State Graph Reasoning for Multimodal Conversational Recommendation

Yuxia Wu, Lizi Liao*, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian*, Tat-Seng Chua

.

***Abstract*—Conversational recommendation system (CRS) attracts increasing attention in various application domains such as retail and travel. It offers an effective way to capture users' dynamic preferences with multi-turn conversations. However, most current studies center on the recommendation aspect while over-simplifying the conversation process. The negligence of complexity in data structure and conversation flow hinders their practicality and utility. In reality, there exist various relationships among slots and values, while users' requirements may dynamically adjust or change. Moreover, the conversation often involves visual modality to facilitate the conversation. These actually call for a more advanced internal state representation of the dialogue and a proper reasoning scheme to guide the decision making process.**

In this paper, we explore multiple facets of multimodal conversational recommendation and try to address the above mentioned challenges. In particular, we represent the structured back-end database as a multimodal knowledge graph which captures the various relations and evidence in different modalities. The user preferences expressed via conversation utterances will then be gradually updated to the state graph with clear polarity. Based on these, we train an end-to-end *S*tate *G*raph-based *R*easoning model (SGR) to perform reasoning over the whole state graph. The prediction of our proposed model benefits from the structure of the graph. It not only allows for zero-shot reasoning for items unseen in training conversations, but also provides a natural way to explain the policies. Extensive experiments show that our model achieves better performance compared with existing methods.

***Index Terms*—Recommendation systems, conversation, knowledge graph**

I. Introduction

CONVERSATIONAL recommendation system (CRS) has become an emerging research topic in information seeking. It integrates the strength of recommendation systems and conversation techniques. In general, recommendation systems predict users' preferences towards items by analyzing their past behaviors such as click history, visit log, and ratings on items, *etc*, which are widely applied in many domains such as While with the help of multi-turn conversations, the system can further capture the detailed and dynamic preferences of users, which may lead to better recommendation results and user experience [28].

There have been many efforts centered on integrating conversation modelling into recommendation systems. From a broader perspective, pioneering works emerge from tag-based interaction between user and systems where the interaction is mainly realized by tags [5]. To further improve the convenience of developed systems, more efforts focus on multi-turn conversations with natural language as both system's input and output [16, 32, 41]. Generally speaking, existing methods have emphasized on three broad directions: a) Since the key advantage of CRS is being able to ask questions, a line of studies work on learning to ask appropriate attributes/topics/categories of items to narrow down the candidate items [5, 38, 39, 42]; b) Another line of efforts target at learning better strategy for making successful recommendations with less turns of interactions [14, 15, 32]; c) There are also works that further delve into in-depth dialogue understanding and response generation [4, 16, 21, 37, 41].

However, there exist several shortages of these current methods. First, most of these methods directly generate responses via action prediction and entity linking, while ignoring the various relationships among slot values and its relation to the explicit dialogue state representations [14, 16], shown as Fig. 1(a). This would lead to less informative representation of user preference and thus harm the recommendation performance. For example, when one user wants to find a cheap item, it indicates that the user has negative preference on the other price values. Second, the modelling of dynamic user preference change is relatively weak. Although some efforts try to achieve this with graph based methods [7, 15, 23, 37] as shown in Fig. 1(b), they fail to represent the dynamic change in an effective way. For example, the model in [15] simply deleted the mentioned attributes or items negated in historical turns without updating any user or item representations. Meanwhile, the model in [37] only considered the most updated user requirements for each slot without remembering any former denied ones.[7] Last but not the least, most of the existing works focus on textual conversations. However, there is a growing demand for multimodal conversations to facilitate recommendation in domains like e-commerce retail and travel. Although there are some initial works [19, 25, 28], the oversimplified usage of image information hinders the image

This work was supported in part by the scholarship from China Scholarship Council (CSC) under Grant 202006280325; in part by the NSFC, China under Grants 61902309, 61701391, and 61772407; in part by ShaanXi Province under Grant 2018JM6092; in part by the Fundamental Research Funds for the Central Universities, China (xxj022019003); in part by China Postdoctoral Science Foundation (2020M683496); in part by the National Postdoctoral Innovative Talents Support Program, