candidate
Compared to the candidate, the one I want is shorter length and darker.
Recommender
Estimation
Candidate
Matching
Next Candidate

Figure 3.6: Multi-modal conversational recommendation diagram with a user simulator and a recommender system.

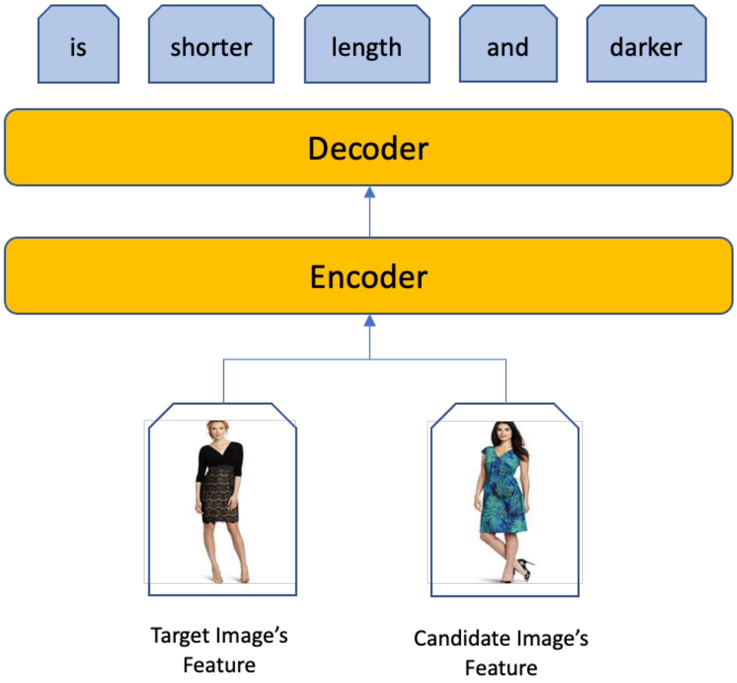


Figure 3.7: Architecture of a user simulator.

Table 3.1: Statistics of *Shoes*, *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees*.

	Shoes			Dresses			Shirts			Tops & Tees		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Relative Captioning Triplets	10,751	-	-	11,970	4,034	-	11,976	4,076	-	12,054	3,924	-
Images$_{caption}$	-	-	-	7,182	2,454	-	8,555	2,966	-	8,387	2,808	-
Images$_{origin}$	10,000	-	4,658	11,452	3,817	3,818	19,036	6,346	6,346	16,121	5,374	5,374

Yu et al., 2020; R. Zhang et al., 2019). Such user simulators have been generally formulated as *relative captioners* for fashion recommendation (Guo et al., 2018; H. Wu et al., 2021) that can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images (i.e. targets as users' preferences and candidate as recommendations). In Figure 3.6, the simulated user (i.e. user simulator) is formulated as a relative captioner with an encoder-decoder structure to generate natural-language feedback given both the target and candidate images. Both the LSTM-based and Transformer-based encoder-decoder approaches for relative captioning have been explored (Guo et al., 2018; Yu, Shen, & Jin, 2019; R. Zhang et al., 2019). Figure 3.7 illustrates the inputs and outputs of a relative captioner. For instance, Guo et al. (2018) applied long short-term memory network (LSTM) based models (Rennie, Marcheret, Mroueh, Ross, & Goel, 2017), such as *Show, Tell* (Vinyals, Toshev, Bengio, & Erhan, 2015), to generate the relative captions as natural-language critiques about the recommendations. Furthermore, H. Wu et al. (2021) proposed a vision-language Transformer-based model for relative captioning (denoted as VL-Transformer), which has shown a better performance in generating relative captions compared to the previous LSTM-based models. Such a natural-language feedback generation process with the user simulator is very similar to a scenario of a shopping conversation session between a shopping assistant and a customer. Such a user simulator can be used for both training and evaluating the multi-modal conversational recommendation models. The success of the relative captioning task is defined as the higher quality of the generated natural language critiques given a pair of images compared to the ground truths (i.e. relative captions) that are collected by crowdsourcing.

3.3.2 Datasets

Fashion is a typical domain that involves multiple modalities, such as images and textual descriptions of products. In the fashion scenario, there are mainly two relative captioning datasets, namely *Shoes*[1] (Berg, Berg, & Shih, 2010; Guo et al., 2018) and *Fashion IQ*[2] (H. Wu et al., 2021), for optimising the user simulators to generate expressive natural-language feedback close to the real users' behaviours. In particular, the *Fashion IQ* dataset includes three subsets containing different fashion categories, i.e. *Dresses*, *Shirts*, *Tops & Tees*. Indeed, both datasets are among the few that provide *relative captions* (created by human assessors) of image pairs that

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

can be used for training and testing the user simulator, as well as the images of the fashion products for training and testing the recommendation models. The *Shoes* and *Fashion IQ* datasets have previously been used by Guo et al. (2018); Yu, Shen, and Jin (2019); R. Zhang et al. (2019) and H. Wu et al. (2021), respectively, for the multi-modal conversational recommendation task, and we replicate their setup (relative captioner, user simulator, etc.). All datasets provide triples (i.e. $\langle a_{target}, a_{candidate}, o_{caption} \rangle$) for training/testing the user simulators. 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.

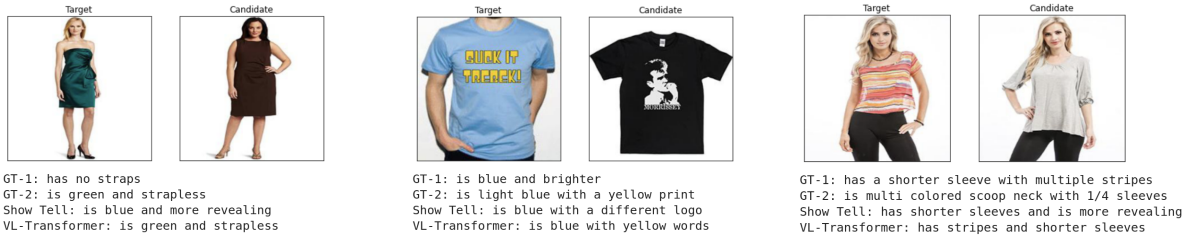
Indeed, these datasets are among the few that provide well pre-processed *relative captions* of image pairs that can be used for training and testing the user simulator, as well as the images of the fashion products for training and testing the recommendation models. For instance, in the *Shoes* dataset, there are $10,751$ relative captions (with one caption per pair of images about their visual differences) and $3,600$ discriminative captions (with one caption per image about their discriminative visual features) for training a user simulator. The *Shoes* dataset also provides $10,000$ images for training the recommender systems, and $4,658$ images for testing. Meanwhile, in the *Dresses* dataset, there are $7,347$ pairs of accessible images with two captions per pair. In particular, the relative captions of the $5,478$ pairs from the *Dresses* dataset are used for training a user simulator, and the relative captions of the $1,869$ pairs for testing. We also extract $7,182$ unique images from the $5,478$ pairs for training the recommender systems, and $2,454$ unique images from the $1,869$ pairs for testing. Here we denote the extracted datasets as $\text{Images}_{caption}$, while the original datasets as Images_{origin}. The recommendation models are evaluated when recommending target images from the test sets, starting from a randomly selected candidate image for the initial dialog turn. Each target image from the test sets represents a user session with the system.

3.3.3 User Simulator Comparison

To demonstrate how close the VL-Transfomer user simulator behaves in comparison to real human captions, we provide a quantitative analysis of the VL-Transformer user simulator for relative image captioning. We evaluate the relative captioning models (i.e. user simulators) on the validation set due to the fact that the test sets for relative captioning are not released in the *Fashion IQ* datasets. Table 3.2 shows the relative captioning effectiveness of the VL-Transformer model and another existing state-of-the-art baseline user simulator model, Show Tell (Guo et al., 2018; Vinyals et al., 2015), for generating natural-language critiques given a pair of images. Effectiveness is measured in terms of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (C.-Y. Lin, 2004) for measuring the overlap of n-grams (contiguous sequences of n words) between the generated text and the reference summaries and Consensus-

Table 3.2: The relative captioning effectiveness of the VL-Transformer relative captioning model compared to the Show Tell baseline model on the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets. The best results for each dataset and measure are in bold. * denotes a significant difference compared to the VL-Transformer in terms of a paired t-test ($p < 0.05$).

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



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

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

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

(a) Dresses (b) Shirts (c) Tops & Tees

Figure 3.8: Examples of different user simulators.

based Image Description Evaluation (CIDEr) (Vedantam, Lawrence Zitnick, & Parikh, 2015) for computing a consensus-based similarity score that considers both n-gram matching and the diversity of the generated captions, on the *Dresseses*, *Shirts*, and *Tops & Tees* datasets. The best overall performing results for each dataset are highlighted in bold in Table 3.2. Comparing the results in the table, we observe that, overall, the VL-Transformer model achieves significantly better performances than the LSTM-based model (i.e. Show Tell) across all metrics on all the *Fashion IQ* datasets. The superior performance of the VL-Transformer model indicates that the Transformer-based model (i.e. VL-Transformer) aligns more closely with the behaviours of real human users than the LSTM-based model (i.e. Show Tell).

Figure 3.8 presents an example of the generated natural-language critiques given a target image and a candidate image on each dataset: (a) Dresses, (b) Shirts, and (c) Tops & Tees. There are two shown ground truths[3] (i.e. GT-1 and GT-2) for each pair of images, each followed by the generated captions by Show Tell and VL-Transformer. From the generated captions on each dataset, it can be observed that the relative caption generated by the VL-Transformer model is more expressive and more close to the ground truths compared to the other model. These results demonstrate that the VL-Transformer user simulator can generate expressive natural-language feedback via relative captioning that is close to the ground truths. Therefore, the use of the VL-Transformer for relative captioning can act as a reasonable surrogate for real human users in generating natural-language feedback.

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

3.3.4 Pros & Cons of User Simulators

Here, we summarise the advantages and limitations of the user simulators with relative captioning for training and evaluating the multi-modal conversational recommender systems, as well as possible future directions to improve the rigour and usefulness of the user simulators.

Pros Compared to the real human users and the annotated entire multi-modal conversation datasets, user simulators are more flexible, cheap, and time-saving to use. In particular, the user simulators for relative captioning can generate natural-language feedback automatically that is close to real users' responses corresponding to the recommendations. Such immediate responses with the simulated users can leverage the performances of the closed-loop systems without real human users. Furthermore, the simulated users enable us to make a fair comparison of multiple multi-modal conversational recommender systems using reproducible experiments, as well as to analyse the behaviours of the users during the conversational recommendation processes, to augment the training data for conversational recommender systems, and to train the reinforcement learning algorithms (Guo et al., 2018; H. Wu et al., 2021; R. Zhang et al., 2019). In particular, the user simulator can generate virtual rewards, such as the *ranking percentile reward* (Guo et al., 2018) (i.e. the percentage of items with a rank lower than the target item among all items) and the *visual reward* (Y. Wu, Macdonald, & Ounis, 2021) (i.e. the Euclidean distance between the target and candidate image representations). Such virtual rewards are important for the success of training the RL-based multi-modal conversational recommender systems (Guo et al., 2018; R. Zhang et al., 2019).

Cons The previous formulations of user simulators for relative captioning in multi-modal conversational recommendations assumed that all the users are cold-start and only consider the desired features of a single target item in the natural-language feedback (Guo et al., 2018; H. Wu et al., 2021). Such assumptions limit the realism of user simulators compared to a real interaction process. In particular, such user simulators have demonstrated their limitations as shown in the use cases, such as repetitive feedback without memory, incorrect natural-language descriptions, limited type of feedback with only preferred features, a single preferred item, a single domain of fashion products. For instance, previous researchers have shown that the recommender systems might present repeated/violated recommendations (R. Zhang et al., 2019) to the same users. Due to the assumptions of the user simulators for relative captioning, simulated users might give repetitive feedback while forgetting about what have been expressed in the previous interaction turns. However, the real users would get annoyed and say something different, rather than repeating the same feedback. Therefore, further research on the user simulators are needed to improve the rigor and usefulness of the user simulation.

Possible Future Directions Here we elicit some possible directions for user simulators in the future to address the above limitations, such as positive/negative feedback, verbal/non-verbal relevance feedback, user initiatives, memory, personalisation, and multi-interests:

- *Positive/Negative Feedback.* The user simulators only consider the users' preferred features in the target items as positive feedback, other than the undesired features in the recommendations as negative feedback. It is necessary to capture the users' negative feedback that contains the explicit dislikes of the users and to avoid recommending items with those undesired features in the future recommendation processes.
- *Verbal and Non-Verbal Relevance Feedback.* The user simulators only consider the users' verbal relevance feedback (i.e. critiques) by describing the preferred features. The user simulators should be able to express their preferences explicitly with verbal feedback (such as natural-language critiques) or implicitly with non-verbal feedback (such as likes/dislikes).
- *User Initiative.* The user simulators take a passive-initiative to respond to the recommendations. However, the users are usually actively ask questions when they are unclear about the recommendations rather than giving critiques only.
- *Memory.* The user simulator should be capable of memorising the previous natural-language feedback and avoiding repeating the same expressions.
- *Personalisation.* Both the Show-Tell and VL-Transformer user simulators consider the simulated users as cold-start users. However, a personalised initial recommendation based on users' interaction histories might leverage the users' experience during the conversational recommendation processes by considering the users' long-term preferences and short-term interests.
- *Multi-Interests.* The user simulators consider the simulated users as cold-start users with only a single target item. However, real human users usually have different needs by keeping many different target items in their minds (such as an outfit of fashion products).

Therefore, we argue that the user simulators for relative captioning can act as a reasonable proxy of real human users by simulating the users' natural-language feedback for training and evaluating the cross-modal interactive recommender systems. We first presented implementations of the existing user simulators with an encoder-decoder structure to demonstrate that the user simulator for relative captioning are suited for the multi-modal conversational recommendation task in fashion. In addition, the use cases for both relative captioning with different user simulators, i.e. Show-Tell and VL-Transformer, on the *Fashion IQ* datasets was presented to demonstrate the effectiveness and appropriateness of the user simulators. For future work, it is worth exploring the simulation of users' natural-language feedback considering positive/negative feedback, verbal and non-verbal relevance feedback, user initiatives, memory, personalisation, and multi-interests in the multi-modal conversational recommendation processes. In the next section, we describe the opportunities in the multi-modal conversational recommendation scenarios.

3.4 Opportunities Within the Framework

After describing the functionality of each component of our multi-modal conversational recommendation framework, we now discuss the main opportunities within the framework. Recall that in Section 3.1, we have identified five limitations in the literature: partial observations, multi-modal sequence dependency, coupling of representation learning and policy optimisation, personalisation, and the realism of simulated conversations with positive/negative feedback. In this section, we seek for opportunities from either a recommender system side or a user side for addressing the above mentioned limitations. In Section 3.4.1, we discuss the architecture of recommendation models and the formulation of optimising the recommendation model. In Section 3.4.2, we describe the opportunities of different user behaviours.

3.4.1 Recommender System Side

In this thesis, one of the main targets of our proposed framework is to effectively capture the users' preferences cross multiple interaction turns from both the users' natural-language feedback and the systems' visual recommendations. The effectiveness of the recommendation models can be improved by either modifying the models' architectures, or applying more advanced optimisation approaches.

Model Architectures The architecture of the recommendation model can be further divided into three parts: multi-modal encoding, multi-modal composition and state tracking.

- **Multi-modal encoding**: Due to the multi-modal nature of the task, it is necessary to have a good understanding of the text and image contents. In particular, in Chapter 4, we first follow MBPI (Guo et al., 2018) to encode the textual sentence with a one-hot encoding and a 1D convolutional layer (1D-CNN), and encode the image based on the ImageNet pre-trained ResNet101 (K. He et al., 2016). Then, we improve the textual representations with a pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019b) in Chapters 5 & 8. Further, in Chapters 6 & 7, we leverage a pre-trained vision and language model, called CLIP (Radford et al., 2021), for both image encoding and text encoding. Different from ResNet and BERT for image and text encoding, CLIP can provide unified representation vectors for each modality (i.e./ both text and images) with the same dimensionality.

- **Multi-modal composition**: Both images of products and users' textual feedback can contain the users' preferred attributes. To this end, the encoded text and image representations are usually fused with a concatenatation operation and/or a followed linear layer (in Chapter 4). We also adopted an advanced representative composition network (Text Image

Residual Gating (TIRG) (Vo et al., 2019)) to better combine image and text representaions in Chapter 6.

- **State tracking**: Multi-modal conversational recommendation involves multiple interaction turns. To this end, it is necessary to capture the evolving of the users' preferences over time. The existing conversational recommender systems usually directly adopt the GRU/LSTM or Transformer models as a state tracker. In Chapter 5, we leverage a RNN-enhanced Transformer structure to better track the dialog states.

Optimisation Approaches As mentioned in Section 2.4, there are various recent reinforcement learning approaches that can be adapted to better formulate and optimise the multi-modal conversational recommendation task.

- **Self-Supervised Reinforcement Learning (SSRL)**: In Chapter 4, we formulate the multi-modal conversational recommendation task with the Supervised Q-learning (SQN) framework (Xin et al., 2020) by taking Q-learning as a regulariser that can examine the quality of the estimated states with the users' historical feedback and improve the quality of the following recommendations.
- **Goal-Oriented Reinforcement Learning (GORL)**: In Chapter 6, we formulate the multi-modal conversational recommendation task with goal-oriented reinforcement learning to effectively optimise the recommendation policy via goal-oriented rewards for pursing the textual and visual goals.
- **Hierarchical Reinforcement Learning (HRL)**: In Chapter 7, we propose a personalised multi-modal conversational recommendation model based on HRL with the Options framework to more effectively incorporate the users' preferences from both their past and real-time interactions.

3.4.2 User Side

In addition to the improvements from the recommender system side, user simulators with more various behaviours can make the multi-modal conversational recommendation task more realistic. In this thesis, we only focus on three different types of user behaviours in addition to the default positive natural-language feedback, such as positive & negative feedback, verbal & non-verbal relevance feedback, and past preferences & current needs. We leave the other types of user behaviours, such as memory and multiple interests (Section 3.3), as interesting future work.

- **Verbal & Non-Verbal Relevance Feedback:** In Chapter 6, we leverage both verbal and non-verbal relevance feedback to further improve the performance of the multi-modal conversational recommendation model.
- **Past Preferences & Current Needs**: In Chapter 7, we explore how the more effectively incorporate the users' preferences from both their past and real-time interactions.
- **Positive & Negative Feedback**: In Chapter 8, we investigate the impact of positive and negative natural-language feedback on the peformance of the multi-modal conversational recommendation models.

Indeed, different user behaviours can also change the architecture of the recommendation models, as well as their optimisation approaches. To this end, it is worthy but challenging to consider more types of user behaviours and make the multi-modal conversational recommendation task more realistic. We believe that our thesis can enlighten the exploration in multi-modal conversational recommendations.

3.5 Conclusions

In this chapter, we first investigated the multi-modal conversational recommendation techniques in the literature in Section 3.1. In particular, we described two routes for developing multi-modal conversational recommender systems, by either enhancing uni-modal systems with multi-modal information or enhancing the image retrieval models with multi-turn interactions. We also illustrated the limitations of the existing multi-modal conversational recommendation models and distinguish our main contributions from the existing work. Then, in Section 3.2, we introduced the framework of the multi-modal conversational recommendation task with cold-start or warm-start users, as well as how to measure the success of the tasks. Moreover, in Section 3.3, we illustrated the users simulators for training and evaluating the multi-modal conversational recommender systems, including the formulation of user simulators, datasets, evaluation metrics, and pros & cons. Finally, we discussed the main opportunities within the framework from both the recommender system side and the user side. Next, starting from Chapter 4, we discuss the studies for addressing the challenges described in Section 3.1 and taking the opportunities described in Section 3.4.

Chapter 4

Partial Observability in Natural-Language Feedback

In our thesis statement (as stated in Section 1.3), we postulated that we can effectively incorporate the users' preferences over time, in the form of partial observations, by modelling the multi-modal conversational recommendation process with (self-)supervised Q-learning. Therefore, in this chapter, we propose a novel multi-modal conversational recommendation model, called the Estimator-Generator-Evaluator (EGE) model based on its three distinctive functional components, which apply partially observable reinforcement learning (i.e. Q-learning for POMDP) for multi-modal conversational recommendations. This chapter is mainly based on our work (Y. Wu et al., 2021) "Partially Observable Reinforcement Learning for Dialog-based Interactive Recommendation" published in the proceedings of the 15th ACM Conference on Recommender Systems (RecSys 2021)[1]. Note that we also refer to multi-modal conversational recommendation as dialog-based interactive recommendation or multi-modal interactive recommendation throughout this thesis.

A dialog-based interactive recommendation task is where users can express natural-language feedback when interacting with the recommender system. However, the users' feedback, which takes the form of natural-language critiques about the displayed recommendation at each iteration, can only allow the recommender system to obtain a partial portrayal of the users' preferences, as argued in Section 3.1. Indeed, such partial observations of the users' preferences from their natural-language feedback make it challenging to correctly track the users' preferences over time, which can result in poor recommendation performances and a less effective satisfaction of the users' information needs when in presence of limited iterations. Reinforcement learning, in the form of a partially observable Markov decision process (POMDP, see Section 2.4), can simulate the interactions between a partially observable environment (i.e. a user) and an agent (i.e. a recommender system). To alleviate such a partial observation issue, we propose a novel

[1] DOI: https://doi.org/10.1145/3460231.3474256

dialog-based recommendation model, the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP, to effectively incorporate the users' preferences over time. Specifically, we leverage an Estimator to track and estimate users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations, and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. Following previous work (see Section 3.1), we train our EGE model by using a user simulator (described in Section 3.3) which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Thorough and extensive experiments conducted on two recommendation datasets – addressing images of fashion products (namely Dresses and Shoes, introduced in Section 3.3) – demonstrate that our proposed EGE model yields significant improvements in comparison to the existing state-of-the-art baseline models. The results conform with our thesis statement with **Research Topic 1** in Section 1.3.

4.1 Motivations

Recently, interactive recommender systems (IRS) have received much attention due to their flexible recommendation strategies and their natural multi-step decision-making processes. A typical interactive recommender system continuously recommends items to users and receives various types of users' feedback, such as clicks, ratings, or textual replies (see Sections 2.1 & 2.3). In particular, natural-language feedback allows an interactive recommender system to obtain richer information relating to the users' current preferences, thereby leading to a more suitable recommendation compared to clickthrough data and ratings. Figure 4.1 shows an example of interactive recommendation based on natural-language feedback, i.e. dialog-based interactive recommendations. In this use case, the user gives natural-language critiques about the system's recommendation at each interaction turn and aims to quickly find the target item, while the system recommends the top-1 item according to the user's natural-language feedback.

Such an interactive recommendation task has been formulated and modelled using reinforcement learning (RL) approaches (see Section 2.4 & 3.1). In the reinforcement learning framework, the interactive recommendation task is usually formulated as a Markov decision process (MDP, see Section 2.4.1) with an assumption that the environment's states (i.e. the users' preferences) are *fully observable*. Such RL-based interactive recommender systems have demonstrated their benefits in fitting the users' dynamic preferences and maximising the expected long-term cumulative rewards from users when achieving the optimal strategies. For instance, the Supervised Q-learning (SQN) framework (Xin et al., 2020) (i.e. a joint learning framework with both a supervised learning layer and a Q-learning layer) was shown to outperform neural recommendation models using supervised learning, such as GRU4Rec (Hidasi et al., 2016), Caser (Tang & Wang, 2018) and SASRec (Kang & McAuley, 2018), by taking the Q-learning layer as a regulariser to introduce reward-driven properties (such as long-term user

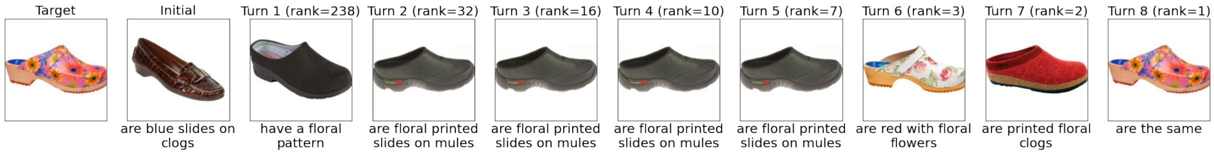


Figure 4.1: An example of dialog-based interactive recommendations. The first image is the target item desired by the user (labeled with "Target"), while the second image (labeled with "Initial") is the initial recommendation proposed by the system randomly. Then, the user gives natural-language critiques about the recommendation at each turn, while the recommender system updates the ranking list and recommends the top-1 item according to the user's comments. The rank of the target item is also presented above the images at each turn. When the target item is recommended, the user will give a comment as "are the same" and the rank is 1.

engagement (Zou et al., 2019)) to the recommendation process.

Despite the expressiveness of natural-language feedback in dialog-based interactive recommendations, the users' feedback can only allow the recommender system to obtain a partial portrayal of the users' preferences. For instance, Figure 4.1 shows an example of a dialog-based interactive recommendation process between the user (simulator, see Section 3.3) and the system generated by a RL-based interactive recommender system (called Model-based Policy Improvement (MBPI)) (Guo et al., 2018). Each natural-language comment in terms of the current recommendation only contains partial visual features of the target item, such as "blue slides" at the initial turn and "red with floral flowers" at the 6th turn. Such partial observations of the users' preferences from their natural-language feedback can drive the recommender system towards a degenerate preference estimation that ignores certain features in the historical observations (Gangwani et al., 2020), i.e. historical natural-language feedback and historical recommendations. For instance, in Figure 4.1, "red" clogs are recommended due to the last comment "red with floral flowers" at the 7th interaction turn, while the single "red" colour in the recommended image does not address the initial user comment "blue slides on clogs" and the other comments with "floral". In addition to this so-called violated recommendation issue, *repeated recommendations* can also be observed in the example IRS in Figure 4.1. Although the rank of the actual target shoe in the ranking list indicated above each suggested image is increasing from the 2nd turn to the 5th turn, the top-1 recommendation remains the same and hence receives identical feedback from the user (simulator). Both these violated and repeated recommendations can hurt the users' experience thereby increasing their disappointment in the interaction processes with the recommender system. Indeed, such partial observations of the users' preferences from their natural-language feedback make it challenging to correctly track the users' preferences over time, which can result in a poor performance of the recommender system in satisfying the users' information needs when in the presence of limited iterations (as can be observed from the literature (Guo et al., 2018)). Although R. Zhang et al. (2019) proposed a reward-constrained recommendation (RCR) model with constraint-augmented reinforcement learning that can effectively mitigate the aforementioned violation issue, the utility of their RCR model was limited by the issue that extra rounds of resampling and violated recommendation detection are needed when the previous samplings are detected as violated recommendations.

In this chapter, we formulate the dialog-based interactive recommendation task as a partially observable Markov decision process (POMDP, see Section 2.4.1) to simulate the interactions between a partially observable environment (i.e. a user) and an agent (i.e. a recommender system). To correctly estimate the users' preferences from such partially observable situations, we extend the SQN framework (Xin et al., 2020) from a MDP to a POMDP and judge/optimise the quality of the estimated users' preferences with the Q-learning layer (also called an Evaluator). To this end, we propose a novel dialog-based interactive recommendation model, called the Estimator-Generator-Evaluator (EGE) model named after its three distinctive functional components, which apply partially observable reinforcement learning (i.e. Q-learning for POMDP) for dialog-based interactive recommendation to effectively incorporate the users' preferences over time. Specifically, we leverage an Estimator to track and estimate the users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations, and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. To mitigate the impact of repeated recommendations, a post-filter is adopted to remove the repeated recommended items from the ranking list based on the recommendation history. Following previous work (Guo et al., 2018; R. Zhang et al., 2019), we train our EGE model by using a user simulator (see Section 3.3), which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Thorough and extensive experiments conducted on two recommendation datasets – addressing images of fashion products (namely dresses and shoes) – demonstrate that our proposed EGE model yields significant improvements in comparison to the existing state-of-the-art baseline models (i.e. a sequential recommendation model (denoted iGRU) with supervised learning and the Model-based Policy Improvement (MBPI) model).

The main contributions of this chapter are summarised as follows:

- We propose a novel dialog-based interactive recommendation model, the Estimator-Generator-Evaluator (EGE) model, which formulates the dialog-based interactive recommendation task as a partially observable Markov decision process (POMDP) to address the partial observations issue in the users' feedback. Our proposed EGE model differs from the existing MBPI (Guo et al., 2018) and RCR (R. Zhang et al., 2019) models as follows: EGE judges and optimises the quality of the estimated user preferences with a Q-learning layer (i.e. Evaluator) for POMDP based on the users' history feedback, while the MBPI model (with only the Estimator and Generator components) is not able to do so and the RCR model needs to repeatedly sample recommendations and detect violations with extra well-categorised visual attributes of items.

- The EGE model extends the SQN (Xin et al., 2020) framework from a MDP to a POMDP and is trained with a combination of a supervised learning classification loss and a Q-learning prediction loss.

- Extensive empirical evaluations are performed on the dialog-based interactive recommendation task, demonstrating significant improvements over existing state-of-the-art approaches while providing directions for future work.

4.2 Methodology

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

4.2.1 Problem Statement

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

A recommender agent acts according to its policy $\pi(a_t|o_{\leq t}, a_{<t})$, which returns the probability of taking action a_t at time t, and where $o_{\leq t} = (o_1, ..., o_t)$ and $a_{<t} = (a_0, ..., a_{t-1})$ are the observation and action histories, respectively. The recommender agent's goal is to learn a policy π that maximises the expected future return $J = \mathbb{E}_{p(\tau)}[\sum_{t=1}^{T} \gamma^{t-1} r_t]$ over trajectories $\tau = (s_0, a_0, ..., a_{T-1}, s_T)$ induced by its policy. In general, a dialog-based interactive recommender system via a POMDP must condition its actions on the entire history $h_t = (o_{\leq t}, a_{<t}) \in \mathscr{H}$. The users' preferences are represented by a target image representation $x_{+,0}^{img}$. The inputs of a recommender agent at time t are the previously recommended items (i.e. action history $a_{<t} = (a_0, ..., a_{t-1})$) and the corresponding users' feedback (i.e. the observation $o_{\leq t} = (o_1, ..., o_t)$).

4.2.2 The Model Architecture

In dialog-based interactive recommendations, a recommender agent recommends an item (in particular, an image) and a user provides natural-language feedback. Figure 4.2 shows our proposed end-to-end Estimator-Generator-Evaluator (EGE) model with partially observable reinforcement learning for dialog-based interactive recommendations to effectively incorporate the user's preferences over time. The user views the recommended item (a single item at each interaction) and gives natural-language feedback by describing their desired features that the current recommended item lacks. The system then incorporates the user's natural-language feedback and recommends (ideally) more-suitable items, until the desired item is found.

Estimator The goal of the Estimator is to track and estimate the user's preferences (i.e. states) from both the user's natural-language feedback and the latest recommended visual item. The Estimator consists of a text encoder, an image encoder and a gated recurrent unit (GRU) (Chung et al., 2014) as in (Guo et al., 2018). In particular, the text encoder extracts the textual sentence representations of the user's preferences from the current user's natural-language feedback. In the textual sentence representations, each word is represented by a one-hot vector. Similarly, the image encoder extracts image feature representations based on the ImageNet pre-trained ResNet101 model (K. He et al., 2016) as in (Guo et al., 2018). Then, both the image feature representations and the textual representations are concatenated as input to a following linear mapping (i.e. a multilayer perceptron (MLP)) and a GRU to obtain the estimated user's preferences. Given a candidate image a_{t-1} and a user's corresponding natural-language feedback o_t at the t-th dialog turn, the encoded textual representation is denoted by x_t^{txt} and the encoded image representation is denoted by:

$$x_{t-1}^{img} = ResNet(a_{t-1}). \tag{4.1}$$

The estimated user's preferences can be achieved with:

$$s_t = Linear(GRU(Linear([x_t^{txt}, x_{t-1}^{img}]), s_{t-1})). \tag{4.2}$$

The GRU component of the Estimator allows our EGE model to sequentially aggregate the partially observable information from the user's natural-language feedback to the estimated preferences.

Generator The goal of the Generator is to recommend a candidate item for the next action according to the estimated state. Considering the large amount of candidate images in the image database, all images are projected into the feature space (ResNet). If K items are recommended at each time t, we select the top K closest images to the estimated state s_t under the Euclidean

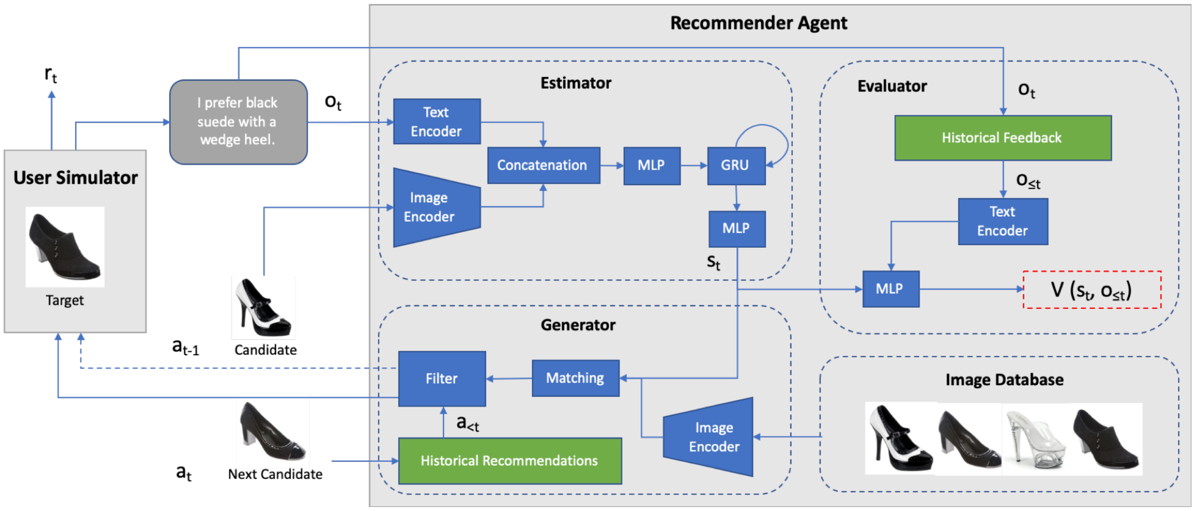


Figure 4.2: The proposed Estimator-Generator-Evaluator (EGE) model for dialog-based interactive recommendations.

distance in the image feature (ResNet) space:

$$a_{t,\leq K} \sim KNNs(s_t), \tag{4.3}$$

where $KNNs()$ is a softmax distribution over the top-K nearest neighbours of s_t and $a_{t,\leq K} = (a_{t,1},...,a_{t,K})$. Furthermore, based on the interaction history $h_t = (o_{\leq t}, a_{<t})$, a post-filter is adopted to remove any candidate items from the ranking list that have previously occurred in the recommendation history $a_{<t}$.

Evaluator The Evaluator is proposed to judge the quality of the estimated state s_t at time t based on Q-learning. It performs the judgement process with the user's historical natural-language feedback $o_{\leq t} = (o_1,...,o_t)$ to regularise the Estimator. Given an estimated state s_t and the textual features $x^{txt}_{\leq t} = (x^{txt}_1,...,x^{txt}_t)$ from the user's historical natural-language feedback, the state values in terms of the user's historical natural-language feedback are computed with:

$$V(s_t, o_i) = Linear(Linear(s_t, x^{txt}_i)), \tag{4.4}$$

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

To summarise, in the EGE model architecture, we maintain the GRU for the state estimation in the Estimator and the KNNs for the candidate matching in the Generator as in the state-of-the-art RL-based approach (Guo et al., 2018), while we propose a Q-learning layer with the historical feedback as an Evaluator to optimise the quality of the estimated state.

4.2.3 The Learning Algorithm

In this work, we adopt a multi-task learning (Goodfellow, Bengio, & Courville, 2016) approach for POMDP to optimise the networks with a combination of a supervised learning classification loss and a Q-learning prediction loss.

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

$$L_s = -\log\left(\frac{e^{y_0}}{e^{y_0} + \sum_{j=1}^{J} e^{y_j}}\right) \tag{4.5}$$

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

$$L_q = \begin{cases} 0.5(V(s_t, o_{\leq t}) - (r_t + \gamma V(s_{t+1}, o_{\leq t+1})))^2, & \text{if } error < 1 \\ |V(s_t, o_{\leq t}) - (r_t + \gamma V(s_{t+1}, o_{\leq t+1}))| - 0.5, & \text{otherwise} \end{cases} \tag{4.6}$$

It is desired that the visual appearance of the recommended item becomes more similar to that of the desired item with increasing user interactions. Thus, at time t, given the recommended item a_t and the desired item representation $x_{+,0}^{img}$, we want to minimise the Euclidean distance. That is, we maximise the following visual reward: $r_t^{vis} = -||ResNet(a_t) - x_{+,0}^{img}||_2$. In addition, we expect that the desired item will be placed at higher ranks with more user interactions. Thus, we also model the *ranking percentile* (Guo et al., 2018) (i.e. the percentage of items with a rank lower than the target item among all items) as a reward r_t^{per} in terms of ranking. We define the reward r_t at time t as $r_t = \alpha r_t^{vis} + (1-\alpha) r_t^{per}$, where $\alpha \in [0,1]$ is a reward weighting factor. A higher value of the reward weighting factor places more emphases on the visual rewawrd.

We jointly train the supervised loss and the RL loss by taking the latter one as a regulariser to introduce reward-driven properties to the recommendation process, in a similar manner to Xin et al. (2020):

$$L_{EGE} = L_s + L_q \tag{4.7}$$

To train our proposed EGE model, we adopt a user simulator (see Section 3.3) as a surrogate for real human users in the training processes. Further details about the used user simulator are provided in Section 4.3.3. When we start to train the proposed framework, the network parameters are randomly initialised. To facilitate an efficient exploration during the following reinforcement learning process, we first pre-train the model with a triplet loss objective, L_{tri},

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

similar to (Guo et al., 2018):

$$L_{tri} = max(0, ||s_t - x_+^{img}||_2 - ||s_t - x_-^{img}||_2 + m) \tag{4.8}$$

where x_+^{img} and x_-^{img} are respectively the representations of the target image and of a randomly sampled image, m is a constant for the margin and $||.||_2$ denotes the L^2-norm. Both the embedded representations of state s_t and item x^{img} are encoded as vectors within a shared space. $||s_t - x^{img}||_2$ measures the distance between the estimated state s_t (i.e. the user's estimated preferences) and an item representation (such as a target item x_+^{img} or a randomly sampled item x_-^{img}). Indeed, the rank of the target image can be improved compared to a random initialisation after the appropriate initial supervised learning process with L_{tri}. Based on the pre-trained model obtained with L_{tri}, the joint loss objective L_{EGE} can further ensure proximity between the target and candidate image representations (L_s), as well as maximise the expected future rewards (L_q), while applying smaller learning rates, resulting in better recommendation performances.

4.3 Experimental Setup

In this section, we evaluate the effectiveness of our proposed EGE model for dialog-based interactive recommendations in comparison to the existing approaches from the literature. Figure 3.5 in Section 3.2.2 shows an example of a recommendation scenario to illustrate how the users can obtain their target items through interaction with the recommender system in the dialog-based interactive recommendation scenario. In particular, the recommender system ranks items based on the ranking scores (i.e. the similarities between the estimated preferences and all items), while the user gives feedback on the single top-ranked recommendation presented to them. Hence, effectiveness can be measured by the percentage of user sessions for which the target item is presented at the top rank by interaction turn M. Furthermore, it is possible that the user may view more of the ranking of items at each interaction turn, down to rank N. Therefore, we define success in the dialog-based interactive recommendation task as higher values in top-heavy metrics such as NDCG@N with a truncation at rank N calculated at the M-th turn, or Success Rate (SR) at the M-th turn. In our experiments, we address three research questions, which are concerned with ascertaining how the Q-learning for POMDP (i.e. the Evaluator) can help the GRU (i.e. the Estimator) to better incorporate the users' accurate preferences over time with the partial observable preferences from the users' natural-language feedback, so as to make better recommendations with KNNs (i.e. the Generator). In particular, our three research questions relate to the Q-learning for POMDP, the historical information and the rewards in the EGE model – namely, how useful the Q-learning for POMDP is, how much historical information is required, and how the rewards are applied:

- RQ4.1: Can our proposed EGE model with Q-learning for POMDP outperform the existing state-of-the-art baseline models in the visually-grounded dialog-based interactive recommendation task?

• RQ4.2: What are the impacts of the historical information in the EGE model on its performance, such as the historical natural-language feedback and the historical recommendations?

• RQ4.3: What are the impacts of the reward-related hyper-parameters of the EGE model on its performance, such as the reward discount factor γ and the reward weighting factor α?

4.3.1 Datasets & Measures

Datasets We perform our experiments on two datasets, namely the *Shoes* and *Dresses* datasets (introduced in Section 3.3.2). In particular, the *Shoes* dataset has previously been used by (Guo et al., 2018) for a dialog-based interactive recommendation task, and we replicate their setup (relative captioner, user simulator, etc.). On both datasets, we apply the same training and testing data split for all recommendation models. In this chapter, we leverage *Shoes* and *Dresses* (see Section 3.3.2) for training and testing the recommender systems. Note that we leverage *Shoes* with the Images$_{origin}$ version that provides $10,000$ images for training the recommender systems, and $4,658$ images for testing and *Dresses* with the Images$_{caption}$ version that extracts $7,182$ unique images from the $5,478$ relevative captioning triplets for training the recommender systems, and $2,454$ unique images from the $1,869$ relevative captioning triplets for testing (see Section 3.3.2). The relevative captioning triplets on both datasets are used for training and testing a user simulator on each dataset. The recommendation models are evaluated when recommending target images from the test sets, starting from a randomly selected candidate image for the initial dialog turn. Each target image from the test sets represents a user session with the system. Moreover, following Guo et al. (2018), as a user simulator, we adopt a *relative captioner* to simulate the user in generating natural-language feedback (described in Section 4.3.3), which has been shown to mimic an actual user behaviour/feedback (Guo et al., 2018).

Metrics The performances of the dialog-based interactive recommender systems are evaluated with metrics including Normalised Discounted Cumulative Gain (i.e. NDCG@N truncated at rank $N = \{5, 10\}$ calculated at the M-th interaction, see Section 2.1.3), Mean Reciprocal Rank (i.e. MRR@N truncated at rank $N = 10$ at the M-th interaction, see Section 2.1.3) and Success Rate (SR, see Section 3.2.2) at the M-th interaction. In particular, SR is the percentage of users for which the target image was retrieved within M turns among all the users with top-1 recommendation. We use all the evaluation metrics (i.e. NDCG@5, NDCG@10, MRR@10 and SR) at the 10th interaction turn for significance testing.

4.3.2 Baselines

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

• The sequential recommendation model (Guo et al., 2018) in a supervised-learning setup (which we denote as iGRU, where “i” stands for “interactive”) is an approach where the recommender agent (with only a GRU and KNNs) is trained with a triplet loss (Guo et al., 2018) to maximise the short-term rewards. The iGRU model is close to the well-established GRU4Rec sequential recommendation model (Hidasi et al., 2016) but using the users’ natural-language feedback and the previous recommended images as the input of the model with an online setup, instead of the logged clickthough data with an offline setup.
• The Model-Based Policy Improvement (MBPI) (Guo et al., 2018)) model is a RL-based approach where the recommender agent (with only a GRU and KNNs) is pre-trained with a triplet loss, and then further trained with a cross entropy loss. In the second training stage, the MBPI model is optimised by maximising the cumulative future rewards given a known environment (i.e. the user simulator). In particular, the MBPI model explores all possible recommendation trajectories in the future interaction turns with the help of the given user simulator and recommends the items with the maximum cumulative future rewards at each turn during this training process.

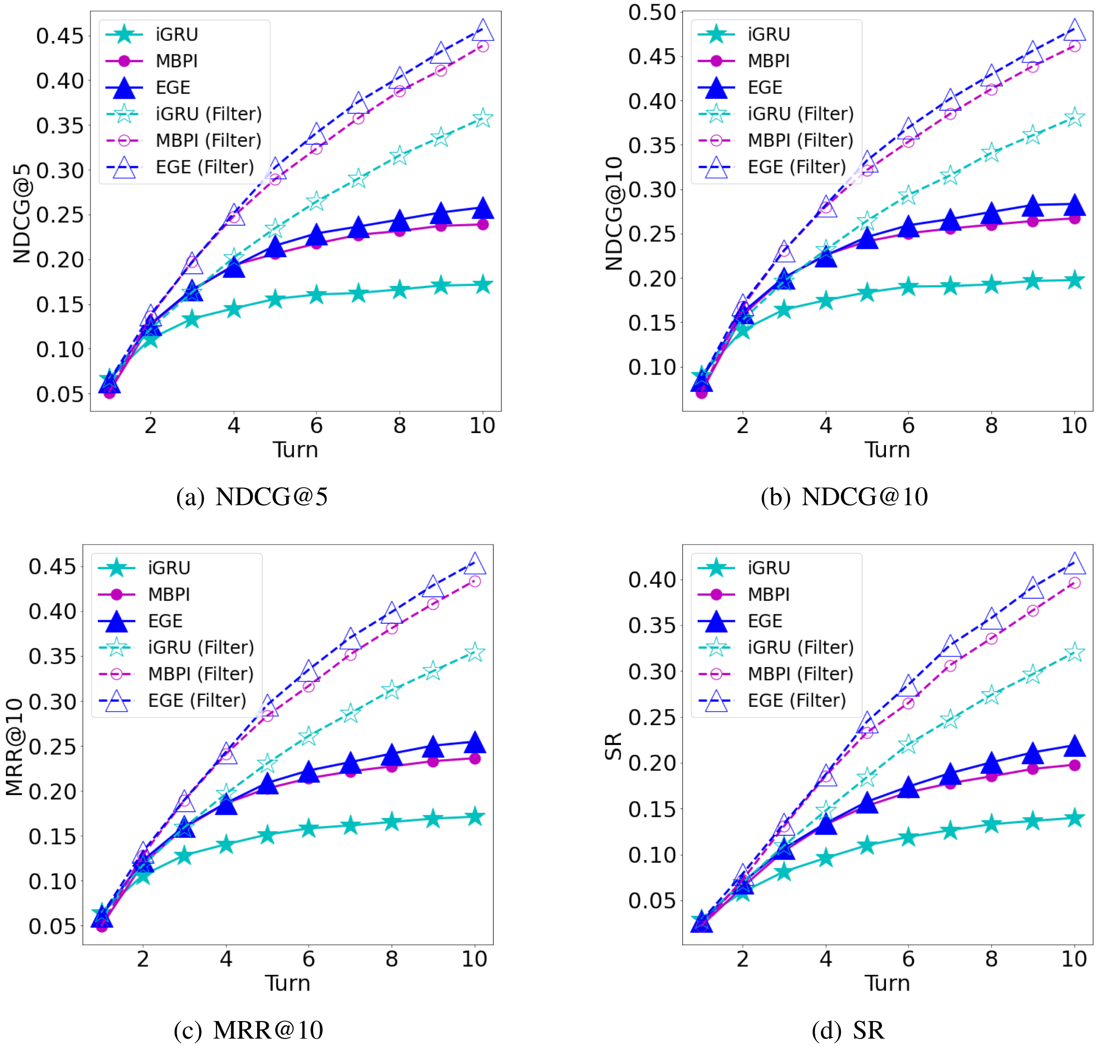
These two baseline models are two existing representative formulations of the dialog-based interactive recommendation task for top-1 recommendation, which are formulated as a sequential modelling problem and a Markov decision process (MDP), respectively. Although there are a few other models with different formulations for the dialog-based interactive recommendation task – such as RCR (R. Zhang et al., 2019) which is formulated as a constrained Markov decision process (CMDP) (Altman, 1999), the augmented cascading bandit (ACB) (Yu, Shen, & Jin, 2019) or the sleeping pairwise ranking bandit (SPRB) (Yu et al., 2020), which are formulated as a multi-armed bandit (MAB) problem (Sutton & Barto, 2018) – these models are not comparable with our scenario due to either requiring extra well-categorised visual attributes of items (RCR) or taking a *category* of the fashion products as the targets (ACB & SPRB).

4.3.3 Experimental Settings

User Simulator To tackle the challenge of training an interactive recommender system online, we adopt a user simulator based on relative captioning (Rennie et al., 2017) as in (Guo et al., 2018), which acts as a surrogate for real human users. The user simulator can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images. Such a natural-language feedback generation process with the user simulator is very similar to a scenario of a shopping conversation session between a shopping assistant and a customer. A user simulator with the *Shoes* dataset was intensively and carefully trained by (Guo et al., 2018) through crowdsourcing relative expressions about the visual differences of the image pairs and manually removing erroneous annotations. Furthermore, the pre-trained user simulator has previously been thoroughly evaluated via both a quantitative evaluation and a user study, thereby serving as a reasonable proxy for real users in our work. Following H. Wu et al. (2021), we also train a user simulator for the *Dresses* dataset. The user simulator for the *Dresses* dataset is selected with the best prediction performance of the relative captioning task on the caption testing split. The pre-trained user simulators are used for both the training and evaluation of the interactive recommendation models.

Setup for Training We first train our proposed EGE model with both user simulators on the *Shoes* and *Dresses* datasets, separately. The network parameters are randomly initialised. Following (Guo et al., 2018), we adopt a two-stage training process to facilitate the efficient exploration during the training with a joint loss L_{EGE}. At the initial stage (i.e. training with a triplet loss objective L_{tri}) and the second stage (i.e. training with a joint loss L_{EGE}) of training, we use Adam (Kingma & Ba, 2014) as the optimiser on both datasets with an initial learning rate 10^{-3} and 10^{-5} (Guo et al., 2018; R. Zhang et al., 2019), respectively. The embedding dimensionality of the feature space is set to 256 and the batch size is 128, following the setting in (Guo et al., 2018). For each batch, we train our model with 10 turns. We consider the top-11 nearest neighbours (considering an initial random item and 10 items during the 10 interactions) for removing the previously recommended items from the ranking list at each interaction with a post-filter, and we pick the top-1 from the post-filtered nearest neighbour list. The number of negative samples (i.e. J) is set at 5, which is considered as a reasonable number for negative sampling, regardless of the dataset size (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). For our proposed EGE model, if not mentioned otherwise, the reward discount factor γ is set to 0.9 while the reward weighting factor α is set to 0.5 due to the EGE model's general good performances with $\gamma, \alpha \in [0,1]$ on both datasets (as shown in Section 4.4.3).

Setup for Evaluation We consider the top-1 nearest neighbour (i.e. $K = 1$) as a recommendation at each interaction turn with or without a post-filter for testing. In particular, when a post-filter is applied, we pick the top-1 item from the post-filtered nearest neighbour list. For the evaluation metrics, we denote the interaction turn $M \in [1,10]$. In particular, we mainly compare the performances of the tested models at the 10th turn (i.e. $M = 10$) with significance tests, which is the maximum interaction number in our study. This is smaller than the values adopted in (Xu et al., 2021; R. Zhang et al., 2019) and is more reasonable in the shopping scenario because the users are more likely to be disappointed if they do not find their desired items after that many turns. If a user obtains the target item in less than 10 turns, we consider the ranking metrics (i.e. NDCG@5, NDCG@10 and MRR@10) for that user to be equal to one for all turns thereafter.



(a) NDCG@5

(b) NDCG@10

(c) MRR@10

(d) SR

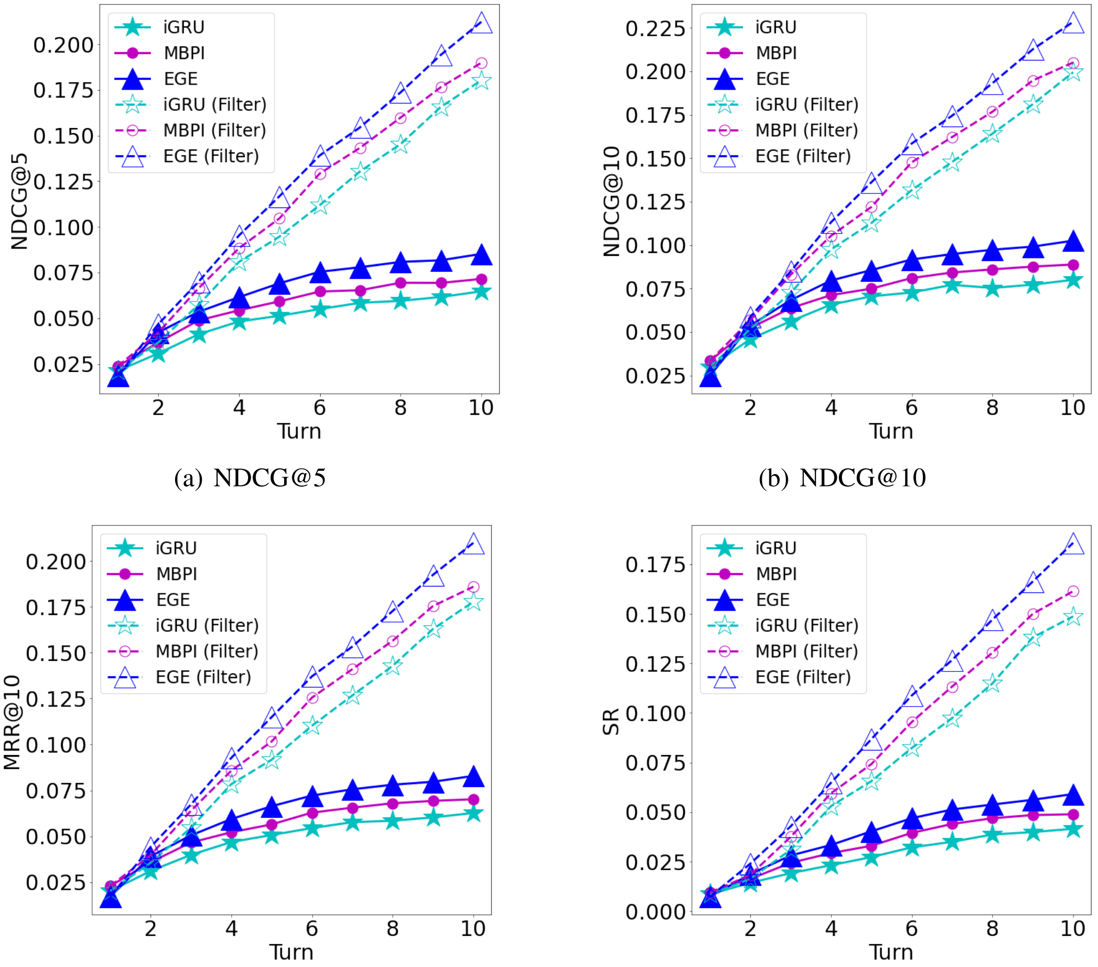
Figure 4.3: Recommendation effectiveness at various interaction turns with top-1 recommendation on the *Shoes* dataset.

4.4 Experimental Results

In this section, we analyse the experimental results with respect to the three research questions stated in Section 4.3, concerning the recommendation effectiveness of our proposed EGE model (Section 4.4.1), impact of the historical information including the historical natural-language feedback and historical recommendations (Section 4.4.2), and the impact of hyper-parameters related to the rewards (Section 4.4.3). We also demonstrate a use case from the logged experimental results to consolidate our findings (Section 4.4.4).

4.4.1 EGE vs. Baselines (RQ4.1)

Figures 4.3 and 4.4 show the recommendation effectiveness of our proposed EGE model and



(a) NDCG@5

(b) NDCG@10

(c) MRR@10

(d) SR

Figure 4.4: Recommendation effectiveness at various interaction turns with top-1 recommendation on *Dresses*.

the existing state-of-the-art baseline models for top-1 recommendation in terms of NDCG@5 (Figure 4.3 (a) and Figure 4.4 (a)), NDCG@10 (Figure 4.3 (b) and Figure 4.4 (b)), MRR@10 (Figure 4.3 (c) and Figure 4.4 (c)) and Success Rate (SR) (Figure 4.3 (d) and Figure 4.4 (d)), while varying the number of interaction turns on the *Shoes* and *Dresses* datasets, respectively. The solid lines show the models' performances without a post-filter (which prevents the already recommended items from being recommended), while the dashed lines show performances when a post-filter is applied. When a post-filter is applied, the model is labeled with "(Filter)". Comparing the results in Figure 4.3 and Figure 4.4, we observe that our proposed EGE model generally achieves a better overall performance in terms of NDCG@5, NDCG@10, MRR@10 and SR at various interaction turns (except for the initial turn) without/with a post-filter, respectively. In the initial interaction turn, the performance of our proposed EGE model is marginally lower than the iGRU model and marginally higher than the MBPI model on the *Shoes* dataset, while it is marginally lower than the other two on the *Dresses* dataset. As the number of interaction turns increases (≥ 2), the differences between the effectiveness of EGE and iGRU/MBPI on all metrics also increase. The better performance of EGE compared to iGRU can be attributed to the fact that our RL-based EGE model is optimised to maximise the long-term rewards with a Q-learning layer in the Evaluator, while the supervised learning approach (i.e. iGRU) aims to maximise the instant reward (i.e. r_t). Furthermore, by considering the historical information with Q-learning for POMDP, our proposed EGE model can also outperform the MBPI model with a better recommendation effectiveness, thereby mitigating the partial observation issue.

To quantify the improvements of our proposed EGE model compared to the other two baseline models, we measure their performances at the 10th interaction turn with top-1 recommendation. Table 4.1 shows the obtained recommendation performances of the models on the user simulator with a test set at the 10th interaction turn. For top-1 recommendation, we compare the performances of our proposed EGE model with the iGRU and MBPI models without/with a post-filter on both the *Shoes* and *Dresses* datasets, respectively. More specifically, Table 4.1 contains two groups of rows for each dataset. The first group of rows reports the effectiveness of the tested models without a post-filter and the improvements of EGE over the best baseline model. The second group of rows reports the performances of the tested models with a post-filter and shows the improvements, in the same way as the first group. The best overall performing results across the two groups of rows in the table are highlighted in bold in Table 4.1. * denotes a significant difference in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to EGE/EGE (Filter) in each group, respectively. Comparing the results in the first group of rows in the table, we observe that our proposed EGE model achieves better performances of $6-11\%$ and $15-23\%$ at the 10th turn than the best baseline model (i.e. MBPI) across all metrics without a post-filter on the *Shoes* and *Dresses* datasets, respectively, while achieving improvements of $4-6\%$ and $11-16\%$ with a post-filter, respectively. Indeed, the EGE model is significantly better than the iGRU and MBPI models without/with a post-filter for each metric at the 10th turn with top-1 recommendation, except for MBPI on the *Dresses* dataset in terms of MRR@10 and SR.

In answer to RQ4.1, the results demonstrate that our proposed EGE model can outperform the state-of-the-art baseline models (i.e. iGRU and MBPI) overall after the first interaction turn. In particular, it is significantly more effective than both the supervised-learning-based approach (i.e. iGRU) and the RL-based approach (i.e. MBPI) without/with a post-filter at the 10th interaction turn with top-1 recommendation. Therefore, our proposed EGE model with Q-learning for POMDP can effectively mitigate the partial observation issue.

4.4.2 Impact of Historical Information (RQ4.2)

To address RQ4.2, we investigate how historical information affects the performance of our model by considering the users' historical feedback and the agent's historical recommendations. In particular, recall from Section 4.2 that the users' historical feedback is used as the input

Table 4.1: Recommendation effectiveness of our proposed EGE model and the baseline models at the 10th turn on both the *Shoes* and *Dresses* datasets. % Improv. indicates the improvements by EGE/EGE (Filter) over the best baseline model. The best overall results are highlighted in bold. * denotes a significant difference in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to EGE/EGE (Filter) in each group, respectively.

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

of the Evaluator in Figure 4.2 to judge the quality of the estimated state, while the agent's historical recommendations are used in a post-filter to remove the recommended items from the recommendation list. In summary, compared to the MBPI model, our EGE model takes the users' historical feedback into consideration during the training process.

Considering the usefulness of the users' historical natural-language feedback, we compare the effectiveness of our proposed EGE model with the MBPI model on both used datasets. In both Figures 4.3 and 4.4, we observe that EGE consistently outperforms MBPI in terms of NDCG@5, NDCG@10, MRR@10 and Success Rate through the 2nd turn to the 10th turn. In Table 4.1, we also observe that the EGE model is consistently and significantly better than the MBPI model in each group for each metric at the 10th interaction turn for top-1 recommendation, except for MBPI on the *Dresses* dataset in terms of MRR@10 and SR. This suggests that adopting the users' historical feedback in the judgement of the estimated states can benefit the interactive recommendation model.

Furthermore, we compare the performances of all the tested models (including our proposed EGE model) considering the usefulness of the agent's historical recommendations with a post-filter. In Figures 4.3 and 4.4, we observe that all of the tested models that apply a post-filter can consistently outperform those without a post-filter after the initial interaction turn. There is a trend that the gap between a model with a post-filter and the one without a post-filter increases at every interaction turn. Thus, this trend indicates that applying the post-filter on the recommendation list using the agent's historical recommendations demonstrates a *cumulative effect*. This suggests that applying a post-filter with the agent's historical recommendation can always further improve the performances of the interactive recommendation models.

Overall, in response to research question RQ4.2, we find that our proposed EGE model can benefit from both the users' historical feedback and the agent's historical recommendations.

4.4.3 Impact of Hyper-Parameters (RQ4.3)

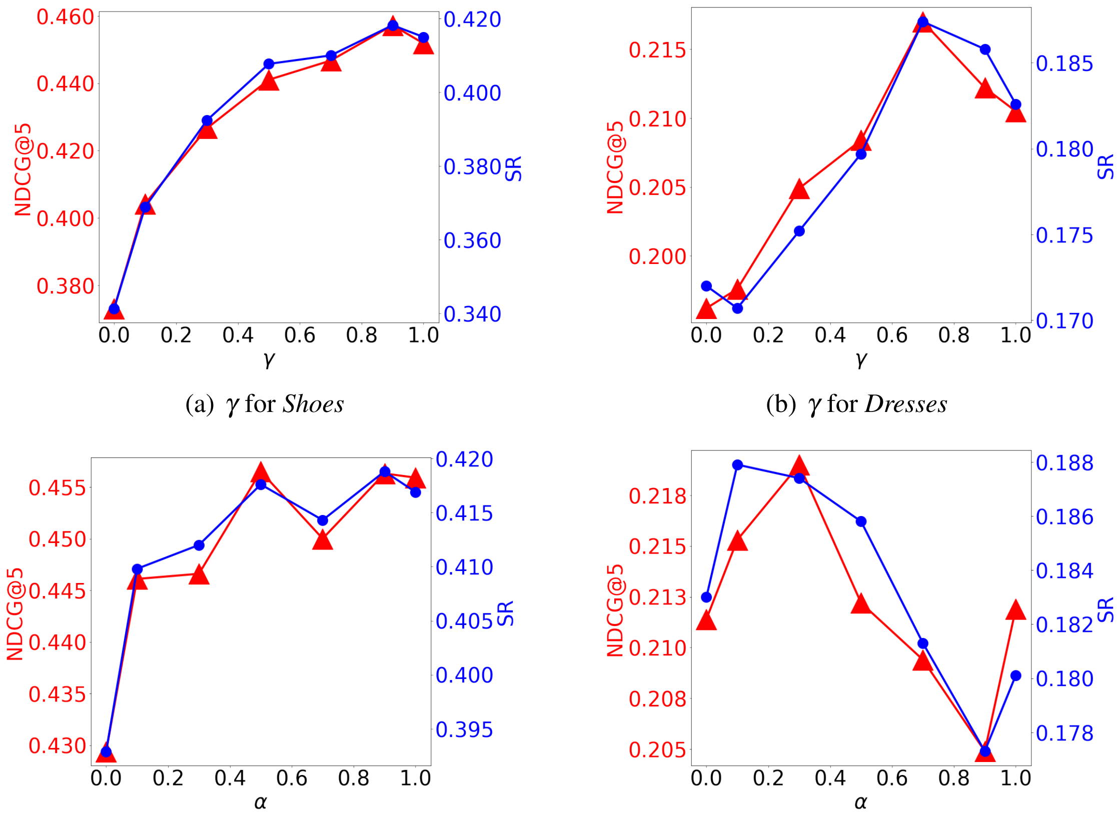
To address RQ4.3, Figure 4.5 depicts the effects of the *reward discount factor* γ and the *reward weighting factor* α on our proposed EGE model while applying a post-filter (denoted EGE (Filter)) in top-1 recommendation on the *Shoes* & *Dresses* datasets.

Effect of the *reward discount factor* γ Figure 4.5 (a) illustrates the NDCG@5 and SR of EGE (Filter) at the 10th turn in top-1 recommendation with different reward discount factors on the *Shoes* dataset. In particular, $\gamma = 0$ means that the models only consider immediate feedback, while $\gamma = 1$ means that the model weights all future rewards equally. We can see that the performance of EGE (Filter) improves when the reward discount factor γ increases from 0, except for $\gamma = 1$. Figure 4.5 (b) demonstrates a similar increasing trend on the *Dresses* dataset and both metrics reach a peak at $\gamma = 0.7$. The generally better performance of the model with $\gamma > 0.1$ than the model with $\gamma = 0$ leads to the conclusion that the Evaluator component does help to improve the overall recommendation effectiveness by considering long-term rewards. On the other hand, the decreased performance of the model with $\gamma = 1$ on the *Shoes* dataset and $\gamma > 0.7$ on the *Dresses* dataset shows that the reward discount factor should be set appropriately.

Effect of the reward weighting factor α Figure 4.5 (c) illustrates the NDCG@5 and SR of EGE (Filter) at the 10th turn in top-1 recommendation on the *Shoes* dataset with different reward weighting factors α, which weight the contributions of the visual reward r_t^{vis} and the ranking percentile reward r_t^{per} to the final rewards. In particular, $\alpha = 0$ means that the model only considers the ranking percentile reward r_t^{per}, while $\alpha = 1$ means that the model only takes the visual reward r_t^{vis} into consideration. We can see that the performance of EGE (Filter) improves when the reward weigthing factor α increases from 0 to 0.5, and varies slightly with $\alpha > 0.5$. Figure 4.5 (d) shows a distinctive trend on the *Dresses* dataset in that both the NDCG@5 and SR metrics first increase and reach a peak at $\alpha = 0.3$ and $\alpha = 0.1$, respectively, and then decease when α increases from 0.3 to 0.9. This trend shows that the visual reward r_t^{vis} is more informative than the ranking percentile reward r_t^{per} in the EGE (Filter) model on the *Shoes* dataset, while the ranking percentile reward r_t^{per} is more important than the visual reward r_t^{vis} on the *Dresses* dataset. Such a difference can be attributed to a domain factor from the datasets in that the images from the *Dresses* dataset usually include a human model to display the clothing while the images from the *Shoes* dataset only contain shoes without a model (as can be observed in the image databases for shoes[3] and dresses[4]). The visual features of the human models can confuse the ResNet component when mapping the dress images to the image feature (ResNet) space. Therefore, the generated dress image embeddings may be affected by the noises from the visual

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

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



(a) γ for *Shoes*

(b) γ for *Dresses*

(c) α for *Shoes*

(d) α for *Dresses*

Figure 4.5: Effects of (a) & (b) the *reward discount factor* γ and (c) & (d) the *reward weighting factor* α at the 10th turn in the top-1 recommendation scenario on the *Shoes* and *Dresses* datasets.

features of the human models, thereby reducing the utility of the visual rewards from the user simulator. To mitigate this issue, our future work will consider more advanced models (Z. Liu, Luo, Qiu, Wang, & Tang, 2016; W. Wang, Xu, Shen, & Zhu, 2018) that aim for effective fashion attribute detection for generating the dress image embeddings, or pre-trained vision and language models, such as CLIP Radford et al. (2021), to provide unified representation verctors for both image and text encodings (as described in the following Chapters 6 and 7).

Overall, in response to RQ4.3, we find that the ranking percentile reward r_t^{per} and the visual reward r_t^{vis} can help our EGE model to improve the recommendation performance.

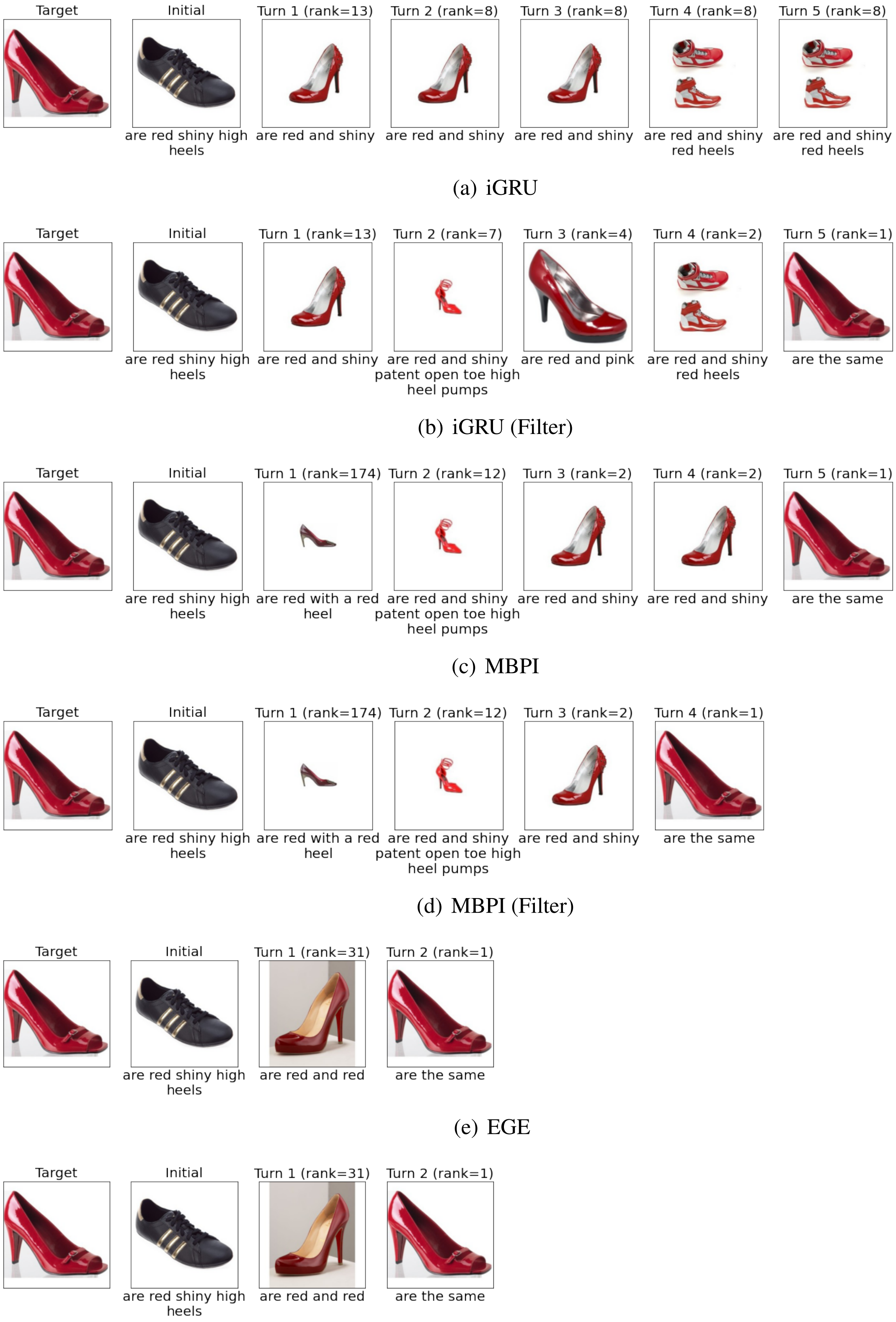
4.4.4 A Use Case

To consolidate the results observed in the above sections, we present a use case of the tested models without/with a post-filter in Figure 4.6: (a) iGRU, (b) iGRU (Filter), (c) MBPI, (d) MBPI (Filter), (e) EGE, (f) EGE (Filter) only on the *Shoes* dataset for top-1 recommendation. In Figure 4.6 (a-f), the first image is the target item desired by the user (labeled with "Target"), while the second image (labeled with “Initial”) is the initial recommendation proposed by the recommender system randomly. For a fair comparison, the initial images are the same across the tested recommender systems given the target image from the testing set. Then, the recommended top-1 items and the user comments in the following turns are presented. The rank of the target item is also presented above the images at each turn (e.g.“Turn 1 (rank=13)” in Figure 4.6 (a) iGRU, where the target image is ranked at the 13th position in the recommendation list at the first interaction turn). When the target item is recommended, the rank is 1 (e.g.“Turn 5 (rank=1)” in Figure 4.6 (b)), and the user simulator will give the comment: “are the same”. We observe that our proposed EGE model is the most effective recommender system among the tested models. Both EGE and EGE (Filter) only need two interactions to display the desired item, while the other tested models require at least 4 interactions given the same target and initial items. For instance, iGRU fails to recommend the target item within 5 interaction turns and recommends the same items repeatedly, even though the rank of the target item is getting higher. Though MBPI is more effective than iGRU, there is also a repeated recommendation at the 4th interaction turn when the post-filter is not applied. Furthermore, we also observe that the ranks of the target item with iGRU/iGRU (Filter) are much higher than the ranks of MBPI/MBPI (Filter) and EGE/EGE (Filter) at the first interaction turn. One possible reason is that the iGRU model is maximising the instant reward while the RL-based models are maximising the future accumulative rewards. In addition, our proposed EGE model is more effective at making use of the user’s natural-language feedback, i.e. “are red shiny high heels". However, both the iGRU (Filter) and MBPI (Filter) models continuously present items that violate the previous user’s feedback. Indeed, iGRU (Filter) recommends the red sport shoes that are violated from the “high heels”, while MBPI (Filter) recommends the black high heels that are contrary to the “red” colour. Note that a use case on the *Dresses* dataset also led to similar results and observations. We omit their reporting in this chapter to avoid redundancy.

4.5 Conclusions

In this chapter, we proposed a novel dialog-based recommendation model, denoted by the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP to effectively incorporate the users’ preferences over time in a partially observable environment. Specifically, we leveraged an Estimator to track and estimate the users’ preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations (with a post-filter to remove repeated recommendations), and an Evaluator to judge the quality of the estimated preferences considering the users’ historical feedback. Following previous work, we trained our EGE model by using a user simulator, which itself is trained to describe the differences between the target users’ preferences and the recommended items in natural language. Our experiments on the *Shoes* and *Dresses* datasets demonstrated that our proposed EGE model achieves significantly enhanced performances compared to the strongest baseline model (i.e. MBPI) – for instance (as shown in Table 4.1), improving by $6-23\%$ when a post-filter is not used, and $4-16\%$ when post-filtering is applied, respectively. Our reported results also showed that the EGE model can benefit from the historical information (i.e. the users' historical feedback and the agent's historical recommendations). The experimental results and analysis provide support for the thesis statement with **Research Topic 1** in Section 1.3.

Next, in Chapter 5, we argue that the existing formulation of interactive recommender systems suffer from their inability to capture the multi-modal sequential dependencies of textual feedback and visual recommendations becuase of their use of recurrent neural network-based or transformer-based models. Therefore, we aim to address the multi-modal sequence dependency issue by leveraging a recurrent-enhanced transformer architecture and introducing a feedback gate to separately process the textual and visual representations.



(a) iGRU

(b) iGRU (Filter)

(c) MBPI

(d) MBPI (Filter)

(e) EGE

(f) EGE (Filter)

Figure 4.6: A use case with different recommendation models on the *Shoes* dataset.

Chapter 5

Multi-Modal Sequence Dependency in State Tracking

In our thesis statement (as stated in Section 1.3), we postulated that we can effectively incorporate the users' preferences over time, with an RNN-enhanced Transformer structure for state tracking, by mitigating the multi-modal sequence dependency issue in the multi-modal conversational recommendation process. Therefore, in this chapter, we propose a novel multi-modal recurrent attention network (MMRAN) model for addressing the multi-modal sequence dependency issue so as to effectively incorporate the users' preferences over the long visual dialog sequences of the users' natural-language feedback and the system's visual recommendations. This chapter is mainly based on our work (Y. Wu, Macdonald, & Ounis, 2022b) "Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation" published in the proceedings of the 16th ACM Conference on Recommender Systems (RecSys 2022)[1].

In Chapter 4, we have addressed the partial observability issue in nautural-language feedback in the dialog-based recommendation task by proposing the Estimator-Generator-Evaluator (EGE) model with Q-learning for POMDP. In addition to the natural-language feedback, the multi-modal conversational recommendation task also involves sequences of visual recommendations across multiple iterations of interactions. However, such multi-modal dialog sequences (i.e. turns consisting of the system's visual recommendations and the user's natural-language feedback) make it challenging to correctly incorporate the users' preferences across multiple turns. Indeed, the existing formulations of interactive recommender systems suffer from their inability to capture the multi-modal sequential dependencies of textual feedback and visual recommendations because of their use of recurrent neural network-based (i.e., RNN-based, see Chapter 4) or transformer-based models. To alleviate the multi-modal sequence dependency issue, in this chapter, we propose a novel multi-modal recurrent attention network (MMRAN) model to effectively incorporate the users' preferences over the long visual dialog sequences of

[1] DOI: `https://doi.org/10.1145/3523227.3546774`

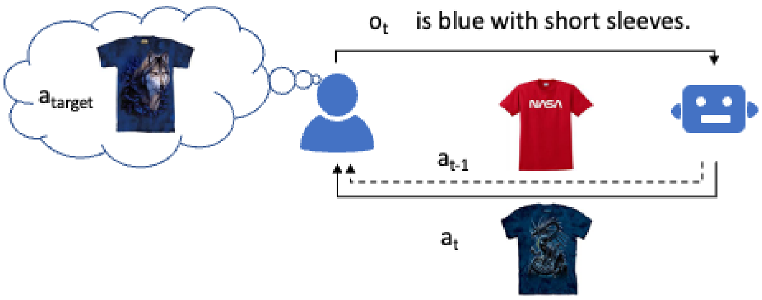


Figure 5.1: Example multi-modal recommendation scenario.

the users' natural-language feedback and the system's visual recommendations. Specifically, we leverage a gated recurrent network (GRN) with a feedback gate to separately process the textual and visual representations of natural-language feedback and visual recommendations into hidden states (i.e. representations of the past interactions) for multi-modal sequence combination. In addition, we apply a multi-head attention network (MAN) to refine the hidden states generated by the GRN and to further enhance the model's ability in dynamic state tracking. Following previous work, we conduct extensive experiments on the Fashion IQ Dresses, Shirts, and Tops & Tees datasets (introduced in Section 3.3) to assess the effectiveness of our proposed model by using a vision-language transformer-based user simulator as a surrogate for real human users (described in Section 3.3). Our results show that our proposed MMRAN model can significantly outperform several existing state-of-the-art baseline models (including the EGE model in Chapter 4). The results conform with our thesis statement with **Research Topic 2** in Section 1.3.

5.1 Motivations

As introduced in Section 3.2, the multi-modal interactive recommendation is specifically concerned with a goal-oriented multi-modal sequence of interactions between users and the recommender system, where users can receive visual recommendations (i.e. the items' images) and express fine-grained natural-language critiques about the recommendations based on their preferences. Figure 5.1 illustrates an example multi-modal interactive recommendation scenario. The multi-modal interactive recommendation task has been previously modelled using recurrent neural networks (RNNs, using a gated recurrent unit (GRU) (Guo et al., 2018; Yu et al., 2020) or a long short-term memory (LSTM) (R. Zhang et al., 2019)) or using a transformer (H. Wu et al., 2021) as *a state tracker* for both *multi-modal sequence combination* (Beard et al., 2018; Gkoumas et al., 2021) (i.e. combining the users' natural-language feedback sequence and the systems' visual recommendation sequence) and *dialog state tracking* (Fu et al., 2020; Liao et al., 2021; Y. Sun & Zhang, 2018) (i.e. eliciting the users' preferences over time). However, the actual neural networks adopted as the state trackers (such as GRUs (Chung et al., 2014), LSTMs (Hochreiter & Schmidhuber, 1997) or transformers (Vaswani et al., 2017)) are all originally designed for *single-modal* sequence modelling tasks (such as natural language processing (Otter et al., 2020)). Therefore, these models typically resort to combining the textual and visual representations with a *concatenation operation* (Guo et al., 2018; H. Wu et al., 2021; R. Zhang et al., 2019), rather than processing the differing multi-modal sequence data separately.

Despite the expressiveness and complementary of visual recommendations and the corresponding natural-language feedback in multi-modal interactive recommendations, the long lengths of the dialog sequences makes it challenging to correctly incorporate the users' preferences over time, thereby resulting in a degraded satisfaction of the users' information needs with inappropriate recommendations. Indeed, the existing formulations of interactive recommender systems suffer from an inability to capture *multi-modal sequence dependencies* between the textual feedback and visual recommendations using either the GRU/LSTM-based models (Goodfellow et al., 2016; Guo et al., 2018; R. Zhang et al., 2019) or the transformer-based model (H. Wu et al., 2021). Specifically, we argue that the inability of these GRU/LSTM-based and transformer-based models at capturing such multi-modal sequence dependencies of the dialog sequences is inherently due to their limitations in *combining multi-modal sequences* or *tracking dialog states* (as we further discuss in Section 5.2).

In this chapter, we alleviate the *multi-modal sequence dependency* issue in multi-modal dialog sequences modelling by addressing the *multi-modal sequence combination* and the *dialog state tracking*, respectively. To better combine the multi-modal dialog sequences than using a concatenation operation, we extend the traditional GRU architecture with an extra *feedback gate* (called a gate recurrent network (GRN), inspired by Donkers, Loepp, and Ziegler (2017); Manotumruksa et al. (2018)) to separately process the textual feedback and the visual items in the visual dialog sequences. To better track the users' dynamic preferences across multiple interaction turns, a multi-head attention network (MAN) is placed on top of our proposed GRN component to refine the GRN's hidden states and to further enhance the model's ability in dialog state tracking, inspired by RNN-enhanced transformers (Z. Wang, Ma, Liu, & Tang, 2019). To this end, we propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation to effectively incorporate the users' preferences over time from the multi-modal dialog sequences of the users' natural-language feedback and the systems' visual recommendations. Following previous work (see Section 3.3), we train and evaluate our MMRAN model by using a vision-language transformer-based user simulator (VL-Transformer), which has been previously shown to be a good surrogate for real users. Our extensive experiments conducted on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees* datasets (see Section 3.3.2) show that our proposed MMRAN model can significantly outperform several existing state-of-the-art baseline models. The main contributions of this chapter are as follows:

- We propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation. Our model separately processes the textual feedback sequences and the visual item sequences for multi-modal sequence combination, and tracks the dialog states using abstract representations of the previous interactions. We show that our proposed MMRAN model is more effective in capturing the dialog sequence information of the natural-language feedback and the visual recommendations compared to the existing baseline models.

• We propose a gated recurrent network (GRN) for extracting the hidden states of the past interactions from the natural-language feedback and the visual recommendations. Our GRN extends the traditional gated recurrent unit (GRU) with a *feedback gate* to capture the correlation between the textual feedback at the current turn and the hidden state of the previous turn.

• We deploy an advanced RNN-enhanced transformer architecture (Z. Wang et al., 2019) for interactive recommendation, in order to effectively track the dialog states with a multi-head attention network (MAN) using the GRN's abstract representations.

• We perform extensive empirical evaluations with our proposed MMRAN model on the multi-modal interactive recommendation task, demonstrating significant improvements over the existing state-of-the-art approaches (including the EGE model in Chapter 4).

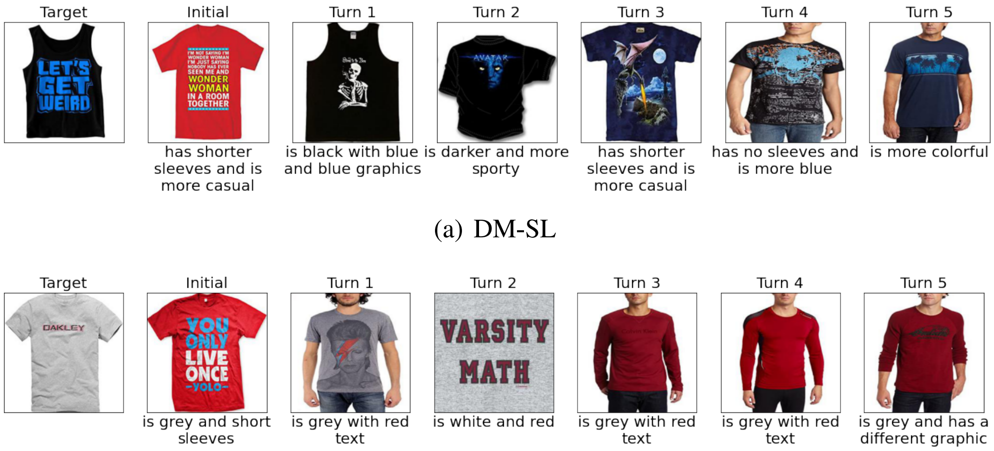
The remainder of this chapter is structured as follows: we first discuss the limitations of the existing multi-modal interactive recommendation models in Section 5.2. We also review the related work and position our contributions in comparison to the existing literature in Section 5.3. Then we detail our proposed MMRAN model in Section 5.4. Afterwards we describe our experimental setup in Section 5.5 and report our experimental results in Section 5.6, respectively. Finally, we summarise our findings in Section 5.7.

5.2 Multi-modal Interactive Recommendation

In this section, we recap the problem of the multi-modal interactive recommendation task (Section 5.2.1). Then, we briefly elicit the limitations of the RNN/transformer-based models in terms of the multi-modal sequence dependency issue (Section 5.2.2).

5.2.1 Preliminaries

We study the multi-modal interactive recommendation task by considering a user interacting with a recommender system using iterative multi-turn interactions through vision and language. At the t-th interaction turn, the recommender system presents a candidate image a_{t-1} selected from a candidate pool $\mathscr{I} = \{a_i\}_{i=0}^{N}$ to the user. The user then provides a natural language critique o_t as feedback, describing the major visual differences between the candidate image and their desired item. Specifically, we assume that the user only gives feedback on the top-ranked candidate item in the ranking list (see Section 3.2). According to the users' natural-language


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

(a) DM-SL

(b) MMIT

Figure 5.2: Examples of multi-modal interactive recommendations obtained from (a) DM-SL (Guo et al., 2018) and (b) MMIT (H. Wu et al., 2021).

feedback o_t and the interaction history up to turn t, $\tau_t = (o_{\leq t}, a_{<t}) \in \mathscr{H}$ (i.e. a set of interaction history), where $o_{\leq t} = (o_1, ..., o_t) \in \mathscr{O}$ (i.e. a set of the users' natural-language feedback) and $a_{<t} = (a_0, ..., a_{t-1}) \in \mathscr{A}$ (i.e. a set of items for recommendation), the recommender system selects another candidate image a_t from the candidate image pool. This vision-language interaction process continues until the user's desired target image a_{target} is recommended or a maximum number of interaction turns, M, is reached, leaving the user unsatisfied.

5.2.2 Multi-Modal Interactive Models

Figure 5.2 shows examples of interactive recommendations obtained from (a) the Dialog Manager (DM) (Guo et al., 2018) model, which is based on a gated recurrent unit (GRU) with a supervised-learning setup (which we denote as DM-SL); and (b) the multi-modal interactive transformer (MMIT) (H. Wu et al., 2021) model based on a transformer. In each example, the recommender gives a random initial recommendation (denoted "Initial") to the user, while the user with a desired target item (denoted "Target") provides natural-language feedback about the recommendation at each turn. Then, the recommender system updates the ranking list of the candidate items for the next recommendation according to the user's feedback.

The GRU/LSTM-based Models

In GRU/LSTM-based models (Guo et al., 2018; R. Zhang et al., 2019) for multi-modal interactive recommendation, due to the inability of GRUs/LSTMs in processing different multi-modal data separately, the representations of the users' natural-language feedback and the systems' visual recommendations are usually combined with a concatenation operation and a multilayer perceptron (MLP), so as to form a single input of the GRUs/LSTMs at each turn. Such a concatenation operation on the multi-modal sequence data causes the combined textual and visual representations to be memorised or forgotten synchronously at each interaction turn in the GRUs/LSTMs-based state trackers (Goodfellow et al., 2016; Guo et al., 2018; R. Zhang et al., 2019) (**Limitation 1**). For instance, in Figure 5.2 (a), a sleeveless shirt is recommended at the 1st turn by the DM-SL model due to the initial comment, "shorter sleeves", compared to the red T-shirt shown at the initial turn, while the sleeveless feature shown in the image at the 1st turn and similarly conveyed by the natural-language at the initial turn is omitted by the recommender for the following recommendations. However, we argue that the users' feedback should have more effect on the hidden state of the GRUs/LSTMs in addition to the combined textual and visual representations, in that the natural-language feedback explicitly conveys the users' information needs while the rejected visual recommendations can be noisy by also containing the users' undesired features.

The Transformer-based Model

In the transformer-based model for multi-modal interactive recommendation (H. Wu et al., 2021), the representations of the users' natural-language feedback and the systems' visual recommendations at all turns are *concatenated* together, while the dialog states (i.e. the estimated users' dynamic preferences) are directly tracked and inferred from all the concatenated textual and visual representations. Although the textual and visual representations at all turns in a multi-modal dialog sequence can fully interact with each other by using a multi-head attention mechanism (Vaswani et al., 2017) in the transformers, we argue that the effectiveness of the transformer-based model is limited as it cannot consider the previous inferred hidden states in an iterative manner (i.e. abstract representations of the past interactions) of the multi-modal dialog sequence as performed by the GRU/LSTM models at each turn (**Limitation 2**). For instance, in Figure 5.2 (b), for the MMIT model, the "red" colour in the comment at the 2nd turn refers to "red text" in the comment at the 1st turn, while it is misunderstood by the recommender system and taken as the colour of the shirt according to the successively recommended "red" shirts from the 3rd turn to the 5th turn.

Summary of Limitations

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

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

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

In summary, the existing multi-modal interactive recommendation models based on only GRUs, LSTMs or transformers are not able to properly process the multi-modal dialog sequences of the natural-language feedback and recommended visual items, which limits these models' ability to incorporate the users' preferences over time. In Section 5.4, we propose a model that addresses these limitations. In the next section, we detail related work in multi-modal interactive recommendation and recurrent neural models.

5.3 Related Work

In this section, we first introduce the gating mechanisms in RNNs. We then discuss the RNN-enhanced transformers.

5.3.1 Gating Mechanisms of Recurrent Models

Traditional RNNs usually suffer from the vanishing gradient problem when processing long sequences (Hochreiter & Schmidhuber, 1997). Recurrent units such as a long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and a gated recurrent unit (GRU) (Chung et al., 2014) are extensions of traditional RNNs, which use gating mechanisms to control the influence of a hidden state of the previous step. While the GRU and LSTM architectures can alleviate the vanishing gradient problem (Chung et al., 2014; Hochreiter & Schmidhuber, 1997), they cannot process different modalities separately at the same time. The representations of the multi-modal sequences are usually combined with a concatenation operation as a single input of the GRUs/LSTMs at each turn (Guo et al., 2018; Tan, Goel, Nguyen, & Ong, 2019; R. Zhang et al., 2019). Many researchers have extended the GRUs/LSTMs to incorporate contextual information associated with the sequence information, such as transition contexts (the time intervals and the geographical distances) by using time and/or spatial-based gates (Manotumruksa et al., 2018; Smirnova & Vasile, 2017; Zhu et al., 2017). In the multi-modal interactive recommendation task, the users' natural-language feedback can be taken as contextual information associated with the visual recommendation sequences. However, to the best of our knowledge, such an approach to associate the sources of contextual information has not been investigated for multi-modal interactive recommendation to address **Limitation 1**.

5.3.2 RNN-Enhanced Transformers

In the transformer-based model (H. Wu et al., 2021) for interactive recommendation, dialog states (i.e. the estimated users' dynamic preferences) can only be directly tracked and inferred from all the concatenated textual and visual representations instead of being estimated from the abstract representations in the past interactions of the dialog sequences, as performed by the

GRU/LSTM models at each turn. To alleviate such inherent limitations of the transformers in state tracking, a number of previous studies (Hao et al., 2019; Kim, Lin, Jeon, Min, & Sohn, 2018; J. Lei et al., 2020; Z. Wang et al., 2019) have investigated RNN-enhanced transformers for sequence modelling tasks, such as R-Transformer (Z. Wang et al., 2019), to take the benefits from both the RNNs for abstract representations at each turn and from the transformers for the whole sequence's overall feature interactions in sequence modelling. Z. Wang et al. (2019) proposed an RNN-enhanced transformer model with a sliding window (called an R-Transformer) to benefit from the advantages of both an RNN and a transformer's multi-head attention mechanism. Three layers (i.e. RNNs with a sliding window, a multi-head attention layer, and a feedforward layer) were arranged hierarchically. In particular, the RNNs process sequences using a sliding window and generate the hidden states of the past interactions sequentially, while the multi-head attention layer captures the dialog states among the RNNs' hidden states of the previous turns, and the feedforward layer conducts non-linear feature transformation. However, these RNN-enhanced transformers have not yet been investigated for multi-modal interactive recommendation in order to address **Limitation 2**.

As discussed in Section 5.2, given the limitations of the GRU, LSTM and transformer-based models, we argue that the existing interactive recommendation models based on only GRUs, LSTMs or transformers are not able to properly process the multi-modal dialog sequences of natural-language feedback and recommended visual items. This limits the ability of such models in incorporating the users' preferences over time. To address **Limitation 1** & **2**, we propose a novel multi-modal recurrent attention network (MMRAN) model for interactive recommendation. Specifically, our model separately processes the textual feedback sequences and the visual item sequences for multi-modal sequence combination so as to address **Limitation 1**, while it tracks the dialog states with the abstract representations of the previous interactions in order to address **Limitation 2**. To the best of our knowledge, this novel structure of our MMRAN model constitutes the first work based on a multi-modal recurrent attention network in multi-modal interactive recommendations.

5.4 Methodology

We now define our proposed **M**ulti-**M**odal **R**ecurrent **A**ttention **N**etwork (MMRAN) model and introduce its components. Figure 5.3 shows the architecture of MMRAN, which aims to effectively incorporate the users' preferences over time. The architecture consists of three parts: text & image encoders, a gated recurrent network (GRN), and a multi-head attention network (MAN). We also describe the training of the MMRAN model using multi-turn interactions with a user simulator.

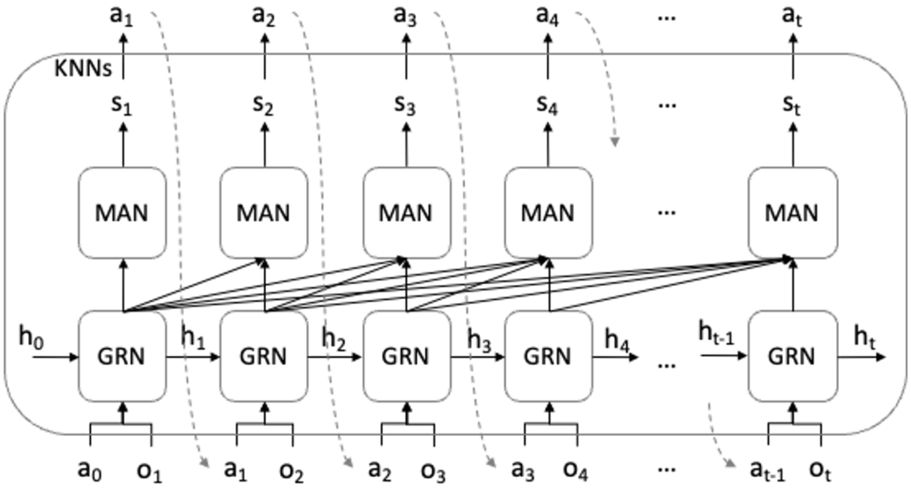


Figure 5.3: The multi-modal recurrent attention network (MMRAN) model.

Text & Image Encoders The text encoder (denoted $TxtEnc(\cdot)$) consists of a 1D convolutional layer (1D-CNN) and a subsequent linear layer as in (Guo et al., 2018), where the user's natural-language feedback o_t (with each word represented by a one-hot vector) is extracted into a textual sentence representation $TxtEnc(o_t)$. Although there are many advanced pre-trained transformer-based language models (such as BERT, see Section 3.4.1) for processing the natural-language feedback, we adopt a one-hot vector for each word with a pre-defined vocabulary (Guo et al., 2018) of fashion-related terms when generating textual sentence representations, thus allowing fair comparisons with existing works (Guo et al., 2018). Furthermore, a pre-defined fashion vocabulary is much smaller and is more concentrated on fashion features than BERT. Similarly, the image encoder (denoted $ImgEnc(\cdot)$) consists of the ImageNet pre-trained ResNet101 model (K. He et al., 2016) and a subsequent linear layer, as in (Guo et al., 2018), where a candidate image a_{t-1} is extracted into image feature representations $ImgEnc(a_{t-1})$. To simplify the notations, in the following we directly use o_t and a_{t-1} as their representations, respectively. Then, both the visual and textual representations are passed to a gated recurrent network (GRN) and a multi-head attention network (MAN) to estimate the user's preferences.

The Gated Recurrent Network (GRN) To address **Limitation 1** and effectively incorporate the users' preferences from the multi-modal dialog sequences of the users' natural-language feedback and the recommended visual items, inspired by (Donkers et al., 2017; Manotumruksa et al., 2018), we propose a gated recurrent network (GRN) with a *feedback gate* for *multi-modal sequence combination*. Figure 5.4 shows the architecture of our proposed gated recurrent network (GRN). Our GRN extends the traditional gated recurrent unit (GRU) with an extra gate (i.e. a feedback gate β_t) to directly impose more effect on the hidden state in addition to the combined textual and visual representations, in that the natural-language feedback explicitly conveys the users' information needs. The estimated hidden states of the user's preferences can be achieved with $h_t = GRN(h_{t-1}, a_{t-1}, o_t)$. In particular, the proposed feedback gate β_t controls the influences of the current textual feedback o_t at each state as follows:

$$\beta_t = \sigma(W_{\beta,h}h_{t-1} + W_{\beta,o}o_t + b_\beta) \tag{5.1}$$

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

$$c_t = W_{c,a}a_{t-1} + W_{c,o}o_t + b_c \tag{5.2}$$

$$z_t = \sigma(W_z c_t + U_z h_{t-1} + b_z) \tag{5.3}$$

$$r_t = \sigma(W_r c_t + U_r h_{t-1} + b_r) \tag{5.4}$$

$$\tilde{h}_t = \tanh(W_h c_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{5.5}$$

$$h_t = (1 - \beta_t) \odot o_t + [(1 - z_t)h_{t-1} + z_t\tilde{h}_t] \tag{5.6}$$

where c_t is an initially inferred multi-modal representation of the visual recommendation a_{t-1} and the corresponding natural-language feedback o_t. z_t, r_r are update and reset gates, respectively. $\tilde{h}_t$ is a candidate hidden state. $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and hyperbolic tangent functions, respectively. U_z, U_r and U_h are the weight matrices that capture the recurrent connections between every two adjacent hidden states h_{t-1} and h_t. $\odot$ denotes the element-wise product. W and b with subscripts are, respectively, the transition matrices and the corresponding biases. By including the natural-language feedback o_t through an aggregation operation (Equation (5.6)), o_t has more effect on the hidden state h_t. In addition, a sliding window with size $N_{sliding\_window}$, as in (Z. Wang et al., 2019), can be used to limit the length of the multi-modal dialog sequences considered at each turn. We investigate its impact on the model's performance in Section 5.6.2.

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

The Multi-head Attention Network (MAN) To address **Limitation 2** and further track the dialog states among the GRN's hidden states of the previous turns, we adopt a multi-head attention network (MAN) architecture that enables our MMRAN model to consider the entire history of the multi-modal interactions during each interaction turn. The multi-head attention mechanism

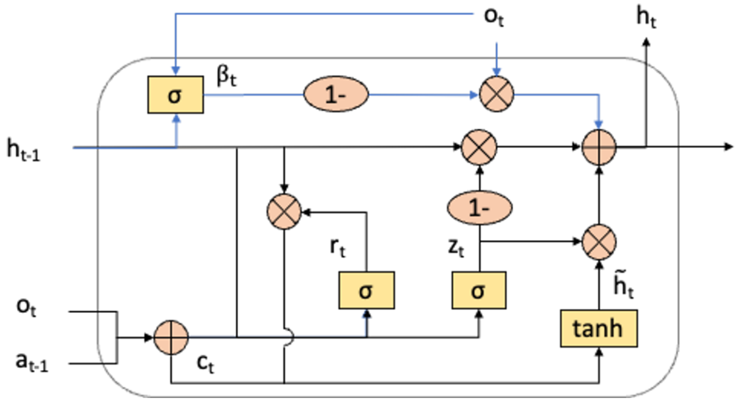


Figure 5.4: Our proposed GRN architecture.

in transformers has been shown to be extremely effective to learn the long-term dependencies in the sequence modelling, since it allows a direct connection between every pair of its input representations (Vaswani et al., 2017; Z. Wang et al., 2019). More specifically, in the multi-head attention mechanism, each input representation at each turn will attend to all the other input representations in the past interactions, thereby obtaining a set of attention scores that are used to refine its representations. In particular, the estimated hidden states of the users' preferences h_i (where $i \in [1,t]$) are further encoded with a multi-layer transformer encoder $TranEnc(\cdot)$ (with N_{layers} layers), which includes the multi-head attention mechanism (with N_{heads} attention heads). The refined hidden states are defined as follows:

$$h_1^{'}, ..., h_t^{'} = TranEnc(h_1, ..., h_t) \tag{5.7}$$

The estimated final state of the user's preferences is obtained as $s_t = Linear(ReLu(Mean(h_1^{'}, ..., h_t^{'})))$. For top-$K$ candidate recommendation, the closest images to the estimated state s_t under the Euclidean distance are recommended: $a_t \sim KNNs(s_t)$, where $KNNs(\cdot)$ is a softmax distribution over the K nearest neighbours of s_t.

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

Training with A User Simulator To avoid collecting and annotating entire multi-modal conversations, which is expensive, time-consuming, and does not scale (S. Zhang & Balog, 2020), we adopt an existing vision-language transformer-based user simulator (VL-Transformer) (H. Wu et al., 2021) as a reasonable proxy for real human users for training and evaluating our proposed MMRAN model. The user simulator considers the differences in the image features of the candidate image $a_{candidate}$ and the target image a_{target} to produce a relative caption:

$$w_{\leq i} = f([ResNet(a_{candidate}), ResNet(a_{target})]) \tag{5.8}$$

where $w_{\leq i} = (w_0, ..., w_i)$ is the word sequence generated for the caption (i.e. o_t), $f(\cdot)$ is the relative captioning network and $ResNet(\cdot)$ is the ImageNet pre-trained ResNet101 model (K. He et al., 2016) to obtain the prominent set of visual attributes from each image. The features of the candidate and target image pairs are concatenated to form a set of relative features, $[ResNet(a_{candidate}), ResNet(a_{target})]$. Furthermore, we train our proposed MMRAN model with a triplet loss objective, $L_{triplet}$, similar to (Guo et al., 2018; H. Wu et al., 2021):

$$L_{triplet} = \max(0, ||s_t - a_+||_2 - ||s_t - a_-||_2 + m) \tag{5.9}$$

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

5.5 Experimental Setup

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

- RQ5.1: Does our proposed MMRAN model outperform the existing state-of-the-art baseline models in the multi-modal interactive recommendation task with natural-language feedback?
- RQ5.2: Does the GRN structure address Limitation 1 and thereby improve the MMRAN models' ability to incorporate the users' preferences from the multi-modal dialog sequences?
- RQ5.3: Does the MAN structure address Limitation 2 so as to improve the MMRAN models' ability to effectively track dialog states?

5.5.1 Datasets & Measures

Datasets We perform experiments on the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets (introduced in Section 3.3.2). On these three datasets, both relative captions of image pairs and the images of the fashion products (Images$_{origin}$) are available for training and testing the user simulators (i.e. relative captioners) and the recommendation models, respectively. The statistics of the Fashion IQ datasets are summarised in Table 3.1.

Measures We measure the effectiveness of the multi-modal interactive recommendation models at the M-th turn interaction with top-heavy metrics, such as NDCG@N (i.e. Normalised Discounted Cumulative Gain truncated at rank $N = 10$, see Section 2.1.3), MRR@N (i.e. Mean Reciprocal Rank truncated at rank $N = 10$, see Section 2.1.3), and SR (i.e. Success Rate that is the percentage of the succeeded users among all the users with top-1 recommendation, see Section 3.2.2). In particular, both NDCG@N and MRR@N measure the quality of the ranking list at each turn, while SR measures the efforts for finding the target items over multi-turn interactions. We apply all the evaluation metrics (i.e. NDCG@10, MRR@10, SR) at the 5th interaction turn for significance testing.

5.5.2 Baselines

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

• **RNNs (GRUs/LSTMs)**: The Dialog Manager model (Guo et al., 2018) is a multi-modal interactive recommendation model based only on a *GRU* as the state tracker. There are two variants of the Dialog Manager model in terms of their learning approaches: Dialog Manager with a supervised-learning setup (denoted DM-SL) and Dialog Manager with a model-based reinforcement learning setup (denoted DM-RL). The DM-SL model is trained with a triplet loss (i.e. Equation (5.9)) to maximise the short-term rewards, while the DM-RL model is further trained with a cross entropy loss to maximise the cumulative future rewards by exploring all possible recommendation trajectories in the future turns given a known environment (i.e. a user simulator) (Guo et al., 2018). In addition, as a further possible baseline, we envisage that an LSTM can also act as a state tracker in the Dialog Manager model with a supervised-learning setup (denoted DM-LSTM). Furthermore, we apply the Estimator-Generator-Evaluator (EGE) model (see Chapter 4) as another GRU-based baseline model, which uses reinforcement learning with a partially observable Markov decision process (POMDP).

• **Transformers**: The multi-modal interactive transformer (MMIT) model (H. Wu et al., 2021) applies only a *transformer*. The MMIT model directly attends to the entire multi-modal interaction history of both the users' previous textual feedback and the system's visual recommendations. The MMIT model is also trained with a triplet loss as per DM-SL.

• **RNN-Enhanced Transformers**: R-Transformer (Z. Wang et al., 2019), a typical RNN-enhanced transformer, can be adapted as a strong baseline model based on a GRU and a transformer. There are two variants of the R-Transformer model: R-Transformer with a window size 3 (Z. Wang et al., 2019) (which we denote as R-T$_{Local}$) and R-Transformer without a sliding window (which we denote as R-T$_{Global}$). The R-T$_{Local}$ and R-T$_{Global}$ models are also trained with a triplet loss similar to DM-SL.

The baseline models based on RNNs (i.e. DM-LSTM, DM-SL and DM-RL) and Transformers (i.e. MMIT)) are the two representative formulations of the existing multi-modal interactive recommendation task, which are formulated as a sequential modelling problem with an RNN (such as a GRU or a LSTM) or a transformer, respectively. The RNN-enhanced transformer models (i.e. R-T$_{Local}$ and R-T$_{Global}$) adapted from the literature (Z. Wang et al., 2019) can provide stronger baselines using more advanced network structures. In addition, the GRN component in MMRAN can also be adapted as a multi-modal interactive recommendation model to estimate the users' preferences and to make recommendations independently (which we denote as MMRAN w/o MAN).

5.5.3 Experimental Settings

Setup for User Simulator We first train the existing VL-Transformer user simulator (H. Wu et al., 2021) for relative captioning on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees* datasets, separately. The network parameters are randomly initialised. We use the Adam (Kingma & Ba, 2014) optimiser with an initial learning rate of 10^{-4}. The batch size is 16, and the maximum number of epochs is 30. The dimensionality of the embeddings and hidden states is 512. Section 3.3.3 provides a comparison of the VL-Transformer user simulator with another recent user simulator called Show Tell (Vinyals et al., 2015) to demonstrate how close the VL-Transfomer user simulator behaves in comparison to real human captions.

Setup for Recommender Training Next, we train our proposed MMRAN model using the VL-Transformer user simulator trained on the *Fashion IQ Dresses, Shirts, Tops & Tees* datasets, respectively. The recommendation models' parameters are randomly initialised. We use Adam with a learning rate of 10^{-3} (Guo et al., 2018; R. Zhang et al., 2019). The embedding dimensionality of the feature space is set to 256 and the batch size to 128 following the setting in (Guo et al., 2018). For each batch, we train the model with 10 interaction turns as in (Y. Wu et al., 2021). The maximum number of epochs for training is 20. For the recommendation task, early stopping (Goodfellow et al., 2016) is used to avoid overfitting. The training terminates when average NDCG@10 over all the interaction turns on the validation sets stops improving for 5 epochs, or when the maximum number of training epochs is reached. For our proposed MMRAN model, we consider all the previous textual feedback and visual recommendations at each turn.

Setup for Recommender Evaluation We evaluate the interactive recommendation models for top-K (i.e. $K = 1$) recommendation with multi-turn interactions $M \in [1,5]$ on the above three datasets, respectively. The previously recommended items are removed from the ranking list at each turn with a post-filter to avoid repeated recommendations, as in (Y. Wu et al., 2021). For a fair comparison, we mainly compare the effectiveness of the tested models at the 5th

Table 5.1: The multi-modal interactive recommendation effectiveness of our proposed MMRAN model and the baseline models at the 5th turn on the three used datasets. % Improv. indicates the improvements by MMRAN over the best baseline model. The best overall results are highlighted in bold. * and † denote a significant difference in terms of a paired t-test (Holm-Boferroni correction, $p < 0.05$), compared to MMRAN and MMRAN w/o MAN in each dataset, respectively.

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

turn (i.e. $M = 5$) using the paired t-test (applying Holm-Bonferroni for multiple comparison correction (Holm, 1979)). When a user successfully finds the target item in less than 5 turns, we consider the ranking metrics (i.e. NDCG@10 and MRR@10) for that user to be equal to one for all turns thereafter.

5.6 Experimental Results

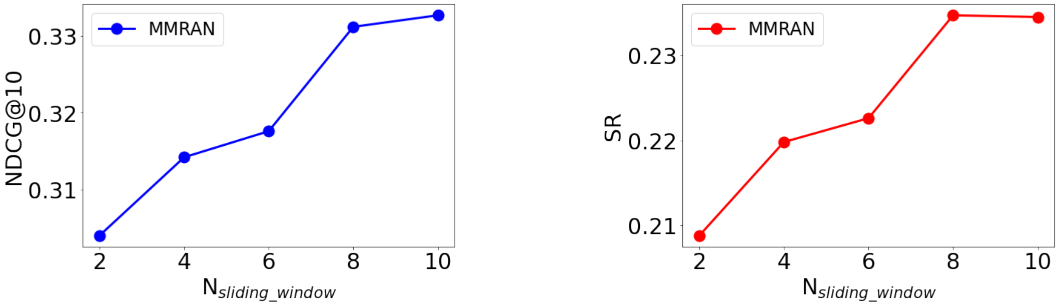
We now analyse the experimental results to answer the three research questions that are stated in Section 5.5, concerning the effectiveness of our proposed MMRAN model for multi-modal interactive recommendations with natural-language feedback (Section 5.6.1), the impact of the GRN structure for multi-modal sequence combination (Section 5.6.2) and the impact of the MAN structure for dialog state tracking (Section 5.6.3). We also show a use case from the logged experimental results to consolidate our findings (Section 5.6.4).

5.6.1 MMRAN vs. Baselines (RQ5.1)

To answer RQ5.1, we assess the effectiveness of our MMRAN model by comparing them with seven strong recommendation approaches in the literature. Table 5.1 shows the obtained recommendation performances of the baseline models (i.e. DM-LSTM, DM-SL, DM-RL, EGE, MMIT, R-T$_{Local}$ and R-T$_{Global}$ in the first part) as well as the MMRAN model variants (in the second part) with the same test sets of the *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets at the 5th interaction turn. The best overall performances across the three groups of columns in the table are highlighted in bold in Table 5.1. In each group, * and † denote, respectively, significant differences compared to MMRAN and MMRAN w/o MAN (i.e. GRN only), in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$) (Holm, 1979). Among the tested baseline models and our proposed models, there are three different types of state trackers: RNNs (i.e. GRUs/LSTMs/GRN, such as DM-LSTM, DM-SL, DM-RL, EGE, MMRAN w/o MAN), Transformers (such as MMIT), and RNN-Enhanced Transformers (such as R-T$_{Local}$, R-T$_{Global}$, and MMRAN). Comparing the results in the table, we observe that our proposed MMRAN model achieves better performances of 24-28%, 17-20% and 18-22% at the 5th turn than the best baseline model across all metrics on the *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees*, respectively. Indeed, our proposed MMRAN model is significantly better than DM-LSTM, DM-SL, DM-RL, EGE, MMIT, R-T$_{Local}$ and R-T$_{Global}$ for each metric at the 5th turn with top-1 recommendation. In answer to RQ1, the results demonstrate that our proposed MMRAN model does overall outperform the previous state-of-the-art baseline models. In particular, it is significantly more effective than all of the GRU/LSTM-based models (i.e. DM-LSTM, DM-SL, DM-RL and EGE) owing to its superior multi-modal sequence combination with GRN and enhanced dialog state tracking with MAN, the transformer-based model (i.e. MMIT) due to improved multi-modal sequence combination with GRN, and the RNN-enhanced transformer models (i.e. R-T$_{Local}$ and R-T$_{Global}$) due to a better multi-modal sequence combination with a feedback gate in GRN. Therefore, these results demonstrate that our proposed MMRAN model, with the multi-modal recurrent attention network, can effectively incorporate the users' preferences over time.

5.6.2 Impact of GRN (RQ5.2)

To address RQ5.2, the second part of Table 5.1 examines the comparative performances of the MMRAN and MMRAN w/o MAN models (i.e. GRN only) with different components for tracking and estimating the users' preferences (i.e. state trackers). First, focusing on GRN, we observe that the MMRAN w/o MAN model performs significantly better than the DM-LSTM and DM-SL models in terms of all metrics on the three datasets. The significantly better performance of the GRN component indicates that the extra feedback gate can enhance the GRU's ability in combining the multi-modal sequences (i.e. textual feedback sequences and visual recommendation sequences). In addition, we also observe that MMRAN with a GRN component in the state tracker performs significantly better than the R-T$_{Local}$ and R-T$_{Global}$ baselines, which all use a GRU component in the state tracker. This suggests, as we argued in Section 5.1, that imposing more effect of the users' natural-language feedback on the hidden state of the GRUs in addtion to the combined textual and visual representations can benefit the interactive recommendation model. Furthermore, Figure 5.5 illustrates the NDCG@10 and SR performances of MMRAN at the 5th turn in top-1 recommendation on the *Fashion IQ Dresses* dataset with different sliding window sizes $N_{sliding\_window}$ that limit the lengths of the multi-modal dialog sequences at each turn. We can see that the performance of MMRAN improves when the value of the sliding window size $N_{sliding\_window}$ increases from 2 to 10, except for $N_{sliding\_window} = 10$



(a) NDCG@10

(b) SR

Figure 5.5: Effects of the sliding window size $N_{sliding\_window}$ over the multi-modal dialog sequences on our proposed MMRAN at the 5th turn on *Fashion IQ Dresses*.

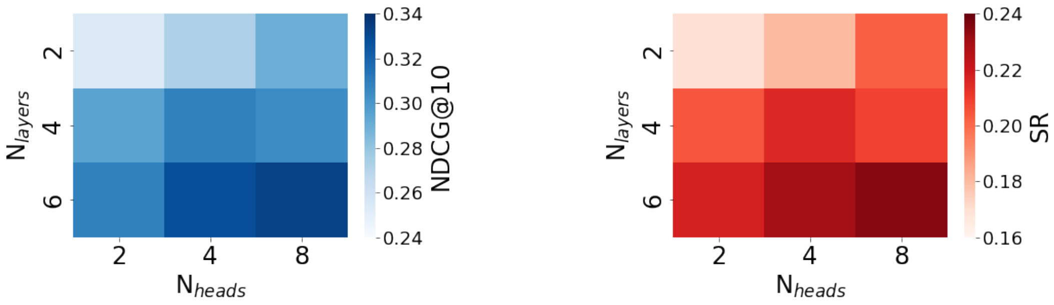
in term of SR. Further ablation studies on the *Fashion IQ Shirts* and *Tops & Tees* datasets also led to similar results and observations. We omit their reporting in this chapter because of space constraints.

Overall, for RQ5.2, we conclude that the GRN component with a natural-language feedback gating mechanism enhances the model's ability to combine the multi-modal sequences so as to address **Limitation 1**, thereby better incorporating the users' information needs from the multi-modal dialog sequences than the traditional GRU network.

5.6.3 Impact of MAN (RQ5.3)

To address RQ5.3, Figure 5.6 depicts the effects of the MAN's layers N_{layers} and the MAN's attention heads N_{heads} on our proposed MMRAN in terms of NDCG@10 and SR at the 5th turn on *Fashion IQ Dresses*. We can see that the performance of MMRAN improves when the MAN's layers N_{layers} and the MAN's attention heads N_{heads} increase from 2, respectively, except for $N_{layers} = 4$ and N_{heads} from 4 to 8. Meanwhile, our proposed MMRAN model can achieve the best performance with $N_{layers} = 6$ and $N_{heads} = 8$ as in (H. Wu et al., 2021). Moreover, we note that the MMRAN model with both the GRN and MAN components significantly outperforms MMRAN w/o MAN, suggesting that the additional MAN component with the multi-head attention mechanism further refines the hidden states (which are generated by the previous GRN component) by tracking the dialog states of the users' preferences among the multi-modal dialog sequences. The better effectiveness of MMRAN with both GRN and MAN added up suggests that MMRAN can benefit from their joint combination. Furthermore, we also observe that the MMRAN model significantly outperforms the transformer-based MMIT model, suggesting that adopting the hidden states of GRN as the representations of the past interactions is more effective than using the original textual and visual representations as the inputs of the transformer's multi-head attention.

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



(a) NDCG@10

(b) SR

Figure 5.6: Effects of the MAN's layers N_{layers} and the MAN's attention heads N_{heads} on our proposed MMRAN model at the 5th turn on *Fashion IQ Dresses*.

sequences, while addressing **Limitation 2**.

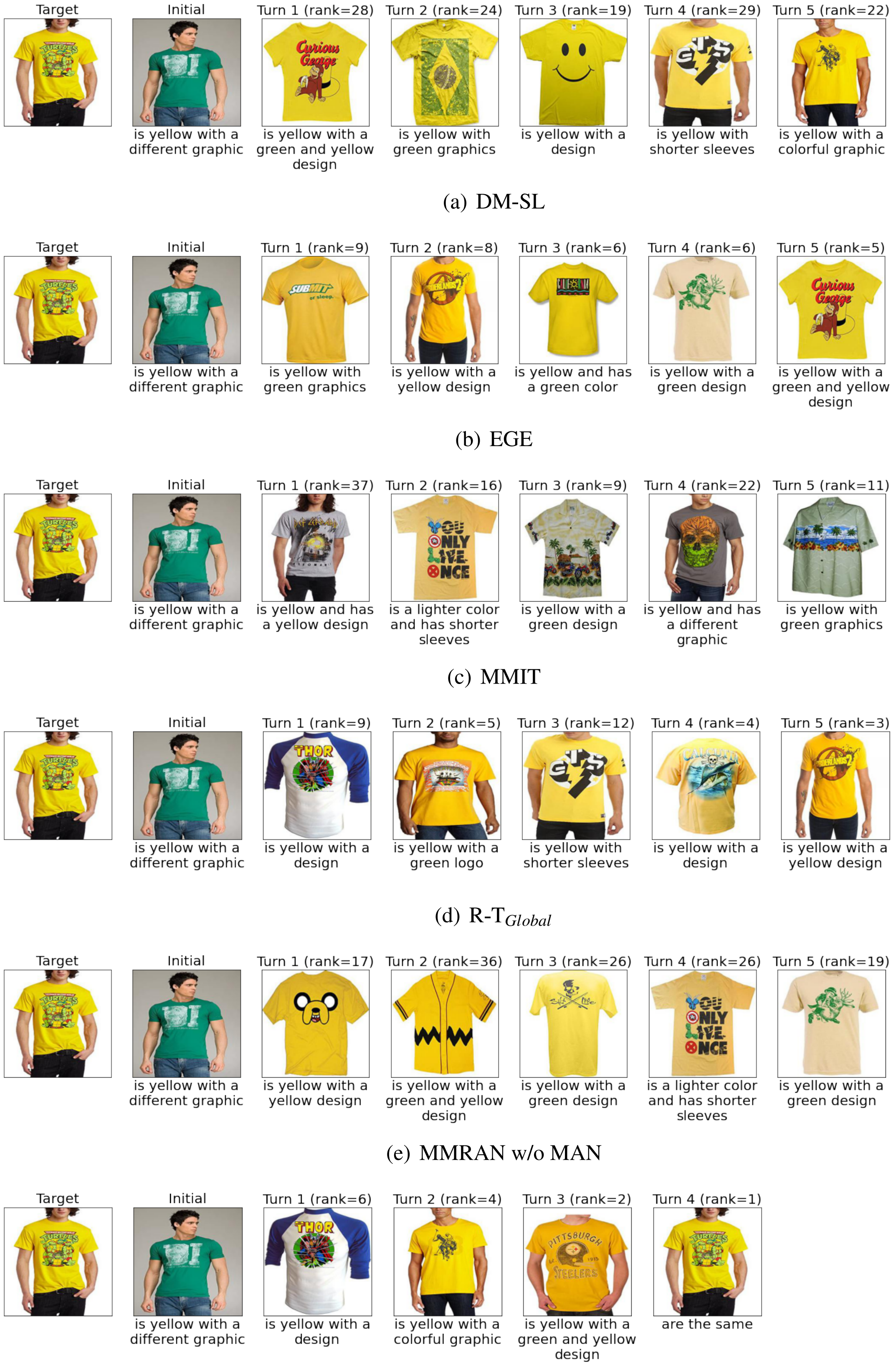
5.6.4 A Use Case

To consolidate the results observed in the chapter, we present a use case of multi-modal interactive recommendation in Figure 5.7 for the *Fashion IQ Shirts*. Figure 5.7 shows the interaction process for top-1 recommendation across the six tested models: (a) DM-SL, (b) EGE, (c) MMIT, (d) R-T$_{Global}$, (e) MMRAN w/o MAN and (f) MMRAN. For a fair comparison, the initial images are the same across the tested models given the target image from the testing set. When the target item is recommended, the rank is 1 (e.g. "Turn 1 (rank=1)" in Figure 5.7 (f)), and the user simulators will give the comment: "are the same". We observe that our proposed MMRAN model is the most effective among the tested models. In particular, DM-SL/EGE with only a GRU, MMIT with only a transformer and R-T$_{Global}$ based on a GRU and a transformer all fail to recommend the target item within 5 interaction turns, while our proposed MMRAN model needs only 4 interaction turns to recommend the target item. Furthermore, we also observe that the rank of the target item with the MMRAN w/o MAN model is relatively higher than the rank of DM-SL at the 5th interaction turn. In addition, both the MMRAN and MMRAN w/o MAN models can generally maintain a reasonable recommendation during the multi-turn interaction process with a shirt that is "yellow with a design". Though the recommendation with MMRAN at the 1st turn is not a really "yellow" shirt, it contains features from both the initial recommendation (i.e. the "green" colour) and the corresponding natural-language feedback (i.e. the "yellow" colour) in its "different graphic", while maintaining the highest rank of the target item (i.e. "rank=6") among all the tested models at the 1st turn. Indeed, the MMRAN model can better capture the "green and yellow design" features from the users' feedback than the other tested models. Similar results and observations were seen for the *Fashion IQ Dresses* and *Tops & Tees* datasets, but are omitted for reasons of space.

5.7 Conclusions

In this chapter, we proposed a novel multi-modal recurrent attention network (MMRAN) model for multi-modal interactive recommendation to effectively incorporate the users' preferences over time. Specifically, we leveraged a gated recurrent network (GRN) with a feedback gate to separately process the natural-language feedback and visual recommendations into hidden states (i.e. representations of the past interactions) for multi-modal sequence combination, as well as a multi-head attention network (MAN) to refine the previously generated hidden states by the GRN component to further track the dialog states of the users' preferences. Following previous work, we trained our MMRAN model by using a vision-language transformer-based user simulator (VL-Transformer), which itself is trained to describe the differences between the target users' preferences and the recommended items in natural language. Our experiments on three *Fashion IQ* datasets demonstrated that our proposed MMRAN model achieves significantly enhanced performances compared to the strongest baseline models on each used dataset - for instance, improvements of 24-28%, 17-20% and 18-22%, respectively. Our reported results showed that the MMRAN model benefits from the capability of GRN in combining multi-modal dialog sequences and from the MAN's structure to effectively track the dialog states. The experimental results and analysis provide support for the thesis statement with **Research Topic 2** in Section 1.3.

Next, in Chapter 6, we argue that the existing formulation of interactive recommender systems suffer from a coupling issue of policy optimisation and multi-modal composition representation learning. Therefore, we aim to address the coupling issue to effectively incorporate the users' preferences over time in multi-modal conversational recommendation by leveraging goal-oriented reinforcement learning and a composition network.


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

(a) DM-SL

(b) EGE

(c) MMIT

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

(e) MMRAN w/o MAN

(f) MMRAN

Figure 5.7: A use case for multi-modal interactive recommendation with different models on *Fashion IQ Shirts*.

Chapter 6

Coupling of Policy Optimisation & Representation Learning

In our thesis statement (stated in Section 1.3), we hypothesised that we can effectively incorporate the users' preferences over time, with a composition network and a multi-task learning approach, by decoupling the policy optimisation and the multi-modal composition representation learning with goal-oriented reinforcement learning. Therefore, in this chapter, we propose a novel 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. This chapter is mainly based on our work (Y. Wu, Macdonald, & Ounis, 2023) "Goal-Oriented Multi-Modal Interactive Recommendation with Verbal and Non-Verbal Relevance Feedback" published in the proceedings of the 17th ACM Conference on Recommender Systems (RecSys 2023)[1].

In the previous chapters, we have addressed the partial observability issue (see Chapter 4) and the multi-modal sequence dependency issue (see Chapter 5) by incorporating the users' preferences from both the users' natural-language feedback and the system's visual recommendations. In addition to the multi-modal information across the multi-turn interactions, the users can also express/indicate their preferences with other types of behaviours, such as likes and dislikes. In particular, 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 asso-

[1] DOI: `https://doi.org/10.1145/3604915.3608775`

ciated 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 (see Section 3.1), we train and evaluate our GOMMIR model by using user simulators (described in Section 3.3) that can generate natural-language feedback about the recommendations as a surrogate for real human users. Experiments conducted on four well-known fashion datasets (see Section 3.3.2) demonstrate that our proposed GOMMIR model yields significant improvements in comparison to the existing state-of-the-art baseline models (including the EGE model in Chapter 4 and the MMRAN model in Chapter 5). The results conform with our thesis statement with **Research Topic 3** in Section 1.3.

6.1 Motivations

As described in Section 3.1, the multi-modal interactive recommendation task (MMIR) usually involves information with various modalities, such as natural language and images. In addition to the visual-language modalities (as discussed in Chapters 4 & 5), users can indicate their positive/negative opinions by clicking like/dislike buttons when viewing a ranked list of visual recommendations (such as images of fashion products). To this end, in this chapter, we aim 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. Figure 6.1 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 (described in Section 2.4). Indeed, such approaches have demon-

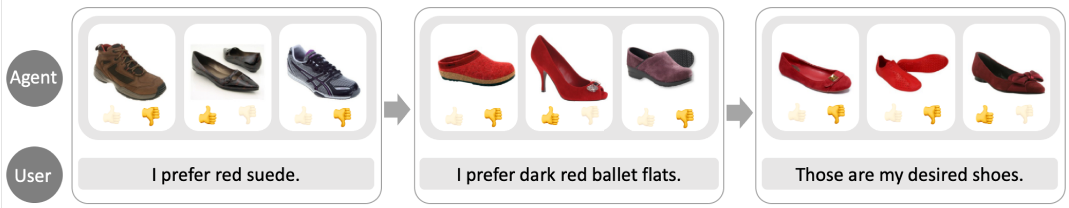


Figure 6.1: An example of multi-modal interactive recommendation with both verbal and non-verbal relevance feedback.

strated 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 (Guo et al., 2018; H. Wu et al., 2021)) 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) (Eysenbach et al., 2022). In particular, prior research often found that learning representations in an end-to-end fashion in DRL is usually unstable (Laskin, Lee, et al., 2020; Yarats et al., 2020) due to the coupling issue. Indeed, the policy optimisation processes of the existing DRL-based interactive recommendation models are 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 (Guo et al., 2018; H. Wu et al., 2021) for multi-modal feature composition between text (encoded with GloVe (Pennington et al., 2014) or BERT (Devlin et al., 2019a)) and image (encoded with ResNet (K. He et al., 2016)) 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) (Vo et al., 2019) and CLIP for Conditioned image retrieval (CLIP4Cir) (Baldrati, Bertini, Uricchio, & Del Bimbo, 2022a, 2022b)) have been recently proposed by various text-image retrieval models (Y. Chen, Gong, & Bazzani, 2020; Ge et al., 2021; Vo et al., 2019). We propose to leverage such approaches as an extra multi-modal composition representation learning task using multi-task learning (Laskin, Srinivas, & Abbeel, 2020) 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 (Afsar, Crump, & Far, 2022; X. Chen et al., 2021; Xin et al., 2020). However, the recommendation policy optimisation functions adopted by existing interactive recommendation agents (Guo et al., 2018; Y. Wu et al., 2021; R. Zhang et al., 2019) are mainly based on both (1) a sampled softmax (M. Chen et al., 2019) with randomly sampled negatives from the whole candidate pool (Guo et al., 2018; Y. Wu et al., 2021), and (2) an uninformative reward function that considers only the critiqued items (Guo et al., 2018; Y. Wu et al., 2021; R. Zhang et al., 2019) and/or a sparse reward function defined as a binary credit (success or fail) for reaching the desired item (R. Zhang et al., 2019). Due to the "goal-oriented" nature of the multi-modal interactive recommendation task, goal-oriented reinforcement learning (GORL) (Colas et al., 2022; M. Liu et al., 2022) 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 (M. Liu et al., 2022). In this chapter, 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 chapter, 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 (M. Liu et al., 2022) based on a policy gradient method (i.e. REINFORCE (M. Chen et al., 2019)) 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) (Vo et al., 2019) 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 (Laskin, Srinivas, & Abbeel, 2020) 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 Chapters 4 and 5, 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, see Section 3.3.2) demonstrate that our proposed model yields significant improvements in comparison to the existing state-of-the-art baseline models (including the EGE model in Chapter 4 and the MMRAN model in Chapter 5).

The main contributions of this chapter 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 (including the EGE model in Chapter 4 and the MMRAN model in Chapter 5).

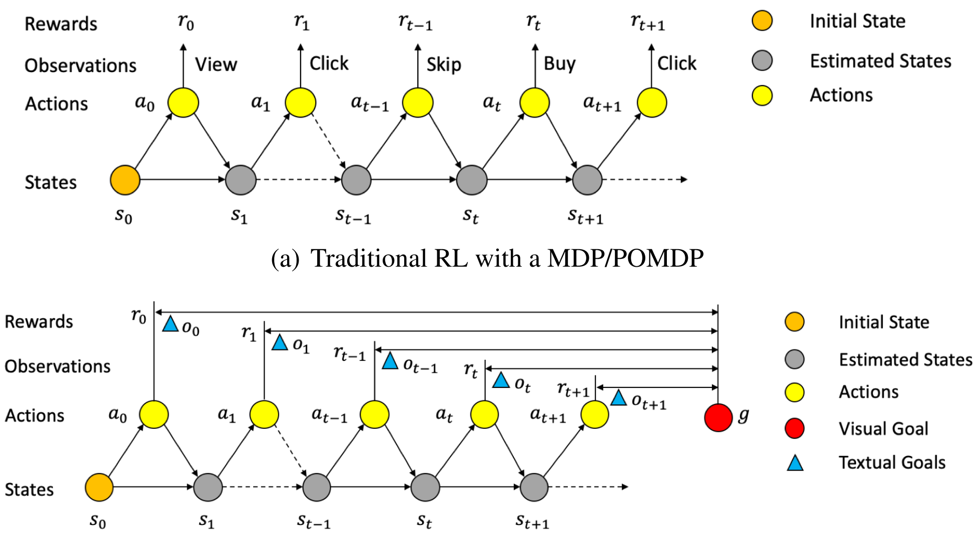
The remainder of the chapter is organised as follows: In Section 6.2, we review the related work, and position our contributions in comparison to the existing literature; Section 6.3 defines the problem formulation and presents our proposed GOMMIR model; Our experimental setup and results are presented in Sections 6.4 and 6.5, respectively; Section 6.6 summarises our findings.

6.2 Related Work

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

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) (Guo et al., 2018), partially observable Markov decision processes (POMDPs) (Y. Wu et al., 2021), constrained Markov decision processes (CMDPs) (R. Zhang et al., 2019) or multi-armed bandits (Yu et al., 2020) so as to effectively incorporate the users' information needs across multiple turns. However, the policy optimisation adopted by existing interactive recommendation agents (Guo et al., 2018; Y. Wu et al., 2021; R. Zhang et al., 2019) is generally ineffective due to random negative sampling (Guo et al., 2018; Y. Wu et al., 2021) and sparse/non-informative rewards (as discussed in Section 6.1). 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 (M. Liu et al., 2022). A goal is defined as "a cognitive representation of a future object" (Colas et al., 2022), 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, K. Wang et al. (2021) 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, D. Zhao et al. (2020) 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 (Batmaz et al., 2019; Deldjoo et al., 2022; Hu, Huang, Zhang, & Liu, 2022; X. Zhao, Zhang, et al., 2018). In particular, non-verbal relevance feedback is often used to model the users' behaviours and to indicate their preferences. For instance, X. Zhao, Zhang, et al. (2018) 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) (Gao et al., 2021; Jannach et al., 2021). For instance, existing conversational recommendation models either allow the users to describe their preferred attributes as positive feedback (Guo et al., 2018; Hu et al., 2022; Y. Sun & Zhang, 2018; H. Wu et al., 2021; Yu, Shen, & Jin, 2019; Yu et al., 2020; Y. Zhang, Chen, Ai, Yang, & Croft, 2018) (e.g., "I prefer dresses with longer sleeves.") or to provide disliked attributes as negative feedback (Y. Wu, Macdonald, & Ounis, 2022a) (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 (Bi, Ai, Zhang, & Croft, 2019; W. Lei et al., 2020; Xu et al., 2021). In this chapter, we consider both verbal (e.g., natural-language critiques) and non-verbal (e.g., likes/dislikes) relevance feedback



(a) Traditional RL with a MDP/POMDP

(b) GO-POMDP for MMIR

Figure 6.2: Traditional RL with a MDP/POMDP (Huang et al., 2022) and GO-POMDP for MMIR.

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 (see Section 3.1 and Chapters 4 & 5) 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 (Vo et al., 2019)) 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.

6.3 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 (Section 6.3.1). Next, in Section 6.3.2, 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 6.3.3).

6.3.1 Preliminaries

GO-POMDP for MMIR

Figure 6.2 (a) shows the traditional RL as a Markov decision process (MDP) or a partially observable Markov decision process (MDP) in formulating interactive/sequential recommendations (Afsar et al., 2022; X. Chen et al., 2021; Huang et al., 2022; Y. Lin et al., 2021). 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 (Huang et al., 2022). 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 6.2 (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 (Y. Wu et al., 2021), 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.

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 $(\mathscr{S}, \mathscr{A}, \mathscr{O}, \mathscr{T}, \mathscr{G}, r, \gamma)$ to describe the multi-modal interactive recommendation process, where: $\mathscr{S}$ is a continuous *state* space to describe the user states; $\mathscr{A}$ is a discrete *action* space that contains candidate items for recommendation; $\mathscr{O}$ 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; $\mathscr{T}$ is a set of conditional transition probabilities between states; $\mathscr{G}$ 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 $g \in \mathscr{G}$ by performing action $a \in \mathscr{A}$ at user state $s \in \mathscr{S}$; $\gamma \in [0,1]$ is the *discount factor* for future rewards.

Figure 6.3 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 ($a_{t,\leq K} = (a_{t,1},...,a_{t,K}) \in \mathscr{A}$) at each turn t. Meanwhile, the user provides non-verbal relevance feedback (e.g., likes/dislikes) and gives natural-language feedback ($o_t \in \mathscr{O}$) 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, $s_{t+1} \sim \mathscr{T}(s_{t+1}|s_t, o_t, a_{t,\leq K})$. 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).

6.3.2 The Model Architecture

Figure 6.4 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 6.3.3.

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 rec-

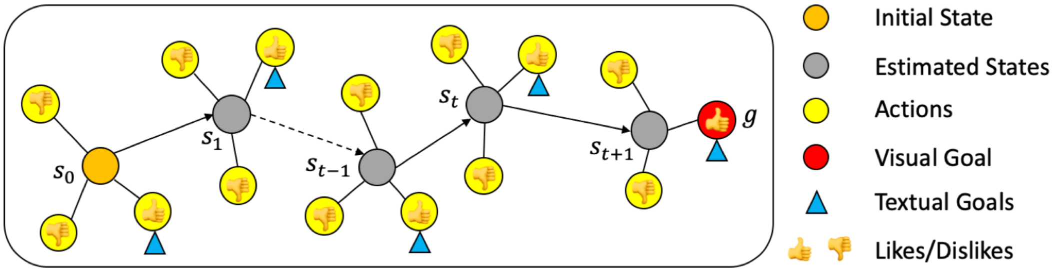


Figure 6.3: The goal-oriented interactive recommendation process with verbal & non-verbal relevance feedback for top-K recommendations.

ommendations 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 (Radford et al., 2021), for both image encoding and text encoding. Different from ResNet and GloVe/BERT for image and text encoding used by previous work in this task (see Section 3.1, and Chapters 4 and 5), 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:

$$o_t' = Norm(Linear(CLIP^{txt}(o_t))). \tag{6.1}$$

Similarly, given a liked image $a_{t,u}$ at the t-th turn, the encoded image representation is denoted by:

$$a_{t,u}' = Norm(Linear(CLIP^{img}(a_{t,u}))). \tag{6.2}$$

For simplicity of notation, we use a_t and o_t directly to denote their representations (i.e. a_t' and o_t'), respectively.

Composition Stage 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) (Vo et al., 2019)) to combine image and text representations 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 (Vo et al., 2019). The multi-modal composition feature $c_t = \psi(a_{t,u}, o_t)$ is computed by:

$$c_t = \psi(a_{t,u}, o_t) = \omega_g f_{gate}(a_{t,u}, o_t) + \omega_r f_{res}(a_{t,u}, o_t) \tag{6.3}$$

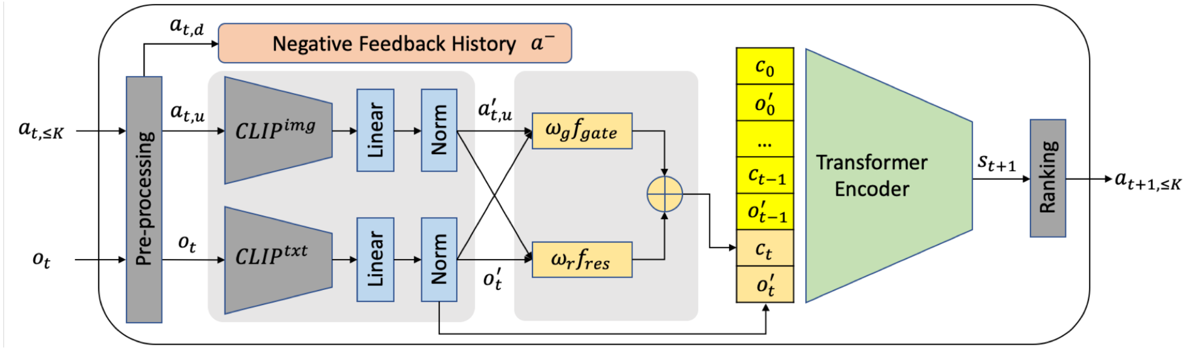


Figure 6.4: The proposed GOMMIR model for multi-modal interactive recommendations.

$$f_{gate}(a_{t,u}, o_t) = \sigma(W_{g2} * ReLU(W_{g1} * [a_{t,u}, o_t])) \odot a_{t,u} \tag{6.4}$$

$$f_{res}(a_{t,u}, o_t) = W_{r2} * ReLU(W_{r1} * [a_{t,u}, o_t]) \tag{6.5}$$

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 (H. Wu et al., 2021; Y. Wu, Macdonald, & Ounis, 2022a, 2022b), 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:

$$s_{t+1} = Linear(Tanh(Mean(TranEnc([c_{\le t}, o_{\le t}])))) \tag{6.6}$$

where $c_{\le t} = (c_0, ..., c_t)$ and $o_{\le t} = (o_0, ..., o_t)$ 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 (Guo et al., 2018; Y. Wu et al., 2021) 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: $a_{t+1,\le K} \sim KNNs(s_{t+1})$, where $KNNs(\cdot)$ is a softmax distribution over the top-K nearest neighbours of s_{t+1} and $a_{t+1,\le K} = (a_{t+1,1}, ..., a_{t+1,K})$. Furthermore, based on the interaction history $h_t = (o_{\le t}, a_{\le t, \le K})$, 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 (Y. Wu et al., 2021).

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 (H. Wu et al., 2021; Y. Wu, Macdonald, & Ounis, 2022a, 2022b). Meanwhile, we leverage the CLIP-based multi-modal encoders and a composition network (i.e. TIRG (Vo et al., 2019)) 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 (Guo et al., 2018; H. Wu et al., 2021; Y. Wu et al., 2021) (as described in Sections 6.1 & 6.2).

6.3.3 Learning Algorithm

We adopt a multi-task learning (Laskin, Srinivas, & Abbeel, 2020) approach for GO-POMDP to optimise the recommendation policy with a policy gradient method (e.g., REINFORCE (M. Chen et al., 2019)) 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 (Mnih et al., 2013)) have demonstrated many advantages in solving DRL problems, they are known to be prone to instability with value function approximations (M. Chen et al., 2019; Sutton, McAllester, Singh, & Mansour, 1999; Xin et al., 2020). Alternatively, policy-based methods (such as REINFORCE) are more stable given a sufficiently small learning rate (M. Chen et al., 2019) compared to value-based methods (such as DQN (Mnih et al., 2013)). Therefore, we rely on a policy gradient method (in particular REINFORCE) and enrich this on-policy method with goals for the MMIR task.

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:

$$\max_\theta J(\pi_\theta) = \max_\theta \mathop{\mathbb{E}}_{\tau \sim \pi_\theta}[R(\tau)] \tag{6.7}$$

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 τ = $((o_0, a_{0,\leq K})$, ..., $(o_T, a_{T,\leq K}))$.

We define the loss for optimising the recommendation policy based on the gradient of $J(\pi_\theta)$ with REINFORCE. Specifically, the gradient of Equation (6.7) can be computed as follows:

$$\nabla_\theta J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_{t,\leq K}|s_t) R(\tau)] \tag{6.8}$$

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:

$$\log \pi_\theta(a_{t,\leq K}|s_t) = \log(\frac{e^{\kappa(s_t,g)}}{e^{\kappa(s_t,g)} + \sum_{j=1}^{J} e^{\kappa(s_t,a_j^-)}}) \tag{6.9}$$

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 (Guo et al., 2018; H. Wu et al., 2021; Y. Wu et al., 2021). 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:

$$r(s_t, a_{t,\leq K}, g) = \sum_{i=1}^{K} \kappa(a_{t,i}, g) = \kappa(a_{t,u}, g) + \sum_{d=1}^{K-1} \kappa(a_{t,d}, g) \tag{6.10}$$

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.

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:

$$\max_\phi L(\psi_\phi) = \sum_{t=0}^{T} \max_\phi \, (0, l_2(c_t, g) - l_2(c_t, a^-) + \varepsilon_1) \tag{6.11}$$

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 (6.9). 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 6.1 & 6.2), as follows:

$$\max \mathscr{L}_{GOMMIR} = \max_{\theta} J(\pi_\theta) + \max_{\phi} L(\psi_\phi) \tag{6.12}$$

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 (Guo et al., 2018):

$$\max_{\theta} L(\pi_\theta) = \sum_{t=0}^{T} \max_{\theta} \left(0, l_2(s_t, g) - l_2(s_t, a^-) + \varepsilon_2\right) \tag{6.13}$$

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:

$$\max \mathscr{L}_{Pre-train} = \max_{\theta} L(\pi_\theta) + \max_{\phi} L(\psi_\phi) \tag{6.14}$$

Based on the pre-trained model obtained with $\mathscr{L}_{Pre-train}$, the joint loss objective $\mathscr{L}_{GOMMIR}$ 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.

6.4 Experimental Setup

In this section, we evaluate the effectiveness of our proposed GOMMIR model in comparison to the existing approaches from the literature. Figure 3.5 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 applied in Chapters 4 and 5, 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:

- RQ6.1: 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?
- RQ6.2: How do the components designed for composition representation learning and goal-oriented policy optimisation in the GOMMIR model affect the performance?

• RQ6.3: 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?

6.4.1 Datasets & Setup

Our proposed approaches are evaluated on four well-known fashion datasets, namely the *Shoes* and *Fashion IQ Dresses, Shirts, Tops & Tees* datasets (see Section 3.3.2), to verify the generalisation of the recommendation performance of our proposed GOMMIR model. Note that we leverage *Shoes* with the Images$_{origin}$ version for training/testing the recommender systems, and *Dresses*, *Shirt*, and *Tops & Tees* with the Images$_{caption}$ version that extracts unique images from the relevative captioning triplets for training/testing the recommender systems (see Section 3.3.2). The statistics of the four datasets are summarised in Table 3.1.

We pre-train our GOMMIR model with a multi-task supervised learning setting (as per Equation (6.14)) for initialisation, and then further optimise GOMMIR with a joint supervised and reinforcement learning setting with Equation (6.12)[2]. Following previous work (see Section 3.1), we use Adam (Kingma & Ba, 2014) with learning rates $\eta_1 = 10^{-3}$ and $\eta_2 = 10^{-5}$ with Equation (6.14) and Equation (6.12), respectively, for optimising the GOMMIR model's parameters. The similarity kernel $\kappa(\cdot)$ in Equation (6.9) 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[3]. The batch size is set to 128 and the number of negative samples (i.e. J) is set to 5. The maximum number of epochs for training is 20. 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.

6.4.2 Online Evaluation

An interactive recommender system is a type of closed-loop system (see Section 2.2) 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

[2] The code and datasets for this chapter are publicly available in `https://github.com/yashonwu/gommir` [3] `https://github.com/openai/CLIP`

models[4] (i.e. the Show, Attend, & Tell (Xu et al., 2015) model on *Shoes* and the VL-Transformer (see Section 3.3) model on *Fashion IQ Dresses*, *Shirts*, and *Tops & Tees*) as a surrogate for real human users (a.k.a. user simulators), as in (Guo et al., 2018; Y. Wu et al., 2021; R. Zhang et al., 2019). We assume the user desires a visual item and gives both 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[5] (Berg et al., 2010; Guo et al., 2018) for *Shoes* (provided by Guo et al. (2018)), following the setting in Guo et al. (2018), while we use the user simulator checkpoints for *Fahion IQ Dresses, Shirts*, and *Tops & Tees* (provided by Y. Wu, Macdonald, and Ounis (2022b)) following the setting in Y. Wu, Macdonald, and Ounis (2022b). 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.

6.4.3 Evaluation Metrics

We measure the effectiveness of different interactive recommendation models under the two evaluation metrics (see Section 3.2.2): Normalised Discounted Cumulative Gain (NDCG) and Success Rate (SR). In our experiments, we consider NDCG@N, which is truncated at rank $N = 3$ and $N = 10$ and we report the interaction turn $M \in [1, 10]$. 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.

[4] These user simulators were used by the original authors - we replicate their user simulator setups. [5] `https://github.com/XiaoxiaoGuo/fashion-retrieval`

6.4.4 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 6.3.2.

- **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**[6] (Vo et al., 2019): 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 6.3.2.
- **CLIP4Cir**[7] (Baldrati et al., 2022a, 2022b): CLIP4Cir adopts a Combiner network (Baldrati et al., 2022b) 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 (Chapter 4) and DEERS (X. Zhao, Zhang, et al., 2018) are the two baseline models that use DRL algorithms.

- **DM**[8] (Guo et al., 2018): 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**[9] (H. Wu et al., 2021): 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

[6] https://github.com/google/tirg [7] https://github.com/ABaldrati/CLIP4Cir [8] https://github.com/XiaoxiaoGuo/fashion-retrieval [9] https://github.com/XiaoxiaoGuo/fashion-iq

recommendations.

• **MMRAN** (Chapter 5): 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** (Chapter 4): 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** (X. Zhao, Zhang, et al., 2018): 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.

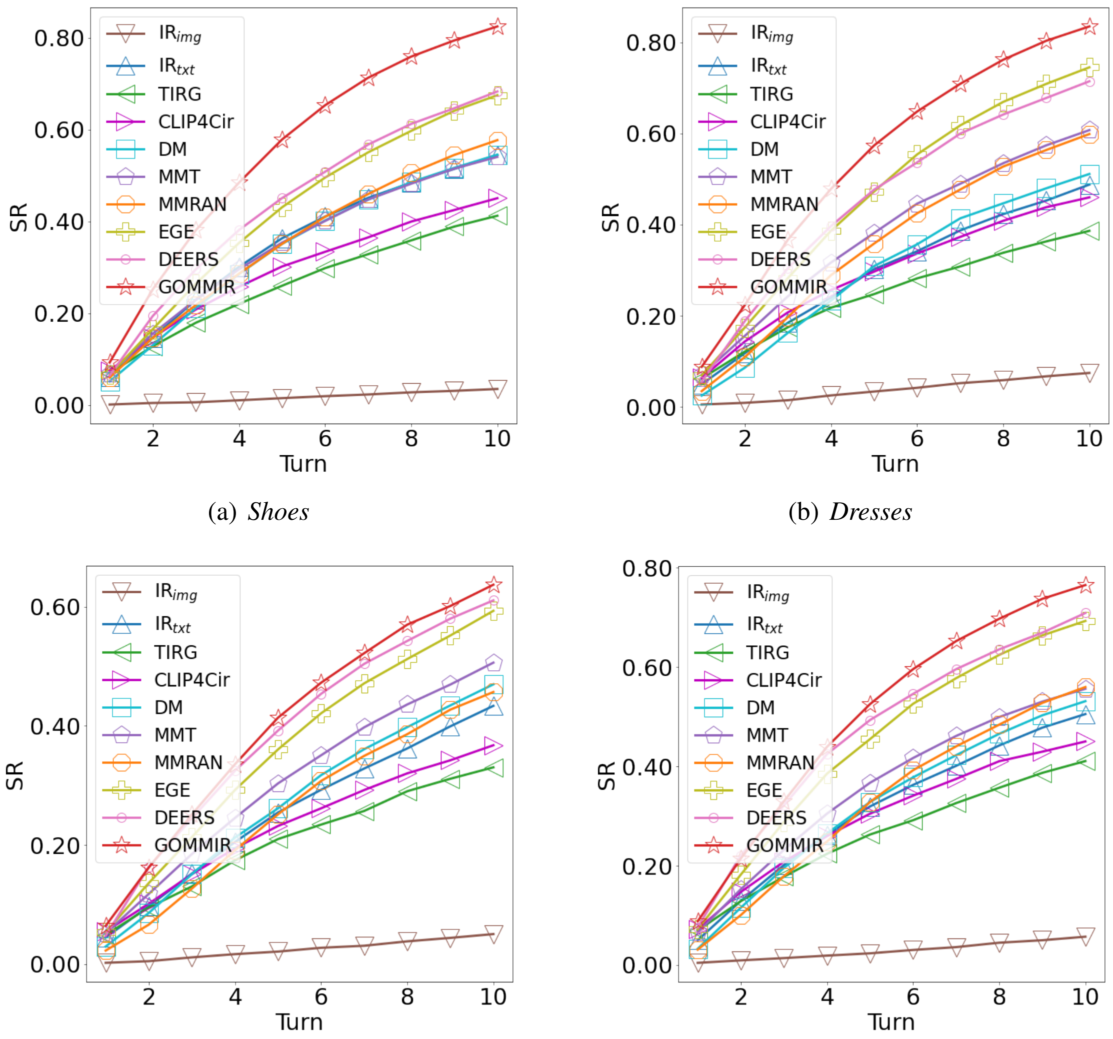
In addition to the above baseline models for the MMIR task, the GOMMIR variants used for the ablation studies (in Section 6.5.2) can also act as strong baselines. For fair comparisons, all of the tested baseline models use CLIP (using the "RN101" checkpoint) for providing the texts and image representations (as described in Section 6.3.2). 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 (W. Lei et al., 2020; Y. Sun & Zhang, 2018), requiring additional attributes of items for learning (Yu, Shen, Zhang, et al., 2019; Yuan & Lam, 2021; R. Zhang et al., 2019) or requiring multi-modal knowledge graph for reasoning (Y. Wu, Liao, et al., 2022).

6.5 Experimental Results

In this section, we analyse the experimental results with respect to the three research questions stated in Section 6.4 to gauge the effectiveness of our proposed GOMMIR model. Specifically, we address the overall effectiveness of our proposed GOMMIR model for the MMIR task (RQ6.1, Section 6.5.1), the impact of the goal-oriented policy optimisation and composition representation learning (RQ6.2, Section 6.5.2), and the effects of the hyper-parameters (RQ6.3, Section 6.5.3). To consolidate our findings, similar to previous chapters, we provide a use case from the logged experimental results in Section 6.5.4.

6.5.1 Performance Comparison (RQ6.1)

Figure 6.5 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



(a) *Shoes*

(b) *Dresses*

(c) *Shirts*

(d) *Tops & Tees*

Figure 6.5: Comparison of the recommendation effectiveness at various interaction turns with top-3 recommendation.

turns on the *Shoes*, *Fashion IQ Dresses*, *Shirts* and *Tops & Tees* datasets. Comparing the results in Figure 6.5, 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 6.1 reports their performances at the 10th interaction turn. The best results of the baseline models and the best overall results are underlined and high-

Table 6.1: 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.

	Shoes			Dresses			Shirts			Tops & Tees		
Models	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

lighted 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 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 RQ6.1, 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.

6.5.2 Impact of Components (RQ6.2)

To address RQ6.2, we investigate the impact of the components designed for both composition representation learning and goal-oriented policy optimisation to tackle the coupling issue. Table 6.2 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 6.2 (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 6.2 show that the full GOMMIR model (i.e. considering the above four aspects in the second part of Table 6.2) 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:

- (1) the hard negative sampling $a^-_{d,j}$ in Equation (6.9),
- (2) the following relevance feedback $a_{t,d}$ in hard negative sampling $a^-_{d,j}$,
- (3) the goal-oriented rewards $r(s_t, a_{t,\leq K}, g)$ in Equation (6.8), and
- (4) the extra rewards of the disliked items $r_{t,d}$ in Equation (6.10).

Table 6.2 (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 (6.9)" 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 (6.8)" optimises the recommendation policy using supervised learning without the goal-oriented rewards. "GOMMIR w/o $r_{t,d}$ in Equation (6.10)" 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 6.2 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 (6.10)" 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 RQ6.2, we find that our proposed GOMMIR model can benefit from explicitly learning the composition representation with an advanced composition network (i.e. TIRG) and

Table 6.2: 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. (6.9)	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. (6.10)	0.8128*	0.8305	0.6369	0.7614

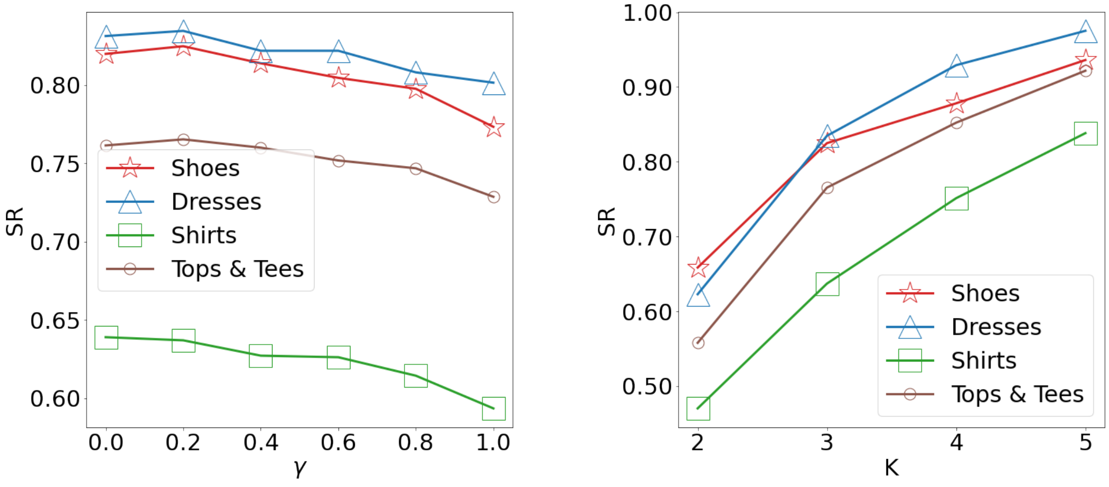
optimising the recommendation policy with hard negative sampling and rewards based on the non-verbal relevance feedback.

6.5.3 Impact of Hyper-Parameters (RQ6.3)

To address RQ6.3, Figure 6.6 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. 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 6.6 (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 6.6 (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



(a) γ for *SR*

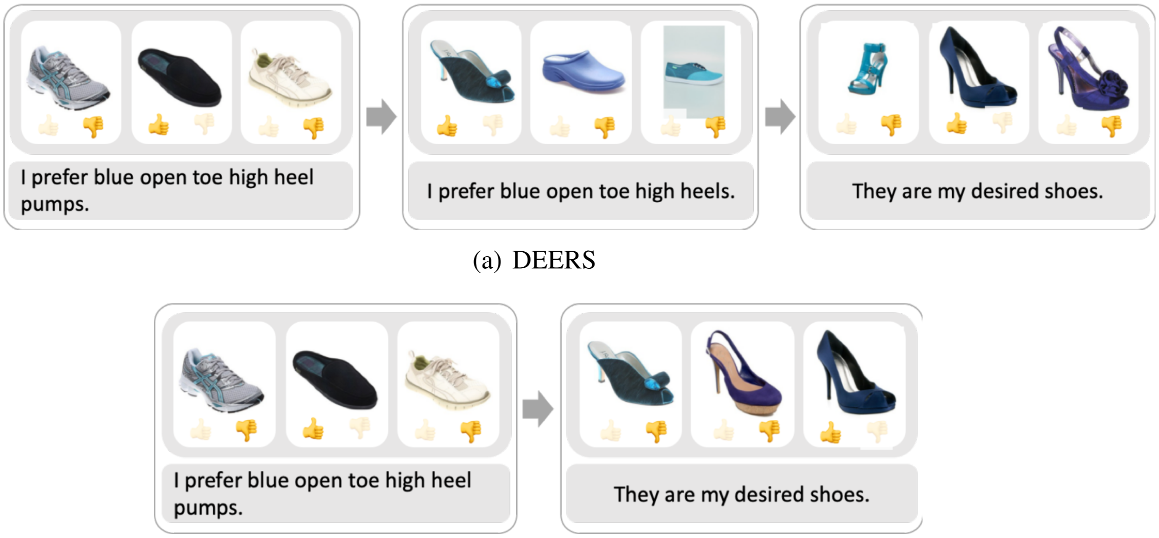
(b) K for *SR*

Figure 6.6: Comparison of the recommendation effectiveness at 10th turn with different γ and K values.

items are exposed to the users and users provide more feedback. Overall, in response to RQ6.3, we find that a lower reward discount factor γ and more exposed top-K items can improve the effectiveness of our GOMMIR model.

6.5.4 Use Case

In this section, we present a use case of the multi-modal interactive recommendation on the *Shoes* dataset in Figure 6.7. 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 6.7 (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 6.5.1. 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 6.7 (a) and (b), the user 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 chapter because of space constraints.


I prefer blue open toe high heel pumps.
I prefer blue open toe high heels.
They are my desired shoes.
I prefer blue open toe high heel pumps.
They are my desired shoes.

(a) DEERS

(b) GOMMIR

Figure 6.7: Example use cases for the interactive recommendation with DEERS and GOMMIR on *Shoes*.

6.6 Conclusions

In this chapter, 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 (see Section 3.1 and Chapters 4 & 5), 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. The experimental results and analysis provide support for the thesis statement with **Research Topic 3** in Section 1.3.

Next, in Chapter 7, we argue that the existing formulation of interactive recommender systems are 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. Therefore, we aim to effectively incorporate both the users' long-term preferences and short-term needs into the personalised recommendations by modelling the multi-modal conversational recommendation process with both the users' interaction history and the users' instant natural-language feedback.

Chapter 7

Personalisation for Cold-Start & Warm-Start Users

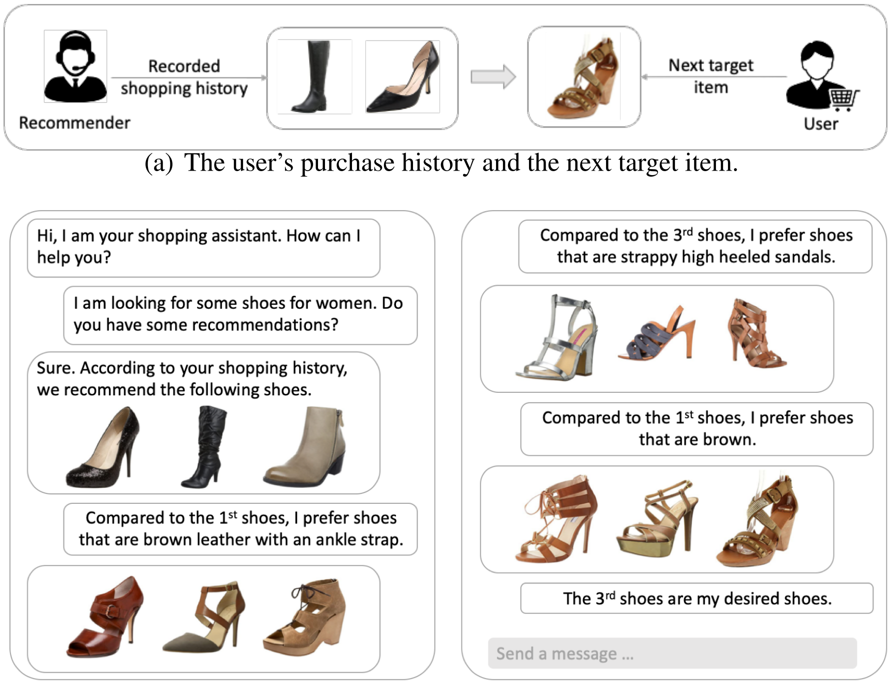
In our thesis statement (as stated in Section 1.3), we hypothesised that we can effectively incorporate both the users' long-term preferences and short-term needs into the personalised recommendations by modelling the multi-modal conversational recommendation process with both the users' interaction history and the users' instant natural-language feedback. Therefore, in this chapter, 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. This chapter is mainly based on our work "Personalised Multi-Modal Interactive Recommendation with Hierarchical State Representations" that has been submitted to TORS and is currently accepted.

In the previous chapters (see Chapters 4, 5, & 6), we have formulated the multi-modal conversational recommendation task with cold-start users without interaction history. In the real-world 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 (see Section 3.2). 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 chapter, 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 (described in Section 3.1), we train and evaluate our PMMIR model using a user simulator (described in Section 3.3) 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 (including our previously proposed EGE model in Chapter 4). The results conform with our thesis statement with **Research Topic 4** in Section 1.3.

7.1 Motivations

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 real-time feedback (often natural-language critiques) according to the quality of the recommendations (described in Section 3.1). In the multi-modal interactive recommendation (MMIR) scenario addressed by this chapter, 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 7.1 shows an example of the personalised multi-modal interactive recommendation with visual recommendations and the corresponding natural-language critiques. Different from the previous chapters (Chapters 4, 5, & 6) that initiate a conversation with randomly sampled recommendation items, the personalised multi-modal interactive recommendation task starts with a personalised initial recommendations based the users' past shopping history. In particular, Figure 7.1 (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 7.1 (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 advances in incorporating the users' current needs (i.e. the target items) from the informative multi-modal information across the multi-turn interactions, exemplified in Chapters 4, 5, & 6, we argue that it is 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



Hi, I am your shopping assistant. How can I help you?

I am looking for some shoes for women. Do you have some recommendations?

Sure. According to your shopping history, we recommend the following shoes.

Compared to the 1st shoes, I prefer shoes that are brown leather with an ankle strap.

Compared to the 3rd shoes, I prefer shoes that are strappy high heeled sandals.

Compared to the 1st shoes, I prefer shoes that are brown.

The 3rd shoes are my desired shoes.

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

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

Figure 7.1: An example of the personalised multi-modal interactive recommendation.

(as used in the previous previous chapters) 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 (Hidasi & Karatzoglou, 2018; Hidasi et al., 2016; Kang & McAuley, 2018; F. Sun et al., 2019) 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) (J. Wang, Ding, & Caverlee, 2021; Y. Zheng, Liu, Li, & Wu, 2021). 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 (Hidasi et al., 2016)) generates the initial personalised recommendations and a multi-modal interactive recommendation model (Chapter 4) 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 (Xin et al., 2020), video (M. Chen, Chang, Xu, & Chi, 2021) and music recommendations (W. Lei et al., 2020). In particular, recent research on multi-modal interactive recommendation (MMIR) has formulated the MMIR task with various DRL algorithms as MDPs (Guo et al., 2018), POMDPs (Y. Wu et al., 2021), CMDPs (R. Zhang et al., 2019) or multi-armed bandits (Yu et al., 2020). However, none of these have been adapted for a personalised recommendation scenario. 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) (Hutsebaut-Buysse et al., 2022; Pateria et al., 2021) 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 (Hutsebaut-Buysse et al., 2022). However, to the best of our knowledge, no prior work has investigated HRL in the multi-modal interactive recommendation task.

In this chapter, 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 (namely REINFORCE (M. Chen et al., 2019)) with a single optimisation objective by maximising the users' future satisfaction with the recommendations (i.e. the cumulative future rewards). Like Chapters 4, 5, & 6, our PMMIR model is trained and evaluated by adopting a user simulator, which is capable of producing natural-language critiques regarding the recommendations. By conducting experiments on two fashion 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 chapter 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 (including the EGE model in Chapter 4). We also show that both cold-start and warm-start users can benefit from our proposed PMMIR model in terms of recommendation effectiveness.

The chapter is structured as follows: Section 7.2 provides a comprehensive review of the related work and highlights the contributions of our research in relation to the existing literature. In Section 7.3, we define the problem formulation and introduce our proposed PMMIR model. The experimental setup and results are presented in Sections 7.4 and 7.5, respectively. Finally, Section 7.6 summarises our findings.

7.2 Related Work

Within this section, we discuss personalisation in interactive recommendation. Then, we describe hierarchical reinforcement learning.

Personalisation in Interactive Recommendation The 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 W. Lei et al. (2020)

(a typical question-based interactive recommendation model (Gao et al., 2021)) leveraged the factorisation machine (FM) (Rendle, 2010) 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 (Hidasi et al., 2016)) with the multi-modal interactive recommendation models (such as EGE (Y. Wu et al., 2021)) 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 (Jannach, Quadrana, & Cremonesi, 2022; Latifi, Mauro, & Jannach, 2021; Quadrana, Karatzoglou, Hidasi, & Cremonesi, 2017; S. Wang et al., 2021) 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. (Quadrana et al., 2017) 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 (Afsar et al., 2022; X. Chen et al., 2021; Y. Lin et al., 2021).

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) (Mnih et al., 2013), REINFORCE (Williams, 1992), and Actor-Critic (Konda & Tsitsiklis, 2000), for coping with the users' dynamic preferences over time and maximising their long-term engagements (Afsar et al., 2022; X. Chen et al., 2021; Y. Lin et al., 2021). In particular, the MMIR task has been formulated with various DRL algorithms as MDPs (Guo et al., 2018), POMDPs (Y. Wu et al., 2021), CMDPs (R. Zhang et al., 2019) or multi-armed bandits (Yu et al., 2020) to simulate the multi-turn interactions between the recommender systems and the users. However, the existing MMIR models (e.g., MBPI (Guo et al., 2018), EGE (Chapter 4), and RCR (R. Zhang et al., 2019)) 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 (Sutton, Precup, & Singh, 1999), Hierarchical of Abstract Machines (HAMs) (Parr & Russell, 1997), and MAXQ value function decomposition (Dietterich, 2000). The existing recommender systems with HRL (Greco et al., 2017; Y. Lin et al., 2022; Xie et al., 2021; D. Zhao et al., 2020) 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 (Greco et al., 2017) 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 (W. Lei et al., 2020) and FPAN (Xu et al., 2021)) 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) (Rendle, 2010) as a controller to generate a set of recommendations (Gao et al., 2021). Different from the standard HRL models, these question-based conversational recommendation models (Gao et al., 2021; W. Lei et al., 2020; Xu et al., 2021) only optimise the meta-controller with RL algorithms (such as REINFORCE (Williams, 1992)) to manage the conversational system, while the controller is separately optimised with supervised learning approaches (such as BPR (Rendle, Freudenthaler, Gantner, & Schmidt-Thieme, 2012)). However, to the best of our knowledge, no prior work has investigated HRL in the multi-modal interactive recommendation task. In this chapter, 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 7.3). 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.

7.3 The PMMIR Model

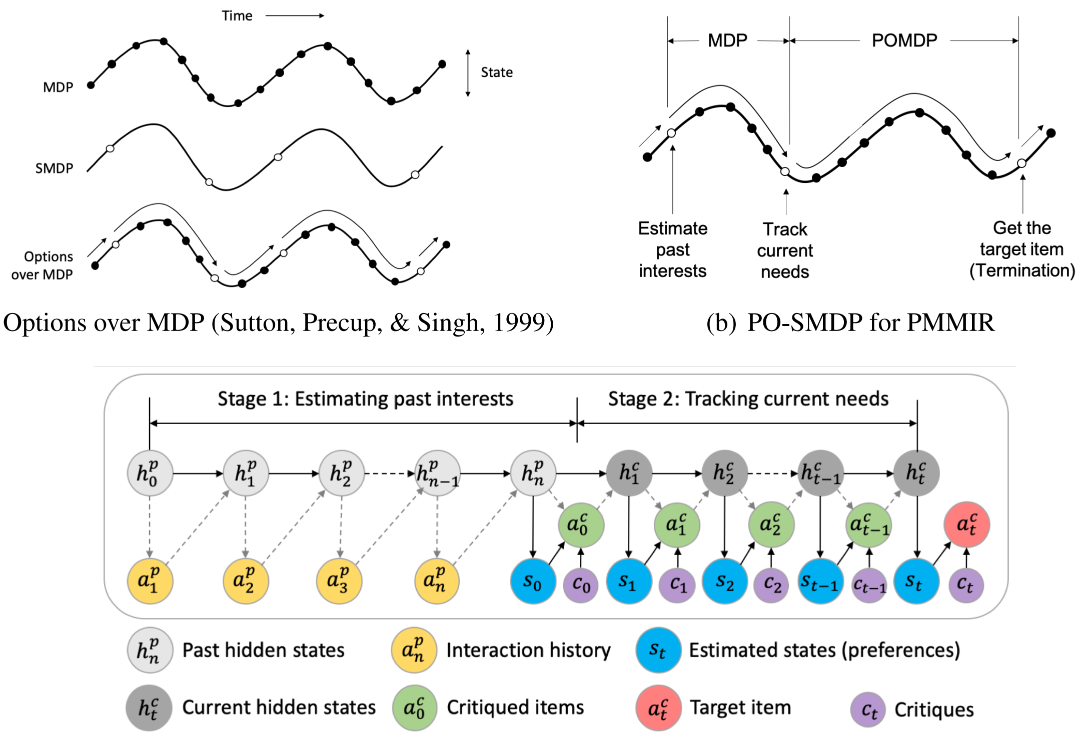
In this section, we begin by formulating the problem of the multi-modal interactive recommendation task using hierarchical reinforcement learning within the framework of partially observable semi-Markov decision processes (PO-SMDP) and we introduce the notations used in our formulation (Section 7.3.1). Then, in Section 7.3.2, 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 7.3.3).

7.3.1 Preliminaries

This chapter focuses on investigating the *personalised* multi-modal interactive recommendation (MMIR) task within a hierarchical reinforcement learning (HRL) formulation, specifically utilising the Options framework (Sutton, Precup, & Singh, 1999) 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 (Chapter 4). Figure 7.2 (b) & (c) illustrate the state transition process with hierarchical state representations for the personalised MMIR task.

PO-SMDP for Personalised MMIR

Figure 7.2 (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 (Sutton, Precup, & Singh, 1999)) 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, Precup, and Singh (1999) 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 chapter, we adopt a partially observable semi-Markov decision process (PO-SMDP, as shown in Figure 7.2 (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



(a) Options over MDP (Sutton, Precup, & Singh, 1999)

(b) PO-SMDP for PMMIR

(c) State Transition Process

Figure 7.2: State transition process with hierarchical state representations for the personalised MMIR task.

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). 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 7.1. 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 7.1. 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.

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 $(\mathscr{S},\mathscr{A},\mathscr{C},\mathscr{O},\mathscr{R},\mathscr{T},\mathscr{P},\gamma)$, where:

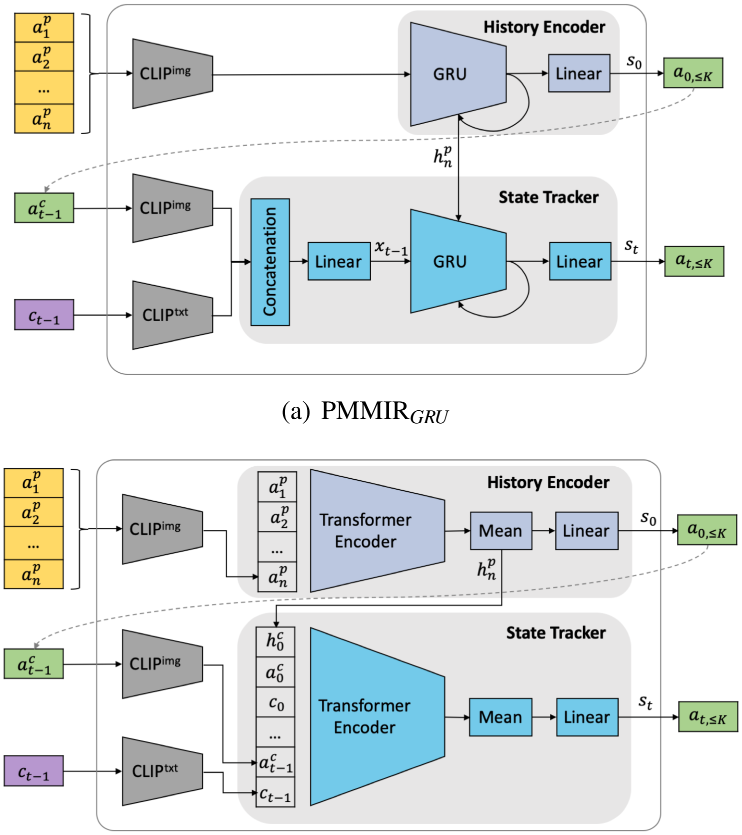
- $\mathscr{S}$ is a set of *states* (i.e. the users' preferences),
- $\mathscr{A}$ is a set of *actions* (i.e. the items for recommendations),

- $\mathscr{C}$ is a set of *observations* (i.e. the users' natural-language critiques),
- $\mathscr{O}$ is a set of *options* (i.e. options for selecting subtasks, either estimating past interests or tracking current needs),
- $\mathscr{R}$ is the *reward function*,
- $\mathscr{T}$ is a set of transition probabilities between states,
- $\mathscr{P}$ 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 $s_t \in \mathscr{S}$. When the recommender system (i.e. the agent) provides a ranking of K items, $a_t \in \mathscr{A}$ $(a_{t,\leq K} = (a_{t,1},...,a_{t,K}))$ and receives a natural-language critique $c_t \in \mathscr{C}$ and a reward $r_t \sim \mathscr{R}(s_t, a_t)$, the estimated preferences s_t change in accordance with the transition distribution, $s_{t+1} \sim \mathscr{T}(s_{t+1}|s_t, a_t, c_t)$. 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 7.2 (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 τ = (($a_{0,\leq K}, c_0$), ..., ($a_{T,\leq K}, c_T$)) 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 their 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.

7.3.2 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 modeling 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 7.3 shows our proposed end-to-end



(a) PMMIR$_{GRU}$

(b) PMMIR$_{Transformer}$

Figure 7.3: The proposed personalised multi-modal interactive recommendation (PMMIR) model with hierarchical state representations.

personalised multi-modal interactive recommendation model (PMMIR) with hierarchical state representations based on GRUs (Figure 7.3 (a) with PMMIR$_{GRU}$) and Transformers (Figure 7.3 (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 (Radford et al., 2021), as the unified visual and textual representations. There are also other alternatives for the multi-modal encoders (Guo et al., 2018; H. Wu et al., 2021), for instance the pre-trained language models (such as GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2019a)) for text and the pre-trained vision models (such as ResNet (K. He et al., 2016) and ViT (Dosovitskiy et al., 2020)) for images. Compared to these alternative encoders, CLIP has the capability of providing a single representation vector for each modality with the same dimensionality. 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) (Chung et al., 2014) as the history encoder (similar to the GRU4Rec (Hidasi et al., 2016) model for sequential recommendations) for encoding the past hidden states as follows:

$$h^{p}_{n} = GRU^{past}(h^{p}_{n-1}, a^{p}_{n}) \tag{7.1}$$

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 (see Section 2.1) as the history encoder (similar to the SASRec (Kang & McAuley, 2018) 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:

$$h^{p}_{n} = Mean(Transformer^{past}(a^{p}_{1:n})) \tag{7.2}$$

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 (Guo et al., 2018; H. Wu et al., 2021) and then a state tracker (either based on a GRU (Guo et al., 2018) or a Transformer encoder (H. Wu et al., 2021)) 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.

$$x_{t-1} = Linear([a^{c}_{t-1}, c_{t-1}]) \tag{7.3}$$

$$h^{c}_{t} = GRU^{current}(h^{c}_{t-1}, x_{t-1}) \tag{7.4}$$

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. $s_t = Linear(h_t^c)$).

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

$$h_t^c = Mean(Transformer^{current}([h_0^c, a_0^c, c_0, ..., a_{t-1}^c, c_{t-1}])) \tag{7.5}$$

The last hidden state h_n^p of the history encoder $Transformer^{past}(\cdot)$ is concatenated as the input of $Transformer^{current}(\cdot)$, that is $h_0^c = h_n^p$. In addition, the hidden state h_t^c at each turn t ($t \neq 0$) is further mapped with a linear layer into the estimated users' current needs (i.e. $s_t = Linear(tanh(h_t^c))$).

Considering the estimated state s_t representing the user's preferences, we adopt a greedy policy (Guo et al., 2018; Y. Wu et al., 2021) 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 Chapter 4.

7.3.3 The Learning Algorithm

To optimise PMMIR, we leverage a two-stage optimisation method following (Guo et al., 2018) with a supervised learning (SL) loss for initialising the policies and then a reinforcement learning (RL) loss for further improving the performances.

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 (Guo et al., 2018) to jointly pre-train the recommendation policies π_ϕ (for estimating the past interests) and π_θ (for tracking the current needs):

$$\max L(\pi_\phi, \pi_\theta) = \sum_{t=0}^{T} \max\ (0, l_2(s_t, a^+) - l_2(s_t, a^-) + \varepsilon) \tag{7.6}$$

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.

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:

$$\max J(\pi_\phi, \pi_\theta) = \max \mathop{\mathbb{E}}_{\tau \sim \pi_\phi, \pi_\theta} [R(\tau)], \text{ where} R(\tau) = \sum_{t=0}^{T} \gamma^t r(s_t, a_{t,\leq K}) \tag{7.7}$$

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

We adopt a policy gradient method (e.g., REINFORCE (Williams, 1992)) 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 (M. Chen et al., 2019) compared to the value-based methods (such as DQN (Mnih et al., 2013)). Specifically, the gradient of Equation (7.7) can be computed as follows:

$$\nabla J(\pi_\phi, \pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\phi, \pi_\theta} [\sum_{t=0}^{T} \nabla \log \pi(a_{t,\leq K}|s_t) R(\tau)] \tag{7.8}$$

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^- ($j \in [1, J]$)):

$$\log \pi(a_{t,\leq K}|s_t) = \log(\frac{e^{sim(s_t, a^+)}}{e^{sim(s_t, a^+)} + \sum_{j=1}^{J} e^{sim(s_t, a_j^-)}}) \tag{7.9}$$

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

Finally, 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:

$$r(s_t, a_{t,\leq K}) = \sum_{i=1}^{K} sim(a_{t,i}, a^+) \tag{7.10}$$

Training Procedure

We also present the training procedure of our PMMIR model for PO-SMDP with REINFORCE in Algorithm 7.1. To facilitate the training processes, a user simulator (see Section 3.3) is adopted as a substitute for real human users. Further information regarding the specific user simulator employed is discussed in Section 7.4.2. As shown in Algorithm 7.1, the recommender policies π_ϕ and π_θ aim to maximise the expected rewards by properly cooperating with each other.

Algorithm 7.1 Training procedure of PMMIR

Input: User-item interaction sequence set $\mathscr{X}$, 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 $\pi_\phi$ & $\pi_\theta$ with Eq. (7.6)
Load all parameters with weights from pre-training
repeat
    Draw a batch of $(a^p_{1:n}, a^{target})$ from $\mathscr{X}$
    Start with $\pi_\phi$ for estimating the past interests
    Generate $h^p_n$ from $a^p_{1:n}$ with Eq. (7.1)/Eq. (7.2)
    Map $h^p_n$ into $s_0$
    Switch into $\pi_\theta$ 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. (7.10)
        if t==0 then
            Calculate $\log \pi(s_t, a_t; \phi)$ with Eq. (7.9)
        else
            Calculate $\log \pi(s_t, a_t; \theta)$ with Eq. (7.9)
        end if
        Estimate and update next state $s_{t+1}$
    end for
    Calculate $R(\tau)$ with Eq. (7.7)
    Perform updates by $\nabla J(\pi_\phi, \pi_\theta)$ with Eq. (7.8)
until converge
return all parameters of policies $\phi$, and $\theta$
```

7.4 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:
• RQ7.1: 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?
• RQ7.2: Can both cold-start and warm-start users benefit from our proposed PMMIR model?
• RQ7.3: What are the impacts of the components of the PMMIR model (such as $h_0^c = h_n^p$ and CLIP backbones) and the introduced hyper-parameters (such as γ & K) on the overall performance?

7.4.1 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 7.1 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* and *Fashion IQ Dresses* datasets (discussed in Section 3.3.2) 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

[1] `http://jmcauley.ucsd.edu/data/amazon/index_2014.html`

[2] `https://nijianmo.github.io/amazon/`

Table 7.1: 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

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.

Setup As described in Algorithm 7.1, we leverage a two-stage training procedure for optimising the PMMIR model[3] following (Guo et al., 2018). In particular, we first pre-train and initialise the PMMIR model with the supervised learning (SL) setting using a learning rate $\eta_{sl} = 10^{-3}$ (Guo et al., 2018) and then further optimise the PMMIR model in the reinforcement learning (RL) setting using a learning rate $\eta_{rl} = 10^{-5}$ (Guo et al., 2018). We use Adam (Kingma & Ba, 2014) with Eq. (7.6) and Eq. (7.8) for optimising the PMMIR model's parameters, respectively. The pre-trained CLIP image and text encoders are loaded with the "ViT-B/32" checkpoint[4], and the visual and textual embedding dimensionalities of the multi-modal feature space are both set to 512. The batch size is set to 128 following the setting in (Guo et al., 2018). The maximum number of epochs for SL & RL training is set to 20 with early stopping, while the maximum number of interaction turns is set to 10. 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 (Y. Wu et al., 2021). The similarity kernel $sim(\cdot)$ in Equation (7.9) 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).

7.4.2 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 (mentioned in Section 3.2.2). Figure 3.5 shows an example of online evaluation with top-K (e.g., $K = 3$) recommendation across multi-turn interactions in

[3] The code and datasets for this chapter are publicly available in `https://github.com/yashonwu/pmmir`

[4] `https://github.com/openai/CLIP`

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 Chapters 4 & 5, 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 (see Section 3.3). Indeed, 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 (Xu et al., 2015) model trained with triplets from *Shoes* by using the checkpoint[5] (Berg et al., 2010; Guo et al., 2018) provided by Guo et al. (Guo et al., 2018). In addition, we adopt the VL-Transformer model introduced in (H. Wu et al., 2021; Y. Wu, Macdonald, & Ounis, 2022b) as a user simulator, specifically trained on triplets extracted from the *Fashion IQ Dresses* dataset, following the setting[6] in (H. Wu et al., 2021; Y. Wu, Macdonald, & Ounis, 2022b) and using the checkpoint from Chapter 5. Both user simulators are deployed by using an image captioning tool (called ImageCaptioning.pytorch[7] (Luo, Price, Cohen, & Shakhnarovich, 2018)). Following previous work (described in Section 3.1 and previous chapters), 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 $N = 3$, described in Section 2.1.3) and Success Rate (SR, described in Section 3.2.2). 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.

[5] https://github.com/XiaoxiaoGuo/fashion-retrieval [6] https://github.com/XiaoxiaoGuo/fashion-iq [7] https://github.com/ruotianluo/ImageCaptioning.pytorch

7.4.3 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 (Hidasi et al., 2016) model for sequential recommendations. It 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) (Guo et al., 2018)) 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** (Chapter 4): 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.

• **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 (Sutton & Barto, 2018) (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 (Kang & McAuley, 2018) model for sequential recommendations, which 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 (H. Wu et al., 2021; Y. Wu, Macdonald, & Ounis, 2022a), 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.

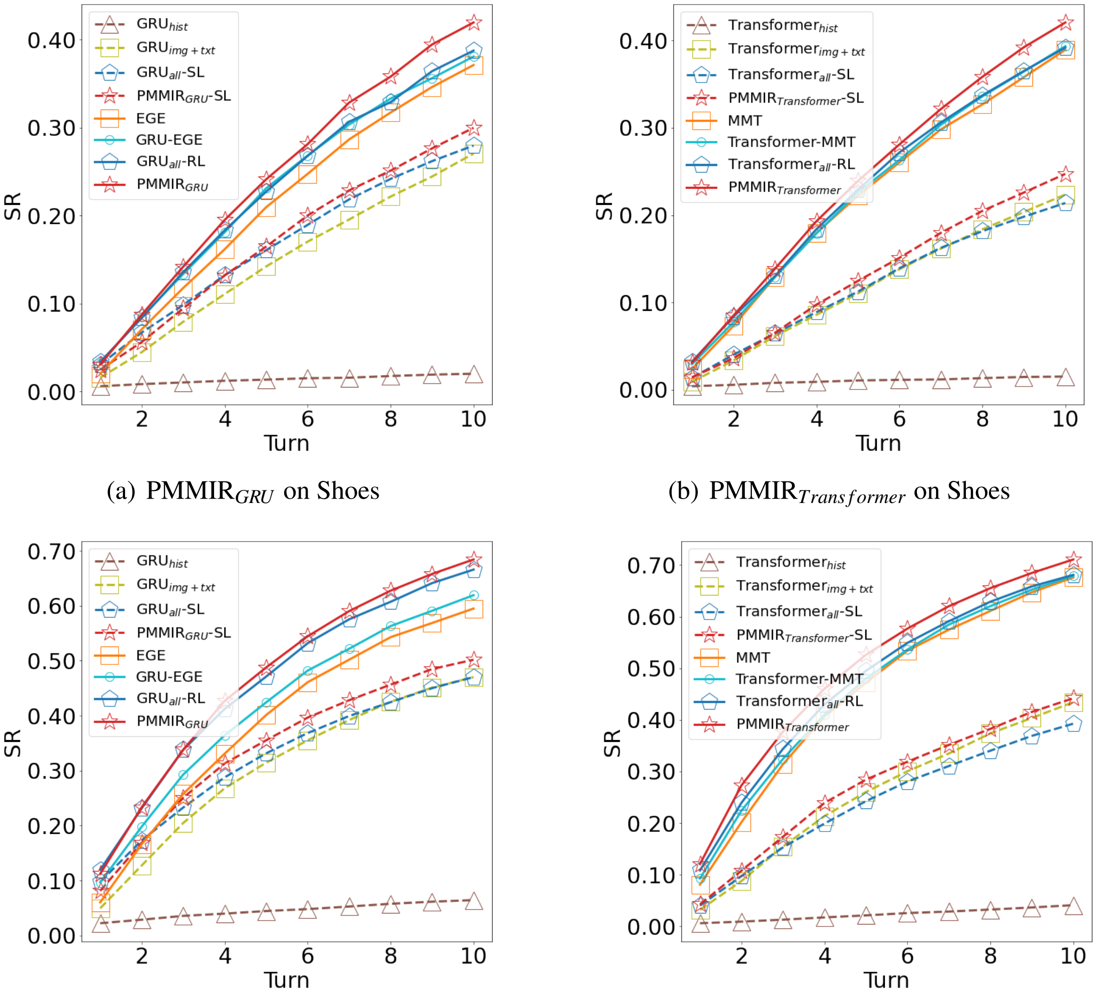
Although there are a few more attention-based/Transformer-based sequential recommendation models (such as BERT4Rec (F. Sun et al., 2019) and Transformers4Rec (de Souza Pereira Moreira, Rabhi, Lee, Ak, & Oldridge, 2021)) and multi-modal interactive recommendation models (such as MMRAN (Chapter 5) with 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 $h_0^c = h_n^p$**: 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/ *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 (Guo et al., 2018; Y. Wu et al., 2021).

• **PMMIR w/ "RN101"**: The "PMMIR w/ RN101" variant replace the ViT-based CLIP checkpoint (i.e. "ViT-B/32") with a ResNet101-based (K. He et al., 2016) 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 7.3.2). Although there are a few more other models with different formulations for the interactive recommendation task (e.g., RCR (R. Zhang et al., 2019), EAR (W. Lei et al., 2020), CRM (Y. Sun & Zhang, 2018), and SGR (Y. Wu, Liao, et al., 2022)), these models are not comparable with our scenario due to requiring additional attributes of items for learning (Haque & Wang, 2022; Yu, Shen, Zhang, et al., 2019; Yuan & Lam, 2021; R. Zhang et al., 2019), requiring a multi-modal knowledge graph for reasoning (Y. Wu, Liao, et al., 2022), or their inability to incorporate both the textual and visual modalities during the recommendation process (W. Lei et al., 2020; Y. Sun & Zhang, 2018).



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

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

Figure 7.4: 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.

7.5 Experimental Results

In this section, we present an analysis of the experimental results in relation to the three research questions outlined in Section 7.4, 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 (RQ7.1, discussed in Section 7.5.1), its performance on both cold-start and warm-start users (RQ7.2, detailed in Section 7.5.2), and the impact of various components and hyperparameters (RQ7.3, covered in Section 7.5.3). To further consolidate our findings, we provide a use case based on the logged experimental results in Section 7.5.4.

Table 7.2: The recommendation effectiveness of our proposed PMMIR model variants (PMMIR_{GRU} and $\text{PMMIR}_{Transformer}$) 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*
$\text{GRU}_{img+txt}$	✗	✓	✓	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												
$\text{Transformer}_{hist}$	✓	✗	✗	SL	0.0104*	0.0107*	0.0149*	0.0150*	0.0213*	0.0228*	0.0411*	0.0422*
$\text{Transformer}_{img+txt}$	✗	✓	✓	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*
$\text{PMMIR}_{Transformer}$-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*
$\text{PMMIR}_{Transformer}$	✓	✓	✓	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

7.5.1 PMMIR vs. Baselines (RQ7.1)

To address RQ7.1, we investigate the performance of our proposed PMMIR model variants (PMMIR_{GRU} and $\text{PMMIR}_{Transformer}$) and the baseline models. Figure 7.4 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 7.4 (a) and (c) represent the results using PMMIR_{GRU}, while Figure 7.4 (b) and (d) correspond to $\text{PMMIR}_{Transformer}$. The x-axis indicates the number of interaction turns. Comparing the results presented in Figure 7.4, 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 7.2 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, and GRU$_{all}$-RL) as well as the Transformer-based models (Transformer$_{hist}$, Transformer$_{img+txt}$, Transformer$_{all}$-SL, MMT, Transformer-MMT, and 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 7.2, 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 RQ7.1, 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.

7.5.2 Cold-Start vs. Warm-Start Users (RQ7.2)

To address RQ7.2, 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) as cold-start users, while those with longer interaction sequences are categorized as warm-start users (as mentioned in Section 7.4.1). This investigation aims to understand how effectively our model adapts to different user scenarios and assess its performance in each case. Table 7.3 presents the performances of our PM-

Table 7.3: 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

MIR 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 7.3, 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 7.3 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

Table 7.4: 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 7.3.

	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*

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 (H. Chen et al., 2022).

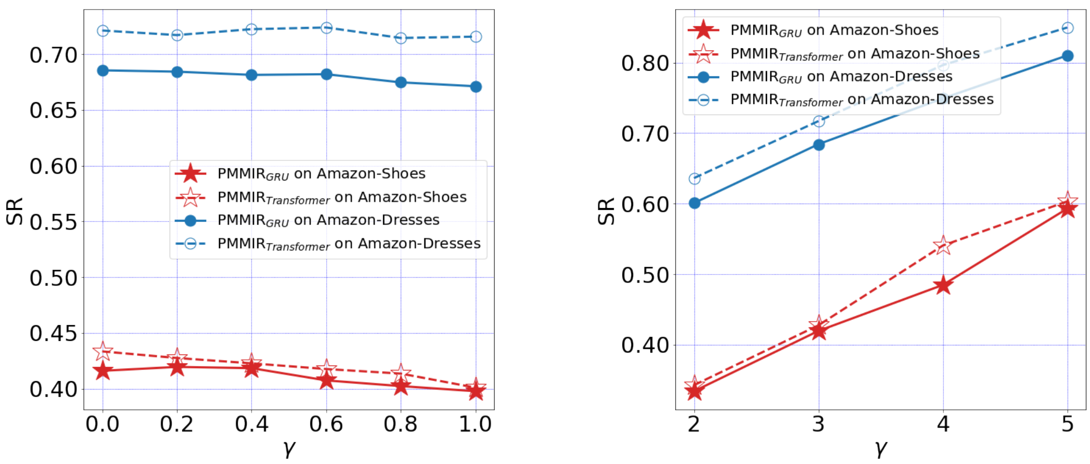
In response to RQ7.2, we find that both cold-start and warm-start users can benefit from our proposed PMMIR model. The warm-start users can generally benefit more with PMMIR$_{GRU}$, while the cold-start users can generally benefit more with PMMIR$_{Transformer}$.

7.5.3 Impact of Components & Hyper-Parameters (RQ7.3)

To address RQ7.3, we investigate the impact of the components and the hyper-parameters of our proposed PMMIR model.

Impact of Components Table 7.4 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 7.5 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 im-



(a) γ for SR

(b) K for SR

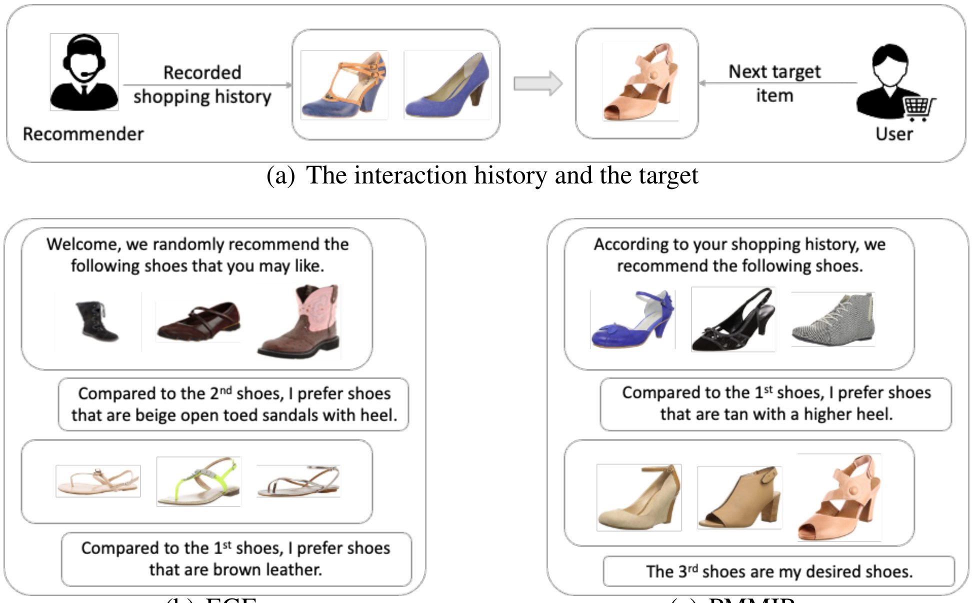
Figure 7.5: Comparison of the recommendation effectiveness at 10th turn with different γ and K values.

portance to all future rewards and considers them on an equal footing. From Figure 7.5 (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 7.5 (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.

Overall, in response to RQ7.3, 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.

7.5.4 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 7.6. 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 7.6 (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


Recommender
Recorded shopping history
Next target item
User
Welcome, we randomly recommend the following shoes that you may like.
Compared to the 2nd shoes, I prefer shoes that are beige open toed sandals with heel.
Compared to the 1st shoes, I prefer shoes that are brown leather.
According to your shopping history, we recommend the following shoes.
Compared to the 1st shoes, I prefer shoes that are tan with a higher heel.
The 3rd shoes are my desired shoes.

(a) The interaction history and the target

(b) EGE

(c) PMMIR$_{GRU}$

Figure 7.6: Example use cases for the multi-modal interactive recommendation task with EGE (without personalisation) and PMMIR$_{GRU}$ (with personalisation) on *Amazon-Shoes*.

increased effectiveness of PMMIR$_{GRU}$ shown in Section 7.5.1. 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 chapter to reduce redundancy.

7.6 Conclusions

In this chapter, 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 (Guo et al., 2018; H. Wu et al., 2021), 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. The experimental results and analysis provide support for the thesis statement with **Research Topic 4** in Section 1.3.

Next, in Chapter 8, we illustrate the realism of simulated conversations by considering positive/negative natural-language feedback in multi-modal conversational recommendation.

Chapter 8

Positive and Negative Natural-Language Feedback

In our thesis statement (as stated in Section 1.3), we metioned the realism of simulated conversations. Therefore, in this chapter, we consider bot positive and negative natural-language feedback in our simulations of users to make the multi-modal conversational recommendation task more realistic. This chapter is mainly based on our work (Y. Wu, Macdonald, & Ounis, 2022a) "Multimodal Conversational Fashion Recommendation with Positive and Negative Natural-Language Feedback" published in the proceedings of the 4th Conference on Conversational User Interfaces (CUI 2022)[1].

In the previous chapters (see Chapters 4, 5, 6, and 7), we have simulated the users' natural-language feedback positively, such as "I prefer blue open toe high heel pumps". However, in a real-world shopping scenario, users can express their natural-language feedback when communicating with a shopping assistant by stating their satisfactions *positively* with "I like" or *negatively* with "I dislike" according to the quality of the recommended fashion products. A multi-modal conversational recommender system (using text and images in particular) aims to replicate this process by eliciting the dynamic preferences of users from their natural-language feedback and updating the visual recommendations so as to satisfy the users' current needs through multi-turn interactions. However, the impact of positive and negative natural-language feedback on the effectiveness of multi-modal conversational recommendation has not yet been fully explored. Since there are no datasets for evaluating conversational recommendation with both positive and negative natural-language feedback, the existing research on multi-modal conversational recommendation imposed several constraints on the users' natural-language expressions (i.e. either only describing their preferred attributes as positive feedback or rejecting the undesired recommendations without any natural-language critiques) to simplify the multi-modal conversational recommendation task. To further explore the multi-modal conversational recom-

[1] DOI: `https://doi.org/10.1145/3543829.3543837`

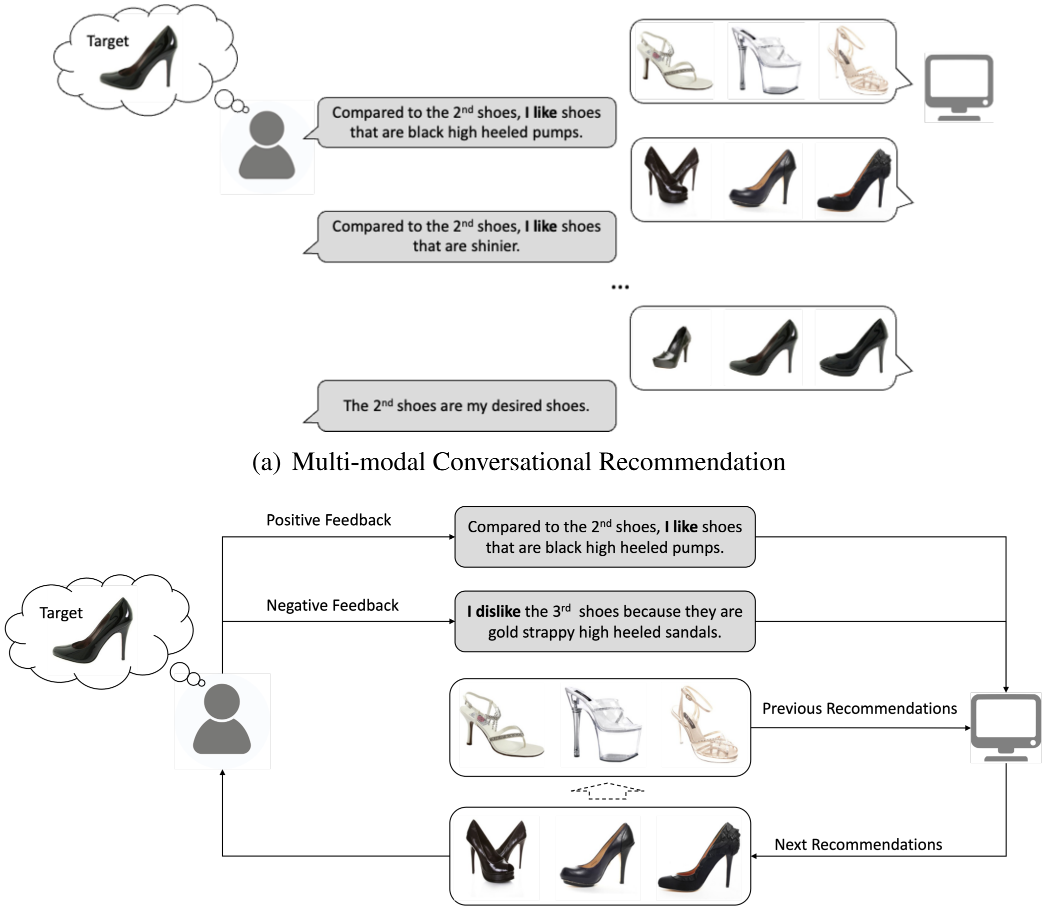
mendation with positive and negative natural-language feedback, we investigate the effectiveness of the recent multi-modal conversational recommendation models for effectively incorporating the users' preferences over time from both positively and negatively natural-language oriented feedback corresponding to the visual recommendations. We also propose an approach to generate both positive and negative natural-language critiques about the recommendations within an existing user simulator. Following previous chapters that simulate the users' feedback with relative captioners (see Section 3.3), we train and evaluate the two existing conversational recommendation models by using the user simulator (see Section 3.3) with *both positive* and *negative* feedback as a surrogate for real human users. Extensive experiments conducted on a well-known fashion dataset (described in Section 3.3.2) demonstrate that positive natural-language feedback (assumed in previous chapters) is more informative relating to the users' preferences in comparison to negative natural-language feedback. The results conform with our thesis statement with **Research Topic 5** in Section 1.3.

8.1 Motivations

In general, the conversational recommender systems have addressed the information asymmetry problem in information seeking, by tracking/eliciting the users' dynamic preferences and take actions (such as recommending items) according to their current needs through multi-turn interactions (described in Section 2.3). Figure 8.1 (a) shows an example of multi-modal conversational recommendation with natural-language feedback (Guo et al., 2018) for fashion products (such as shoes). In this use case (as shown in Chapters 4, 5, 6, and 7), the user gives natural-language feedback (critiques) that describe the differences between the users' preferences (i.e. the *target* item they have in mind) and the system's recommendations at each interaction turn, to obtain items with more preferred features. The conversational recommender system recommends the images of 3 items, based on the users' natural-language critiques.

Such a multi-modal conversational recommendation task is close to a real-world shopping scenario, where the users generally express their natural-language feedback positively or negatively according to the quality of the recommendations when communicating/interacting with the shopping assistants (who may recommend items). In particular, the users might be asked to state their satisfactions using the sentences with "I like" for positive feedback or "I dislike" for negative feedback. Figure 8.1 (b) demonstrates an example of both positive and negative natural-language feedback in the multi-modal conversational recommendation task. The recommender system is expected to update visual recommendations with more preferred features and to avoid recommendations with undesired features according to the users' positive and/or negative natural-language feedback.

Despite the expressiveness of natural-language feedback in conversational recommendation, the impact of positive and negative natural-language feedback on the effectiveness of multi-


Target
Compared to the 2nd shoes, I like shoes that are black high heeled pumps.
Compared to the 2nd shoes, I like shoes that are shinier.
...
The 2nd shoes are my desired shoes.
Positive Feedback
Compared to the 2nd shoes, I like shoes that are black high heeled pumps.
Negative Feedback
I dislike the 3rd shoes because they are gold strappy high heeled sandals.
Target
Previous Recommendations
Next Recommendations

(a) Multi-modal Conversational Recommendation

(b) Positive & Negative Natural-Language Feedback

Figure 8.1: An example of multi-modal conversational recommendation with positive & negative natural-language feedback.

modal conversational recommendation has not yet been fully explored. Due to the lack of multimodal conversations with both positive and negative natural-language critiques about the visual recommendations in terms of the users' preferences, the existing research on multi-modal conversational recommendation imposed several constraints on the users' natural-language expressions, in order to simplify the multi-modal conversational recommendation task. For instance, the users are assumed to either only describe their preferred attributes as positive feedback (Guo et al., 2018; Y. Sun & Zhang, 2018; H. Wu et al., 2021; Y. Wu et al., 2021; Yu, Shen, & Jin, 2019; Yu et al., 2020; Y. Zhang et al., 2018) or just reject the undesired item-level recommendations without any natural-language critiques (Bi et al., 2019; W. Lei et al., 2020; Xu et al., 2021) during the multi-turn interactions. To learn satisfactory recommender systems with enough training data, *user simulators* have been used as surrogates for real human users in the optimisation and evaluation processes (see Sections 3.1 and 3.3). In particular, Guo et al. (2018) proposed a user simulator with only positive natural-language feedback for relative captioning (Rennie et al., 2017). Meanwhile, W. Lei et al. (2020) formulated the conversational recommendation task as answering the questions about the attributes and the recommended items with a binary yes/no response.

In this chapter, we investigate the effectiveness of the recent multi-modal conversational recommendation models for effectively incorporating the users' preferences over time from positively and/or negatively natural-language oriented feedback corresponding to the visual recommendations. To make the conversational recommendation task more realistic by supporting both positive and negative natural-language feedback, we propose an approach to generate both positive and negative natural-language critiques about the recommendations with an existing user simulator for relative captioning (see Sections 3.3). Following previous work, we train and evaluate the two existing multi-modal conversational recommendation models (i.e. Dialog Manager (DM) and Multi-modal Interactive Transformer (MIT) (see Sections 3.1) by using the user simulator with positive and negative feedback as a surrogate for real human users. Extensive experiments conducted on a well-known fashion dataset demonstrate that positive feedback is more informative relating to the users' preferences in comparison to negative feedback. The main contributions of this chapter are summarised as follows:

- We first investigate the effectiveness of the multi-modal conversational recommendation models with both positive and negative natural-language feedback. Different from the previous work relating to positive and negative feedback, the users are assumed to actively express their satisfactions positively with "I like" or negatively with "I dislike" according to the quality of the recommendations, rather than answering questions passively with "yes" or "no".
- We propose an approach to generate both positive and negative natural-language feedback with a user simulator for relative captioning, which enables our research with various combinations of positive and negative natural-language sentences.
- We investigate the impact of different textual encoding mechanisms (i.e. pre-trained contextual embeddings (Devlin et al., 2019b) and one-hot embeddings) on the effectiveness of the multi-modal conversational recommendation models.
- Extensive empirical evaluations are performed on the multi-modal recommendation task, demonstrating different levels of difficulties for incorporating the users' preferences from positive and negative feedback over existing state-of-the-art approaches (i.e. those in previous chapters) while providing directions for future work.

The remainder of the chapter is organised as follows: In Section 8.2, we review the related work and position our contributions in comparison to the existing literature about positive and negative natural-language feedback. Section 8.3 defines the problem statement and extends two recent multi-modal conversational recommendation models for top-K recommendations. Section 8.4 presents the existing user simulator for relative captioning and extends it for generating both positive and negative natural-language feedback. Our experimental setup and results are presented in Sections 8.5 and 8.6, respectively. Section 8.7 summarises our findings and provides possible future work.

8.2 Related Work

In this section, we mainly introduce positive & negative natural-language feedback in the recommendation field.

8.2.1 Positive & Negative Natural-Language Feedback

Positive/negative explicit/implicit feedback (such as ratings, transactions, clicks, and skips) have been intensively investigated in the recommendation field (Batmaz et al., 2019; Chakraborty et al., 2021; Deldjoo et al., 2022; S. Wang et al., 2019; X. Zhao, Zhang, et al., 2018; Zou et al., 2020). For instance, X. Zhao, Zhang, et al. (2018) proposed a deep Q-learning network (DQN) based recommender system with GRUs by incorporating both positive implicit feedback (i.e. clicks) and negative implicit feedback (i.e. skips) from the logged implicit interactions datasets. In recent research, natural-language feedback has been proven to be more informative relating to the users' preferences compared to the non-verbal explicit/implicit feedback. Although natural-language feedback has been intensively investigated in the conversational recommendation field (Gao et al., 2021; Jannach et al., 2021; X. Zhao, Zhang, et al., 2018), these existing research on conversational recommendation imposed several constraints on the users' natural-language feedback to simplify the conversational recommendation task. In particular, the users are assumed to either only describe their preferred attributes as positive feedback (Guo et al., 2018; Y. Sun & Zhang, 2018; H. Wu et al., 2021; Y. Wu et al., 2021; Yu, Shen, & Jin, 2019; Yu et al., 2020; Y. Zhang et al., 2018) or just answer attribute-level questions with a binary yes/no response while rejecting the undesired item-level recommendations without any natural-language critiques (Bi et al., 2019; W. Lei et al., 2020; Xu et al., 2021) during the multi-turn interactions. For instance, the existing multi-modal conversational recommendation models based either on a GRU (Guo et al., 2018; Y. Wu et al., 2021; R. Zhang et al., 2019) or a Transformer Encoder (H. Wu et al., 2021) only consider the users' positive natural-language feedback for describing their desired features in terms of the recommendations, thereby directing the recommender systems towards obtaining a correct desired item. Meanwhile, W. Lei et al. (2020) formulated the conversational recommendation task as answering the questions about the attributes and the recommended items with a binary yes/no response. Furthermore, a multi-round conversational recommender system (called Feedback-guided Preference Adaptation Network (FPAN)) (Xu et al., 2021) was recently proposed to consider the relation between attribute-level and item-level positive and negative feedback signals. The users' feedback is constrained to answer the questions asked by the recommender systems and is also simplified by answering "yes" for acceptance and "no" for rejection in terms of the attribute-level clarification questions and the and item-level recommendations from the recommender systems. However, we argue that users should be able to actively express their positive and/or negative critiques about the recommendations via natural language in addition to answering the recommender systems' questions. Such a constraint with only positive natural-language feedback or a simplification with "yes" or "no" is limited by the conversational recommendation datasets available, which makes the research less realistic in the shopping scenario. To this end, we propose an approach to generate both positive and negative natural-language feedback with the existing user simulator for relative captioning (see Section 3.3).

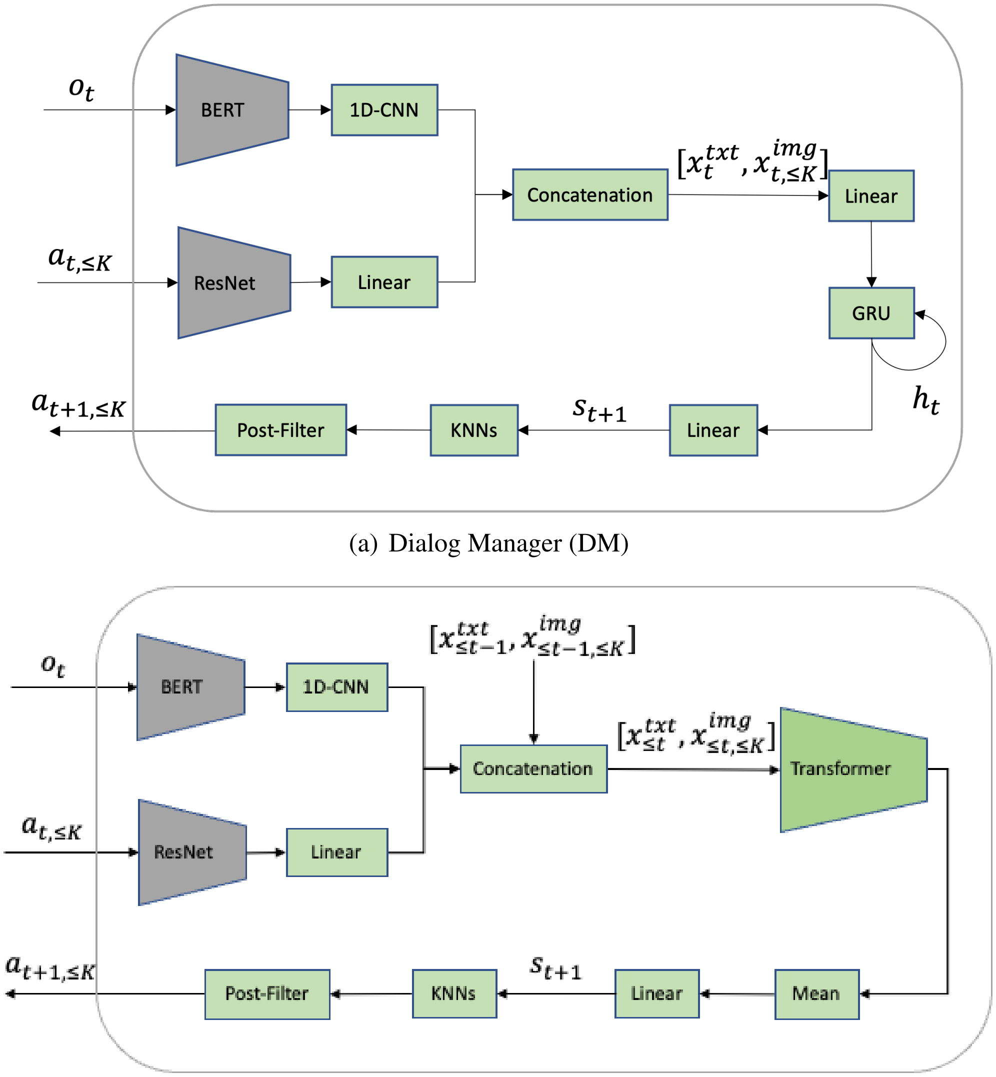
As a consequence, in this chapter, we investigate the effectiveness of the existing multimodal conversational recommendation models with both positive and negative natural-language feedback that describes the users' desired/undesired features in terms of the visual recommendations. To the best of our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

8.3 The Multi-modal Conversational Recommendation Models

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

8.3.1 Preliminaries

We study the multi-modal conversational recommendation task by considering a user interacting with a recommender system via iterative interaction turns with text and images. At the t-th interaction turn, the recommender system presents K candidate images $a_{t,\leq K} = (a_{t,1},...,a_{t,K})$ selected from a candidate pool $\mathscr{I} = \{a_i\}_{i=0}^{N}$ to the user. The user then provides a natural language critique, o_t, as feedback, describing the major differences between the candidate image and the desired image. The natural language feedback can be positive – having the form "Compared to the k-th item, I like ..." (i.e. o_t^+) or negative – such as "I dislike the k-th item because ..." (i.e. o_t^-). Based on the users' positive/negative natural-language feedback and the interaction history up to turn t, $\tau_t = (o_{\leq t}, a_{\leq t,\leq K}) \in \mathscr{H}$, where $o_{\leq t} = (o_1,...,o_t) \in \mathscr{O}$ and $a_{\leq t,\leq K} = (a_{1,\leq K},...,a_{t,\leq K}) \in \mathscr{A}$, the recommender system selects another list of candidate images $a_{t+1,\leq K}$ from the candidate image pool. This vision-language interaction process continues until the target image a_{target} is recommended or the maximum number of interaction turns M is reached.



(a) Dialog Manager (DM)

(b) Multi-modal Interactive Transformer (MIT)

Figure 8.2: Architectures of the multi-modal conversational recommendation models: (a) Dialog Manager (DM) and (b) Multi-modal Interactive Transformer (MIT).

8.3.2 The Model Architectures

Figure 8.2 shows the architectures of two end-to-end models (i.e. Figure 8.2 (a) Dialog Manager (DM) (Guo et al., 2018) and Figure 8.2 (b) Multi-modal Interactive Transformer (MIT) (H. Wu et al., 2021)) for multi-modal conversational recommendations to effectively incorporate the users' preferences over time. The user views the recommended items (K items at each interaction) and provides positive natural-language feedback by describing their desired features that the current recommended items lack. Alternatively, the user can provide negative feedback by describing the undesired features in the current recommended items compared to the user's envisaged target item.

Text & Image Encoders The multi-modal conversational recommendation models track and estimate the user's preferences from both the user's positive and negative natural-language feedback and the latest recommended visual items. The positive and negative natural-language feedback texts are encoded with a text encoder, while the recommended images are encoded with an image encoder (Guo et al., 2018; H. Wu et al., 2021). In particular, the text encoder (which consists of a pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019b), a 1D convolutional layer (1D-CNN) and a subsequent linear layer) encodes the positive and negative natural-language feedback texts into a single textual representation. Alternatively, each word in the sentences can also be represented by a one-hot vector with pre-defined vocabulary (Guo et al., 2018; H. Wu et al., 2021) of fashion-related terms. We adopt the pre-trained BERT model as our default encoding mechanism, while we investigate the impact of different encoding mechanisms (i.e. the one-hot encoding and the BERT encoding) in Section 8.6.3. In a similar manner to the text encoder, the image encoder extracts image feature representations based on the ImageNet pre-trained ResNet101 model (K. He et al., 2016) and subsequently transforms the extracted image feature representations with a linear layer. Then, both the image feature representations and the textual representations are concatenated as input to a subsequent GRU (Guo et al., 2018) or Transformer Encoder (H. Wu et al., 2021) to model the user's estimated preferences.

The State Trackers Given a list of candidate images $a_{t,\leq K} = (a_{t,1}, ..., a_{t,K})$ and a user's corresponding natural-language feedback o_t at the t-th dialog turn, the encoded textual representation is denoted by x_t^{txt} and the encoded image representation is denoted by $x_{t,\leq K}^{img} = ResNet(a_{t,\leq K})$. The concatenated textual and image representations $[x_t^{txt}, x_{t,\leq K}^{img}]$ are further tracked in a gated recurrent unit (GRU) (Chung et al., 2014) as in (Guo et al., 2018). The estimated state of user's preferences can be achieved with

$$s_{t+1} = Linear(GRU(Linear([x_t^{txt}, x_{t,\leq K}^{img}]), h_t)),$$

where $h_t = GRU(Linear([x_{t-1}^{txt}, x_{t-1,\leq K}^{img}]), h_{t-1})$ is the estimated hidden states of the user's preferences. The GRU component allows the model to sequentially aggregate the recommendations and positive/negative feedback information from the recommender system's recommendations and the user's natural-language feedback to the estimated hidden states. Alternatively, a Transformer-based state tracker enables the recommendation model to attend to the entire history of the multi-modal interactions. The estimated state of user's preferences can be achieved with

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

The Top-K Recommendations Based on the estimated state of user's preferences, a list of candidate items can be recommended for the next action. If K items are recommended at each turn $t+1$, we select the top-K closest images to the estimated state s_{t+1} under the Euclidean distance in the image feature (ResNet) space: $a_{t+1,\leq K} \sim KNNs(s_{t+1})$, where $KNNs()$ is a softmax distribution over the top-K nearest neighbours of s_{t+1} and $a_{t+1,\leq K} = (a_{t+1,1}, ..., a_{t+1,K})$. Furthermore, to avoid repeated recommendations during the multi-turn interactions, we adopt a post-filter, as in (Y. Wu et al., 2021), to remove any candidate items from the ranking list that have previously occurred in the recommendation history $a_{\leq t,\leq K}$.

The Triplet Loss Function User simulators (Ekstrand et al., 2021; Guo et al., 2018; H. Wu et al., 2021; S. Zhang & Balog, 2020) are generally used as a surrogate for real human users in the training processes. For a fair comparison, we train the above GRU/Transformer-based models with a triplet loss objective, L_{tri}, similar to (Guo et al., 2018; H. Wu et al., 2021):

$$L_{tri} = max(0, ||s_{t+1} - x_{+}^{img}||_2 - ||s_{t+1} - x_{-}^{img}||_2 + m) \tag{8.1}$$

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

8.4 A User Simulator with Positive and Negative Feedback

To learn satisfactory multi-modal conversational recommender systems with enough training data, user simulators based on vision and language (VL) have been considered as surrogates for real human users in the optimisation and evaluation processes (see Section 3.3). The adoption of such VL-based user simulators helps to avoid collecting and annotating entire multi-modal conversations, which is expensive, time-consuming, and does not scale (S. Zhang & Balog, 2020).

User Simulators for Relative Captioning Such user simulators have been generally formulated as *relative captioners* for fashion recommendation (Guo et al., 2018; H. Wu et al., 2021) that can automatically generate descriptions of the prominent visual differences between any pair of target and candidate images (i.e. a target representing the user's desired item and the candidate representing a recommendation by the system). For instance, Guo et al. (2018) applied long short-term memory network (LSTM)-based models, such as *Show, Tell* (Vinyals et al., 2015), to generate the relative captions as natural-language critiques about the recommendations. These user simulators for relative captioning have been thoroughly evaluated via both a quantitative evaluation and a user study, which showed that the user simulators for relative captioning can serve as a reasonable proxy for real users (Guo et al., 2018).

User Simulators with Positive/Negative Feedback Here, we propose an approach to generate positive and negative natural-language feedback with the existing user simulators for relative captioning (Guo et al., 2018). In the relative captioning task, the relative captioner $cap_{rel}()$ is given a candidate image $a_{t,k}$ ($k \in [1,K]$) and a target image a_{target} and it is tasked with describing the differences of $a_{t,k}$ relative to a_{target} in natural language. To generate both positive and negative feedback, two sentence templates, which state the users' satisfactions positively with "I like" and negatively with "I dislike", are appended to the relative captions from $cap_{rel}()$. In particular, the positive feedback of the image pair (i.e. $a_{t,k}$ and a_{target}) is defined as follows:

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

where each relative caption $cap_{rel}(a_{target}, a_{t,k})$ describes what is missing from the candidate image $a_{t,k}$ to obtain a_{target}. Inversely, the relative captioner $cap_{rel}()$ can also describe the features of the candidate image $a_{t,k}$ that are not contained in the target image a_{target}. Therefore, we propose that negative feedback can thus be instantiated by reversing candidate and target images:

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

and changing the textual prefix from "I like" to "I dislike". It is worth noting that we adopt templates as wrappers to handle users' positive and negative utterances so as to reduce the errors for language understanding and generation. We demonstrate an example of positive & negative feedback with relative captioning in Section 8.6.4.

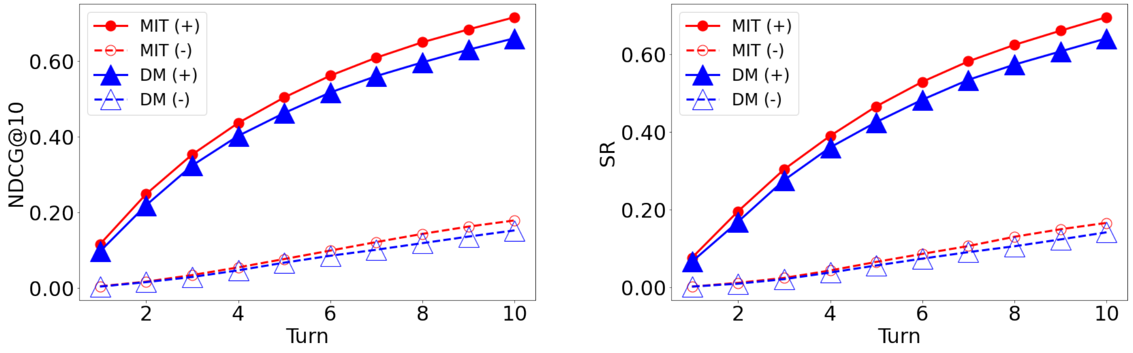
8.5 Experimental Setup For Recommendation

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

- RQ8.1: Is positive natural-language feedback more informative relating to the users' preferences in comparison to negative natural-language feedback?
- RQ8.2: Can the combined positive & negative natural-language feedback enhance the ability of the existing GRU/Transformer-based models in incorporating the users' preferences?
- RQ8.3: What is the impact of the natural-language encoding on the models' performances?

8.5.1 Dataset & Measures

Dataset We perform experiments on the *Shoes* dataset (see Section 3.3.2). The dataset provides $10,751$ pairs of images with relative captions about their visual differences and $3,600$ images with captions about their discriminative visual features for training a user simulator. In



(a) NDCG@10

(b) SR

Figure 8.3: Comparison of the recommendation effectiveness of DM and MIT with single-sentence feedback at various interaction turns with top-3 recommendation on *Shoes*. + & - denote positive and negative natural-language feedback, respectively.

Table 8.1: Multi-modal conversational recommendation effectiveness of the tested models at the 5th & 10th turns on the *Shoes* dataset. The best overall results are highlighted in bold. The best performing results in the first and second parts of the table are underlined, while the best overall performing results are highlighted in bold. † and * respectively denote significant differences in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to the best performing results in the first group and the best overall performing results. + and - denote positive & negative natural-language feedback, respectively.

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

addition, the dataset also contains $10,000$ images for training the recommender systems, and $4,658$ images for testing.

Measures The effectiveness of the multi-modal conversational models is measured by Normalised Discounted Cumulative Gain (i.e. NDCG@N truncated at rank $N = 10$ calculated at the M-th interaction, see Section 2.1.3) and Success Rate (SR, see Section 3.2.2) at the M-th interaction, as in Chapter 4. In particular, SR is the percentage of users who find their target items in the top-K recommendation lists among all the users within M interactions. Furthermore, it is possible that the user may view more of the ranking of items at each interaction turn, down to rank N. We use the evaluation metrics (i.e. NDCG@10 and SR) at the 5th and 10th interaction turn for significance testing.

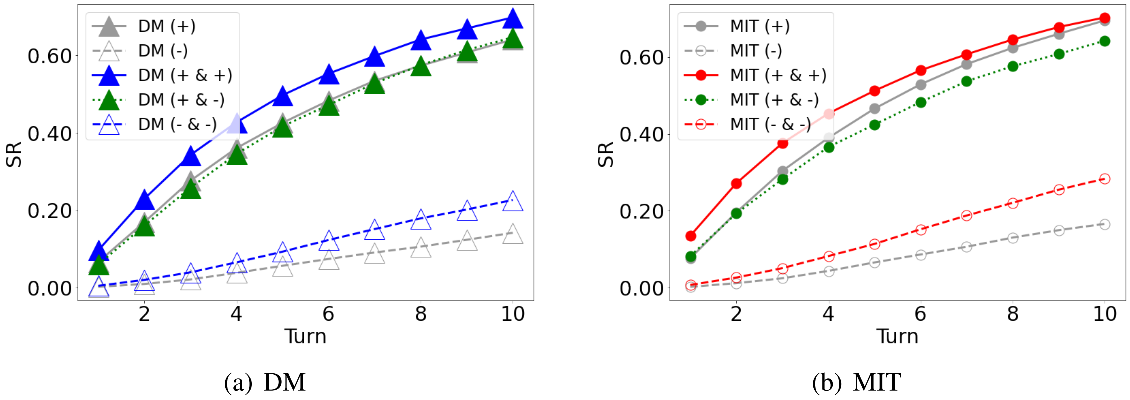
8.5.2 Experimental Settings

Setup for User Simulator A user simulator with the *Shoes* dataset was intensively and carefully trained by (Guo et al., 2018) through crowdsourcing relative expressions about the visual differences of the image pairs that are written by real human users in natural language. Furthermore, the pre-trained user simulator has previously been thoroughly evaluated via both a quantitative evaluation and a user study (Guo et al., 2018), thereby serving as a reasonable proxy for real users in our work. The pre-trained user simulator can generate either positive or negative natural-language feedback with our proposed approach as illustrated in Section 8.4. At each interaction turn, the user simulator gives feedback on the candidate images that are the most and/or least similar to the target image. We consider five types of natural-language feedback in our experiments: single-positive (i.e. +), single-negative (i.e. -), paired-positive (i.e. + & +), paired-negative (i.e. - & -) and mixed (i.e. + & -) feedback. In single-positive (i.e. +), single-negative (i.e. -) and mixed (i.e. + & -) feedback settings, the most similar candidate image receives positive feedback and/or the least similar candidate image receives negative feedback.

Setup for Recommender Systems We then train the models (i.e. DM and MIT) with the user simulator on the *Shoes* dataset. The parameters of the models are randomly initialised. We use Adam (Kingma & Ba, 2014) with a learning rate 10^{-3} (Guo et al., 2018; R. Zhang et al., 2019). We set the embedding dimensionality of the feature space to 256 and the batch size to 128 as in (Guo et al., 2018). For each batch, we train the model with 10 interaction turns as in (Y. Wu et al., 2021). We consider the top-K items (K=3) as a recommendation list at each interaction turn for testing. For the evaluation metrics, we denote the interaction turn $M \in [1, 10]$. If a user obtains the target item in less than 10 interaction turns, we consider the ranking metric (i.e. NDCG@10) for that user to be equal to one for all turns thereafter (see Chapter 6).

8.6 Experimental Results

In this section, we analyse the experimental results respect to the research questions stated in Section 8.5, concerning the effectiveness of the models for multi-modal conversational recommendations with positive and negative natural-language feedback (Section 8.6.1), the impact of the combined positive and negative feedback (Section 8.6.2), and the impact of the textual encoding mechanisms (Section 8.6.3). We demonstrate a use case for generating both positive and negative feedback, as well as a use case from the logged experimental results to consolidate our findings (Section 8.6.4).



(a) DM

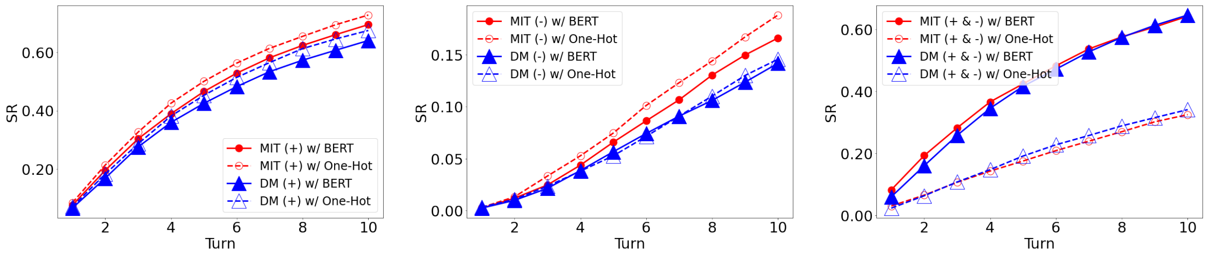
(b) MIT

Figure 8.4: Comparison of the recommendation effectiveness of DM and MIT with various types of natural-language feedback at various interaction turns with top-3 recommendation on *Shoes*.

8.6.1 Positive vs. Negative Feedback (RQ8.1)

Figure 8.3 shows the recommendation effectiveness of the DM and MIT models with positive or negative single-sentence feedback for top-3 recommendation in terms of NDCG@10 (Figure 8.3 (a)) and SR (Figure 8.3 (b)), while varying the number of interaction turns on the *Shoes* dataset. The solid lines show the models' performances with positive natural-language feedback (denoted +), while the dashed lines show performances with negative natural-language feedback (denoted -). Comparing the results in Figure 8.3, we observe that both DM and MIT models generally achieve a better overall performance with positive feedback than negative feedback in terms of NDCG@10 and SR. The better performance of the tested models with positive feedback compared to those with negative feedback indicates that positive natural-language feedback in more informative relating the users' preferences than negative natural-language feedback. In addition, MIT achieves a better overall performance than DM in terms of NDCG@10 and SR at various interaction turns with positive and negative natural-language feedback. Such an observation is aligned with the results reported in (H. Wu et al., 2021) considering positive natural-language feedback only.

Table 8.1 shows the obtained recommendation performances of the tested models (i.e. DM and MIT) with the same test sets of the *Shoes* dataset at the 5th and 10th interaction turns. More specifically, Table 8.1 contains two parts: the first part reports the effectiveness of the models with either positive or negative feedback. The second part reports the effectiveness of the models with different combinations of positive or negative feedback. The best performing results in the first and second parts of the table are underlined, while the best overall performing results are highlighted in bold in Table 8.1. † and * respectively denote significant differences in terms of a paired t-test with a Holm-Bonferroni multiple comparison correction ($p < 0.05$), compared to the best performing results in the first group and the best overall performing results. Comparing the results in the first group of rows in the table, we observe that both DM and MIT achieve a



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

Figure 8.5: Effects of the textual encoding mechanisms on both DM and MIT with different types of natural-language feedback.

significant better overall performance in terms of both NDCG@10 and SR at the 5th and 10th turns with positive feedback (denoted +) than with negative feedback (denoted -) on the *Shoes* dataset, respectively.

In answer to RQ8.1, the results demonstrate that the tested models with positive feedback are significantly more effective than those with negative feedback. Therefore, it can be inferred that positive feedback is more informative relating to the users' preferences than negative feedback. The DM and MIT models can better incorporate the users' preferences from the recommended visual items with positive natural-language feedback than negative natural-language feedback.

8.6.2 Impact of the Combined Feedback (RQ8.2)

Figure 8.4 (a) and Figure 8.4 (b) illustrate the SR of DM and MIT with different types of natural-language feedback (i.e. different combinations of positive and negative feedback) at the various interaction turns, respectively. The gray lines show the DM/MIT model's performances with a single sentence at each interaction turn, while the blue/red and green lines show performances with a pair of sentences at each interaction. Comparing the results in Figure 8.4 (a) and Figure 8.4 (b), we observe that both DM and MIT achieve a better overall performance with paired positive (i.e. + & +) or paired negative (i.e. - & -) natural-language feedback sentences in comparison to the models with a single positive (i.e. +) or single negative (i.e. -) natural-language feedback sentence. Furthermore, the performances of DM and MIT differ with a pair of both positive and negative feedback sentences. In particular, the performance of DM (+) and DM (+ & -) are very close in term of SR at various interaction turns, while MIT (+) outperforms MIT (+ & -) overall except for the initial two interaction turns. The better performance of the models with (+ & +) and (- & -) compared to the models with (+) and (-) can be attributed to the fact that the same type of natural-language feedback at each turn can be aggregated to leverage the information relating to the users' preferences. Meanwhile, the paired positive and negative feedback make it challenging for DM and MIT the elicit the users' preferences from the feedback sentences with opposite sentiments. Furthermore, Table 8.1 demonstrate that DM (+ & +)

and MIT (+ & +) are significantly more effective than those with other types of natural-language feedback at both 5-th and 10-th interaction turns, except for MIT (+) in term NDCG@10 and SR at the 10-th interaction turn.

Overall, in response to RQ8.2, we find that the single type of natural-language feedback (i.e. either paired positive or paired negative feedback) at each turn can be aggregated to leverage the information relating to the users' preferences, while the paired positive and negative feedback make it challenging for DM and MIT to elicit the users' preferences.

8.6.3 Impact of Textual Encoding (RQ8.3)

To address RQ8.3, Figure 8.5 depicts the effects of the textual encoding mechanisms on both DM and MIT with different types of natural-language feedback. Figure 8.5 (a) demonstrates that both DM (+) and MIT (+) with the one-hot encoding using a pre-defined vocabulary of fashion-related terms achieve an overall better performance in comparison to those with the BERT encoding. Figure 8.5 (b) shows that MIT (-) with the one-hot encoding also outperforms MIT (-) with the BERT encoding, while DM (-) with the one-hot encoding and the BERT encoding are almost the same. The better performance of the models with the one-hot encoding compared to the BERT encoding can be attributed to the fact that the pre-defined fashion vocabulary for the one-hot encoding is much smaller and is more concentrated on fashion features than BERT. Furthermore, Figure 8.5 (c) shows that the performances of DM (+ & -) and MIT (+ & -) with the one-hot encoding are dramatically degraded compared to those with the BERT encoding that is able to capture the contextual information between sentences with the pre-trained contextual embeddings. Such a difference can be attributed to the inability of the one-hot encoding in capturing the relations between the positive and negative natural-language feedback.

Overall, in response to RQ8.3, we find that the BERT encoding is surprisingly important to capture the contextual information with the pre-trained contextual embeddings when there are both positive and negative feedback, while the one-hot encoding can enhance the models' performance by using a pre-defined fashion vocabulary that is more concentrated on fashion features than BERT.

8.6.4 Use Cases

A Use Case for Generating Positive & Negative Feedback To show that it is realistic to generate both positive and negative natural-language feedback with our proposed approach (illustrated in Section 8.4), we provide examples of the generated natural language critiques for given target images and candidate images in Table 8.2, on the *Shoes* dataset. There are two ground truths for each pair of the candidate and target images, while following the generated positive and negative natural-language feedback by the aforementioned user simulator for relative captioning. We observe that the user simulator can effectively describe the major differences between the target and candidate images, while using phrases that differ from the ground truth critiques, such as “all white” vs. “solid white” for the positive feedback and “more athletic soles” vs. “chunkier sole” for the negative feedback.

A Use Case for Multi-modal Conversational Recommendation To consolidate the results observed in the the above sections, we present a use case of multi-modal conversational recommendation in Table 8.3 and Table 8.4 on the *Shoes* dataset. Table 8.3 and Table 8.4 show the interaction process for top-3 recommendation with the DM model over positive feedback (i.e. DM (+)) and negative feedback (i.e. DM (-)), respectively. For fair comparison, the initial images are the same across DM (+) and DM (-) given the target image from the testing set. We observe that DM with positive feedback is more effective than negative feedback. In particular, DM with positive feedback only needs 2 interaction turns to display the desired item in addition to the initial random recommendation by capturing the key features from the user’s positive feedback, such as “gold”, “open-toed”, “high heels”, and “straps”. However, DM with negative feedback fails to recommend the user’s desired shoes within 5 interaction turns. Although, DM with negative feedback can successfully capture the “open toe” feature from the rejection of the “closed toe” feature, it is still struggling with the decisions of the colours and the thickness of the platform. The differences in interaction rounds indicate that positive feedback is typically more conducive than negative feedback to capturing user preference information in recommendation systems.

8.7 Conclusions

In this chapter, we first investigated the effectiveness of the multi-modal conversational recommendation models with both positive and negative natural-language feedback. To make the conversational recommendation task more realistic with both positive and negative natural-language feedback, we proposed an approach to generate both the positive and negative natural-language critiques about the recommendations with the existing user simulator for relative captioning. Following previous chapters (see Chapters 4, 5, 6, and 7), we trained and evaluated the two existing conversational recommendation models by using the user simulator with positive and negative feedback as a surrogate for real human users. Our experiments on the *Shoes* dataset demonstrated that positive feedback is more informative relating to the users’ preferences in comparison to negative feedback. Our reported results also showed that the types of users’ natural-language feedback (i.e. different combinations of positive and negative feedback) and the types of textual encoding mechanisms (i.e. pre-trained contextual embeddings and one-hot embeddings) can greatly affect the performance of the both tested models (i.e. DM & MIT).

The experimental results and analysis provide support for the thesis statement with **Research Topic 5** in Section 1.3.

Thus far, we have shown the effectiveness of the various recommender systems, which instantiated our proposed multi-modal conversational recommendation framework in Chapter 3. In particular, starting from Chapter 4 until Chapter 8, we have proposed various techniques to address the challenges (i.e. partial observations in natural-language feedback, multi-modal sequence dependency issue, coupling of policy optimisation and representation learning, personalisation, and the realism of simulated conversations with negative natural-language feedback) within such a framework. Therefore, in the next chapter, we summarise the main contributions and conclusions of this thesis and also discuss possible future directions.

Table 8.2: An example of positive and negative feedback with relative captioning on the *Shoes* dataset.

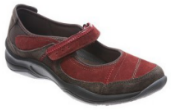
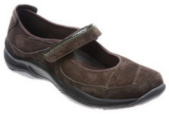
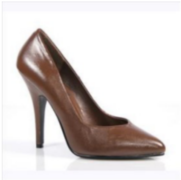
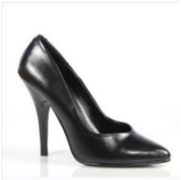
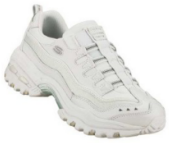
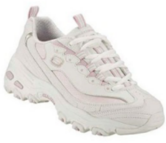
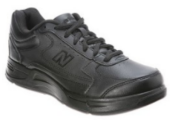
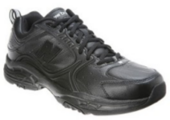
Pair	Target	Candidate	Feedback
1			**Ground Truths:** • Compared to the candidate shoes, the target shoes are shoes with red trim. • Compared to the target shoes, the candidate shoes are brown, not red. **Positive Feedback:** Compared to the candidate shoes, I like shoes that are red and black. **Negative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

Table 8.3: An example use case for multi-modal conversational recommendation for the Dialog Manager model with positive natural-language feedback on the *Shoes* dataset.

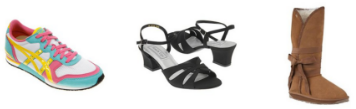
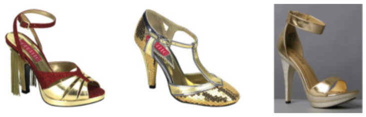
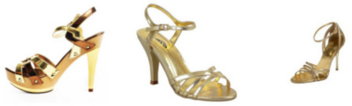
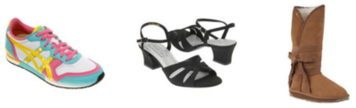
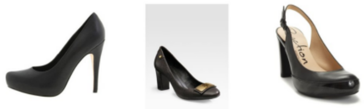
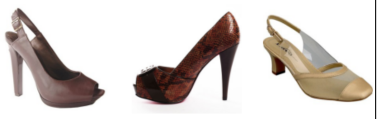
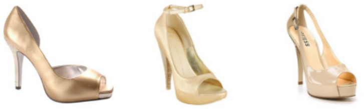
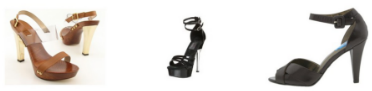
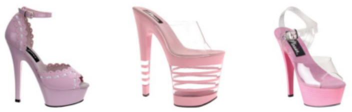
Turn	Top-3 Recommendations	Positive Feedback
0		Compared to the 3rd shoes, I like shoes that are gold open toe high heels.
1		Compared to the 3rd shoes, I like shoes that are open-toed with straps.
2		The 1st shoes are my desired shoes.

Table 8.4: An example use case for multi-modal conversational recommendation for the Dialog Manager model with negative natural-language feedback on the *Shoes* dataset.

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

Chapter 9

Conclusions and Future Work

9.1 Contributions and Conclusions

In this thesis, we focused on developing multi-modal conversational recommender systems by leveraging the multi-modal interactions (including both visual and textual information) between users and recommender systems. Specifically, we leveraged advanced techniques (including multi-modal learning, sequential recommender systems, and reinforcement learning approaches) for reformulating and improving the multi-modal conversational recommendation framework. In Section 1.2, we stated that this thesis was motivated to address the following challenges:

- **Challenge 1:** How to better understand the users' natural-language feedback and the corresponding recommendations with the partial observations of the users' preferences over time;
- **Challenge 2:** How to better tracking the users' preferences over the sequences of the systems' visual recommendations and the users' natural-language feedback;
- **Challenge 3:** How to decouple the recommendation policy (i.e. model) optimisation and the multi-modal composition representation learning;
- **Challenge 4:** How to effectively incorporate the users' long-term and short-term interests for both cold-start and warm-start users;
- **Challenge 5:** How to ensure the realism of simulated conversations, such as positive/negative natural-language feedback.

To address the above challenges (i.e. questions), we first described the multi-modal conversational recommendation framework in Chapter 3 and illustrated the challenges and opportunities within the framework. Furthermore, we have proposed various models from Chapter 4 to

Chapter 8 for addressing the above challenges corresponding to five research topics (mentioned in Section 1.3). In particular, we argued that the tasks of modelling multi-modal conversational recommendations can be effectively achieved by tracking users' preferences with partial observations, mitigating the multi-modal sequence dependency issue, decoupling the composition representation learning from policy optimisation, incorporating both the users' long-term preferences and short-term needs for personalisation, and ensuring the realism of simulated conversations. Below, we will describe our main contributions and conclusions in addressing these challenges:

- **Research Topic 1:** *By modelling the multi-modal conversational recommendation process with (self-)supervised Q-learning in a partially observable environment, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time using the partial observations*. To address **Challenge 1**, we proposed a novel dialog-based recommendation model, denoted by the Estimator-Generator-Evaluator (EGE) model, with Q-learning for POMDP to effectively incorporate the users' preferences over time in a partially observable environment (see Chapter 4). Specifically, we leveraged an Estimator to track and estimate the users' preferences, a Generator to match the estimated preferences with the candidate items to rank the next recommendations (with a post-filter to remove repeated recommendations), and an Evaluator to judge the quality of the estimated preferences considering the users' historical feedback. Our experiments in Chapter 4 validated **Research Topic 1** by showing that our proposed EGE model can achieve significantly enhanced performances compared to the strongest baseline model (i.e. MBPI), as shown in Figures 4.3 & 4.4 and Table 4.1.

- **Research Topic 2:** *By mitigating the multi-modal sequence dependency issue in the multi-modal conversational recommendation process, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time with an RNN-enhanced Transformer structure for state tracking*. To address **Challenge 2**, we proposed a novel multi-modal recurrent attention network (MMRAN) model for multi-modal interactive recommendation to effectively incorporate the users' preferences over time. Specifically, we leveraged a gated recurrent network (GRN) with a feedback gate to separately process the natural-language feedback and visual recommendations into hidden states (i.e. representations of the past interactions) for multi-modal sequence combination, as well as a multi-head attention network (MAN) to refine the previously generated hidden states by the GRN component to further track the dialog states of the users' preferences. Our experiments in Chapter 5 validated **Research Topic 2** by showing that our proposed MMRAN model achieves significantly enhanced performances compared to the strongest baseline models (see Table 5.1). Our reported results showed that the MMRAN model benefits from the capability of GRN in combining multi-modal dialog sequences and from the MAN's structure to effectively track the dialog states (see Figures 5.5 and 5.6).

- **Research Topic 3:** *By decoupling the policy optimisation and the multi-modal composition representation learning with goal-oriented reinforcement learning, the multi-modal conversational recommender system can effectively incorporate the users' preferences over time with a composition network and a multi-task learning approach.* To address **Challenge 3**, 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. Our experiments in Chapter 6 validated **Research Topic 3** by showing that our proposed GOMMIR model achieves better overall performances compared to the best baseline models (see Figure 6.5 and Table 6.1).

- **Research Topic 4:** *By modelling the multi-modal conversational recommendation process with both the users' interaction history and the users' instant natural-language feedback, the multi-modal conversational recommender system can effectively incorporate both the users' long-term preferences and short-term needs into the personalised recommendations*. To address **Challenge 4**, 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. Our experiments in Chapter 7 validated **Research Topic 4** by showing that our proposed PMMIR model variants achieve significantly better performances compared to the best baseline models (see Figure 7.4 and Table 7.2). 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 (see Tables 7.2 and 7.4). In addition, the results show that both cold-start and warm start users can benefit from our proposed PMMIR model (see Section 7.3).

- **Research Topic 5:** *To make the multi-modal conversational recommendation task more realistic, we ensure the realism of simulated conversations by considering positive/negative natural-language feedback.* To address **Challenge 5**, we first investigated the effectiveness of the multi-modal conversational recommendation models with both positive and negative natural-language feedback. To make the conversational recommendation task more realistic with both positive and negative natural-language feedback, we proposed an approach to generate both the positive and negative natural-language critiques about the recommendations with the existing user simulator for relative captioning. Our experiments in Chapter 8 validated **Research Topic 4** by showing that positive feedback is more informative relating to the users' preferences in comparison to negative feedback (see Figure 8.3 and Table 8.1). Our reported results also showed that the types of users' natural-language feedback (i.e. different combinations of positive and negative feedback, see Figure 8.4) and the types of textual encoding mechanisms (i.e. pre-trained contextual embeddings and one-hot embeddings, see Figure 8.5) can greatly affect the performance of the both tested models (i.e. DM & MIT).

In summary, we have validated each of the claims of our thesis statement in Section 1.3. We have shown that we can effectively tracking and estimating the users' dynamic preferences from the multi-modal conversational recommendations by leveraging multi-modal pre-trained models for representation encoding and composition (such as ResNet, BERT, CLIP, and TIRG), nerual networks for dialog state tracking (such as GRU, Transformer, and RNN-enhanced Transformer (see Chapter 5)) and reinforcement learning (such as self-supervised reinforcement learning (SSRL, see Chapter 4), goal-oriented reinforcement learning (GORL, see Chapter 6), and hierarchical reinforcement learning (HRL, see Chapter 7)) approaches. Furthermore, our described multi-modal conversational recommendation framework in Chapter 3 allows to investigate the impact of different types of users' feedback on the multi-modal conversational recommendation task (see Chapter 8). Next, we describe some future research directions for multi-modal conversational recommendations in Section 9.2.

9.2 Directions for Future Work

In this section, we discuss possible future directions that could benefit re-framing and developing more effective and realistic multi-modal conversational recommender systems (MMCRS).

- **Mixed-initiative MMCRSs:** Across Chapters 4 to 8, we have only investigated a critiquing-based multi-modal conversational recommendation framework with "systems recommending visual recommendations, cold/warm-start users providing natural-language feedback". However, the interactions between users and recommender systems can involve complex mixed-initiated question-answering processes (Zamani et al., 2022). In particular, it is more natural and realistic to allow users to ask questions about the details of the recommendations and allow recommender systems to ask clarification questions to more effectively collect the users' preferences.

- **Generation-based MMCRSs:** Across Chapters 4 to 8, we only investigated retrieval-based multi-modal conversational recommendation framework by matching candidates with an estimated preference representation. Recent advances in generative artificial intelligence (GAI), such as ChatGPT and Diffusion models, have greatly enhanced the generation abilities in both image and texts. As we mentioned in Section 2.3, generation-based CRSs have the advantage of being able to generate novel responses, which allows them to handle a wider range of inputs. To this end, it is promising to leverage the recent generative AI techniques (such as Llama 2 (Touvron et al., 2023), reinforcement learning from human feedback (RLHF) (Christiano et al., 2017), and retrieval-augmented generation (RAG) (Lewis et al., 2020)) for modelling the multi-modal conversational recommendation task.

- **LLMs powered autonomous agents as MMCRSs:** In the scope of this thesis, we have only investigated a multi-modal conversational recommendation framework by tracking and estimating the users' preferences from sequences of visual recommendations and natural-language feedback. In particular, the core component of the current MMCRSs is a multi-modal dialog state tracker based on GRU, Transformer or RNN-enhanced Transformer. However, the current MMCRSs are not able to plan the whole conversational recommendation process, memorise the users' all interaction history, and use tools for accessing extra information. Recently, with the rapid development of LLMs, LLMs have been leveraged as autonomous agents to perform tasks automatically with abilities of planning (i.e. task decomposition to "think step by step" and self-reflection by refining past action decisions and correcting previous mistakes), memory (short-term memory as in-context learning and long-term memory as the external vector store), and using tools (by calling external APIs for extra information) (L. Wang et al., 2023; Weng, 2023). To this end, it is interesting and promising to leverage LLMs-powered autonomous agents as MMCRSs.

9.3 Concluding Remarks

This thesis has investigated a challenging task: multi-modal conversational recommendation. In particular, this thesis contributed to re-framing and developing more effective multi-modal conversational recommender systems by leveraging advanced techniques, such as multi-modal learning, deep learning, and reinforcement learning. However, there are still interesting research directions in this area, as we have shown some of them in Section 9.2. This work provides a solid motivation and the groundwork for exploring these further research directions in the future. We believe that the development of realistic and effective multi-modal conversational recommender systems will continue to benefit users by effectively satisfying their information needs.

References

Afsar, M. M., Crump, T., & Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, *55*(7), 1–38.

Aggarwal, C. C., et al. (2016). *Recommender systems* (Vol. 1). Springer.

Akerkar, R., & Sajja, P. (2009). *Knowledge-based systems*. Jones & Bartlett Publishers.

Altman, E. (1999). *Constrained markov decision processes*. CRC Press.

Antognini, D., & Faltings, B. (2021). Fast multi-step critiquing for vae-based recommender systems. In *Proc. RecSys* (pp. 209–219).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baldrati, A., Bertini, M., Uricchio, T., & Del Bimbo, A. (2022a). Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proc. CVPR* (pp. 4959–4968).

Baldrati, A., Bertini, M., Uricchio, T., & Del Bimbo, A. (2022b). Effective conditioned and composed image retrieval combining clip-based features. In *Proc. CVPR* (pp. 21466–21474).

Batmaz, Z., Yurekli, A., Bilge, A., & Kaleli, C. (2019). A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, *52*(1), 1–37.

Beard, R., Das, R., Ng, R. W., Gopalakrishnan, P. K., Eerens, L., Swietojanski, P., & Miksik, O. (2018). Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proc. CoNLL* (pp. 251–259).

Berg, T. L., Berg, A. C., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy web data. In *Proc. ECCV* (pp. 663–676).

Bi, K., Ai, Q., Zhang, Y., & Croft, W. B. (2019). Conversational product search based on negative feedback. In *Proc. CIKM* (pp. 359–368).

Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of library and information systems*, *69*(Supplement 32), 175–186.

Chakraborty, S., Hoque, M., Rahman Jeem, N., Biswas, M. C., Bardhan, D., Lobaton, E., et al. (2021). Fashion recommendation systems, models and methods: A review. *Informatics*, *8*(3), 49.

Chen, C.-M., Wang, C.-J., Tsai, M.-F., & Yang, Y.-H. (2019). Collaborative similarity embedding for recommender systems. In *Proc. WWW* (pp. 2637–2643).

Chen, H., Lin, Y., Pan, M., Wang, L., Yeh, C.-C. M., Li, X., ... Yang, H. (2022). Denoising self-attentive sequential recommendation. In *Proc. RecSys* (pp. 92–101).

Chen, L., De Gemmis, M., Felfernig, A., Lops, P., Ricci, F., & Semeraro, G. (2013). Human decision making and recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(3), 1–7.

Chen, L., & Pu, P. (2012). Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, *22*, 125–150.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., & Chi, E. H. (2019). Top-k off-policy correction for a reinforce recommender system. In *Proc. WSDM* (pp. 456–464).

Chen, M., Chang, B., Xu, C., & Chi, E. H. (2021). User response models to improve a reinforce recommender system. In *Proc. WSDM* (pp. 121–129).

Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., & Tang, J. (2019). Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.

Chen, X., Yao, L., McAuley, J., Zhou, G., & Wang, X. (2021). A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. *arXiv preprint arXiv:2109.03540*.

Chen, Y., Gong, S., & Bazzani, L. (2020). Image search with text feedback by visiolinguistic attention learning. In *Proc. CVPR* (pp. 3001–3011).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Proc. NeurIPS* (Vol. 30).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Colas, C., Karch, T., Sigaud, O., & Oudeyer, P.-Y. (2022). Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, *74*, 1159–1199.

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proc. RecSys* (pp. 39–46).

Dancker, J. (2022). A brief introduction to recurrent neural networks: An introduction to rnn, lstm, and gru and their implementation. *Towards Data Science*. Retrieved from `https://towardsdatascience.com/a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4`

Deldjoo, Y., Nazary, F., Ramisa, A., Mcauley, J., Pellegrini, G., Bellogin, A., & Di Noia, T. (2022). A review of modern fashion recommender systems. *arXiv preprint arXiv:2202.02757*.

Deldjoo, Y., Trippas, J. R., & Zamani, H. (2021). Towards multi-modal conversational information seeking. In *Proc. SIGIR* (pp. 1577–1587).

de Souza Pereira Moreira, G., Rabhi, S., Lee, J. M., Ak, R., & Oldridge, E. (2021). Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proc. RecSys* (pp. 143–153).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT* (pp. 4171–4186).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *JAIR*, *13*, 227–303.

Donkers, T., Loepp, B., & Ziegler, J. (2017). Sequential user-based recurrent neural network recommendations. In *Proc. RecSys* (pp. 152–160).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.

Doyle, J. C., Francis, B. A., & Tannenbaum, A. R. (2013). *Feedback control theory*. Courier Corporation.

Ekstrand, M. D., Chaney, A., Castells, P., Burke, R., Rohde, D., & Slokom, M. (2021). Simurec: Workshop on synthetic data and simulation methods for recommender systems research. In *Proc. RecSys* (p. 803–805).

Eysenbach, B., Zhang, T., Salakhutdinov, R., & Levine, S. (2022). Contrastive learning as goal-conditioned reinforcement learning. In *Proc. NeurIPS*.

Felfernig, A., Friedrich, G., Jannach, D., & Zanker, M. (2015). Constraint-based recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 161–190). Boston, MA: Springer US.

Fu, Z., Xian, Y., Zhang, Y., & Zhang, Y. (2020). Tutorial on conversational recommendation systems. In *Proc. RecSys* (pp. 751–753).

Gangwani, T., Lehman, J., Liu, Q., & Peng, J. (2020). Learning belief representations for imitation learning in pomdps. In *Proc. UAI* (pp. 1061–1071).

Gao, C., Lei, W., He, X., de Rijke, M., & Chua, T.-S. (2021). Advances and challenges in conversational recommender systems: A survey. *AI Open*, *2*, 100-126.

Ge, X., Chen, F., Jose, J. M., Ji, Z., Wu, Z., & Liu, X. (2021). Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proc. MM* (pp. 5185–5193).

Gkoumas, D., Li, Q., Lioma, C., Yu, Y., & Song, D. (2021). What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, *66*, 184–197.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Greco, C., Suglia, A., Basile, P., & Semeraro, G. (2017). Converse-et-impera: Exploiting deep learning and hierarchical reinforcement learning for conversational recommender systems. In *Proc. AI*IA* (pp. 372–386).

Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., & Feris, R. (2018). Dialog-based interactive image retrieval. In *Proc. NeurIPS* (pp. 678–688).

Hao, J., Wang, X., Yang, B., Wang, L., Zhang, J., & Tu, Z. (2019). Modeling recurrence for transformer. *arXiv preprint arXiv:1904.03092*.

Haque, A., & Wang, H. (2022). Rethinking conversational recommendations: Is decision tree all you need? *arXiv:2208.14614*.

Hasselt, H. (2010). Double q-learning. *Proc. NeurIPS*, 2613–2621.

Hausknecht, M., & Stone, P. (2015). Deep recurrent q-learning for partially observable mdps. In *Proc. AAAI* (pp. 29–37).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. CVPR* (pp. 770–778).

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. In *Proc. WWW* (pp. 173–182).

Hidasi, B., & Karatzoglou, A. (2018). Recurrent neural networks with top-k gains for session-based recommendations. In *Proc. CIKM* (pp. 843–852).

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. In *Proc. ICLR.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

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

Hu, C., Huang, S., Zhang, Y., & Liu, Y. (2022). Learning to infer user implicit preference in conversational recommendation. In *Proc. SIGIR* (pp. 256–266).

Huang, J., Oosterhuis, H., Cetinkaya, B., Rood, T., & de Rijke, M. (2022). State encoders in reinforcement learning for recommendation: A reproducibility study. In *Proc. SIGIR* (pp. 2738–2748).

Hutsebaut-Buysse, M., Mets, K., & Latré, S. (2022). Hierarchical reinforcement learning: A survey and open research challenges. *Machine Learning and Knowledge Extraction*, *4*(1), 172–221.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., & Whiteson, S. (2018). Deep variational reinforcement learning for pomdps. In *Proc. ICML* (pp. 2117–2126).

Iovine, A., Narducci, F., & Semeraro, G. (2020). Conversational recommender systems and natural language:: A study through the converse framework. *Decision Support Systems*, *131*, 113250.

Jadidinejad, A. H., Macdonald, C., & Ounis, I. (2020). Using exploration to alleviate closed loop effects in recommender systems. In *Proc. SIGIR* (pp. 2025–2028).

Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, *54*(5), 1–36.

Jannach, D., Quadrana, M., & Cremonesi, P. (2022). Session-based recommender systems. In *Recommender systems handbook* (pp. 301–334). Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422–446.

Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. In *Proc. ICDM* (pp. 197–206).

Kim, S., Lin, S., Jeon, S., Min, D., & Sohn, K. (2018). Recurrent transformer networks for semantic correspondence. *arXiv preprint arXiv:1810.12155*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proc. ICLR*.

Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Proc. NeurIPS* (pp. 1008–1014).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*(8), 30–37.

Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., & Srinivas, A. (2020). Reinforcement learning with augmented data. In *Proc. NeurIPS* (pp. 19884–19895).

Laskin, M., Srinivas, A., & Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. In *Proc. ICML* (pp. 5639–5650).

Latifi, S., Mauro, N., & Jannach, D. (2021). Session-aware recommendation: A surprising quest for the state-of-the-art. *Information Sciences*, *573*, 291–315.

Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T. L., & Bansal, M. (2020). Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Lei, W., He, X., Miao, Y., Wu, Q., Hong, R., Kan, M.-Y., & Chua, T.-S. (2020). Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proc. WSDM* (pp. 304–312).

Levine, S. (2022). Understanding the world through action. In *Proc. CoRL* (pp. 1752–1757).

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proc. NeurIPS* (Vol. 33, pp. 9459–9474).

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proc. WWW* (pp. 661–670).

Liao, L., Long, L. H., Zhang, Z., Huang, M., & Chua, T.-S. (2021). Mmconv: An environment for multimodal conversational search across multiple domains. In *Proc. SIGIR* (pp. 675–

684).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proc. of ACL workshop on text summarization branches out* (pp. 74–81).

Lin, Y., Lin, F., Zeng, W., Xiahou, J., Li, L., Wu, P., … Miao, C. (2022). Hierarchical reinforcement learning with dynamic recurrent mechanism for course recommendation. *Knowledge-Based Systems*, *244*, 108546.

Lin, Y., Liu, Y., Lin, F., Wu, P., Zeng, W., & Miao, C. (2021). A survey on reinforcement learning for recommender systems. *arXiv preprint arXiv:2109.10665*.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), 76–80.

Liu, M., Zhu, M., & Zhang, W. (2022). Goal-conditioned reinforcement learning: Problems and solutions. In *Proc. IJCAI* (pp. 5502–5511).

Liu, T., Fang, S., Zhao, Y., Wang, P., & Zhang, J. (2015). Implementation of training convolutional neural networks. *arXiv preprint arXiv:1506.01195*.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., … others (2023). Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR* (pp. 1096–1104).

Lu, Z., & Yang, Q. (2016). Partially observable markov decision process for recommender systems. *arXiv preprint arXiv:1608.07793*.

Luo, R., Price, B., Cohen, S., & Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. *arXiv:1803.04376*.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Manotumruksa, J., Macdonald, C., & Ounis, I. (2018). A contextual attention recurrent architecture for context-aware venue recommendation. In *Proc. SIGIR* (pp. 555–564).

Manzoor, A., & Jannach, D. (2021). Generation-based vs. retrieval-based conversational recommendation: A user-centric comparison. In *Proc. RecSys* (pp. 515–520).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proc. NeurIPS* (pp. 3111–3119).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proc. DL* (pp. 195–204).

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning*

Systems, *32*(2), 604–624.

Parr, R., & Russell, S. (1997). Reinforcement learning with hierarchies of machines. In *Proc. NeurIPS.*

Pateria, S., Subagdja, B., Tan, A.-h., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, *54*(5), 1–35.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proc. EMNLP* (pp. 1532–1543).

Petrov, A., & Macdonald, C. (2022). A systematic review and replicability study of bert4rec for sequential recommendation. In *Proc. RecSys* (pp. 436–447).

Quadrana, M., Cremonesi, P., & Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, *51*(4), 1–36.

Quadrana, M., Karatzoglou, A., Hidasi, B., & Cremonesi, P. (2017). Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proc. RecSys* (pp. 130–137).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others (2021). Learning transferable visual models from natural language supervision. In *Proc. ICML* (pp. 8748–8763).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rafati, J., & Noelle, D. C. (2019). Learning representations in model-free hierarchical reinforcement learning. In *Proc. AAAI* (Vol. 33, pp. 10009–10010).

Rendle, S. (2010). Factorization machines. In *2010 IEEE International Conference on Data Mining* (pp. 995–1000).

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). Bpr: Bayesian personalized ranking from implicit feedback. *arXiv:1205.2618*.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *Proc. CVPR* (pp. 7008–7024).

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Springer.

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, *6*(12), 310–316.

Shi, Y., Karatzoglou, A., Baltrunas, L., Larson, M., Oliver, N., & Hanjalic, A. (2012). Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proc. RecSys* (pp. 139–146).

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *Proc. ICML* (pp. 387–395).

Simrock, S. (2011). Tutorial on control theory. In *Proc. ICAELEPCS* (pp. 10–14).

Smirnova, E., & Vasile, F. (2017). Contextual sequence modeling for recommendation with recurrent neural networks. In *Proc. of RecSys workshop on Deep Learning* (pp. 2–9).

Smyth, B. (2007). Case-based recommendation. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 342–376). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stefani, R. T., Shahian, B., Savant, C. J., & Hostetter, G. H. (2002). *Design of feedback control systems*. Oxford University Press Oxford.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proc. CIKM* (pp. 1441–1450).

Sun, Y., & Zhang, Y. (2018). Conversational recommender system. In *Proc. SIGIR* (pp. 235–244).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Proc. NeurIPS*.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181–211.

Tan, Z.-X., Goel, A., Nguyen, T.-S., & Ong, D. C. (2019). A multimodal lstm for predicting listener empathic responses over time. In *Proc. FG* (pp. 1–4).

Tang, J., & Wang, K. (2018). Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proc. WSDM* (pp. 565–573).

Taud, H., & Mas, J. (2018). Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, 451–455.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Uppal, S., Bhagat, S., Hazarika, D., Majumder, N., Poria, S., Zimmermann, R., & Zadeh, A. (2021). Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, *77*, 149–171.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Proc. NeurIPS* (pp. 5998–6008).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proc. CVPR* (pp. 4566–4575).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proc. CVPR* (pp. 3156–3164).

Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.-J., Fei-Fei, L., & Hays, J. (2019). Composing text and image for image retrieval-an empirical odyssey. In *Proc. CVPR* (pp. 6439–6448).

Wang, J., Ding, K., & Caverlee, J. (2021). Sequential recommendation for cold-start users with meta transitional learning. In *Proc. SIGIR* (pp. 1783–1787).

Wang, K., Zou, Z., Deng, Q., Tao, J., Wu, R., Fan, C., ... Cui, P. (2021). Reinforcement learning with a disentangled universal value function for item recommendation. In *Proc. AAAI* (pp. 4427–4435).

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... others (2023). A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M. A., & Lian, D. (2021). A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, *54*(7), 1–38.

Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q. Z., & Orgun, M. (2019). Sequential recommender systems: challenges, progress and prospects. In *Proc. IJCAI* (pp. 6332–6338).

Wang, W., Xu, Y., Shen, J., & Zhu, S.-C. (2018). Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proc. CVPR* (pp. 4271–4280).

Wang, Z., Ma, Y., Liu, Z., & Tang, J. (2019). R-transformer: Recurrent neural network enhanced transformer. *arXiv preprint arXiv:1907.05572*.

Weng, L. (2018). Attention? attention! *lilianweng.github.io*. Retrieved from `https://lilianweng.github.io/posts/2018-06-24-attention/`

Weng, L. (2023, Jun). Llm-powered autonomous agents. *lilianweng.github.io*. Retrieved from `https://lilianweng.github.io/posts/2023-06-23-agent/`

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

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., & Feris, R. (2021). Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proc. CVPR* (pp. 11307–11317).

Wu, Y., Liao, L., Zhang, G., Lei, W., Zhao, G., Qian, X., & Chua, T.-S. (2022). State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*.

Wu, Y., Macdonald, C., & Ounis, I. (2021). Partially observable reinforcement learning for dialog-based interactive recommendation. In *Proc. RecSys* (p. 241–251).

Wu, Y., Macdonald, C., & Ounis, I. (2022a). Multimodal conversational fashion recommendation with positive and negative natural-language feedback. In *Proc. CUI* (pp. 1–10).

Wu, Y., Macdonald, C., & Ounis, I. (2022b). Multi-modal dialog state tracking for interactive fashion recommendation. In *Proc. RecSys* (pp. 124–133).

Wu, Y., Macdonald, C., & Ounis, I. (2023). Goal-oriented multi-modal interactive recommendation with verbal and non-verbal relevance feedback. In *Proc. RecSys* (pp. 362–373).

Xie, R., Zhang, S., Wang, R., Xia, F., & Lin, L. (2021). Hierarchical reinforcement learning for integrated recommendation. In *Proc. AAAI* (Vol. 35, pp. 4521–4528).

Xin, X., Karatzoglou, A., Arapakis, I., & Jose, J. M. (2020). Self-supervised reinforcement learning for recommender systems. In *Proc. SIGIR* (pp. 931–940).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML* (pp. 2048–2057).

Xu, K., Yang, J., Xu, J., Gao, S., Guo, J., & Wen, J.-R. (2021). Adapting user preference to online feedback in multi-round conversational recommendation. In *Proc. WSDM* (pp. 364–372).

Yarats, D., Kostrikov, I., & Fergus, R. (2020). Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *Proc. ICLR*.

Yu, T., Shen, Y., & Jin, H. (2019). A visual dialog augmented interactive recommender system. In *Proc. KDD* (pp. 157–165).

Yu, T., Shen, Y., & Jin, H. (2020). Towards hands-free visual dialog interactive recommendation. In *Proc. AAAI* (Vol. 34, pp. 1137–1144).

Yu, T., Shen, Y., Zhang, R., Zeng, X., & Jin, H. (2019). Vision-language recommendation via attribute augmented multimodal reinforcement learning. In *Proc. MM* (pp. 39–47).

Yuan, Y., & Lam, W. (2021). Conversational fashion image retrieval via multiturn natural language feedback. In *Proc. SIGIR* (p. 839–848).

Zamani, H., Trippas, J. R., Dalton, J., & Radlinski, F. (2022). Conversational information seeking. *arXiv preprint arXiv:2201.08808*.

Zhang, R., Yu, T., Shen, Y., & Jin, H. (2022). Text-based interactive recommendation via offline reinforcement learning. In *Proc. AAAI* (Vol. 36, pp. 11694–11702).

Zhang, R., Yu, T., Shen, Y., Jin, H., & Chen, C. (2019). Text-based interactive recommendation via constraint-augmented reinforcement learning. In *Proc. NeurIPS* (pp. 15214–15224).

Zhang, S., & Balog, K. (2020). Evaluating conversational recommender systems via user simulation. In *Proc. KDD* (pp. 1512–1520).

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, *52*(1), 1–38.

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proc. CIKM* (pp. 177–186).

Zhao, D., Zhang, L., Zhang, B., Zheng, L., Bao, Y., & Yan, W. (2020). Mahrl: Multi-goals abstraction based deep hierarchical reinforcement learning for recommendations. In *Proc. SIGIR* (pp. 871–880).

Zhao, X., Xia, L., Zhang, L., Ding, Z., Yin, D., & Tang, J. (2018). Deep reinforcement learning for page-wise recommendations. In *Proc. RecSys* (pp. 95–103).

Zhao, X., Zhang, L., Ding, Z., Xia, L., Tang, J., & Yin, D. (2018). Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proc. KDD* (pp. 1040–1048).

Zhao, X., Zhang, L., Xia, L., Ding, Z., Yin, D., & Tang, J. (2017). Deep reinforcement learning for list-wise recommendations. *arXiv preprint arXiv:1801.00209*.

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., & Li, Z. (2018). Drn: A deep reinforcement learning framework for news recommendation. In *Proc. WWW* (pp. 167–176).

Zheng, L., Noroozi, V., & Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *Proc. WSDM* (pp. 425–434).

Zheng, Y., Liu, S., Li, Z., & Wu, S. (2021). Cold-start sequential recommendation via meta learner. In *Proc. AAAI* (pp. 4706–4713).

Zhu, Y., Li, H., Liao, Y., Wang, B., Guan, Z., Liu, H., & Cai, D. (2017). What to do next: Modeling user behaviors by time-lstm. In *Proc. IJCAI* (pp. 3602–3608).

Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., & Yin, D. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proc. KDD* (pp. 2810–2818).

Zou, L., Xia, L., Gu, Y., Zhao, X., Liu, W., Huang, J. X., & Yin, D. (2020). Neural interactive collaborative filtering. In *Proc. SIGIR* (pp. 749–758).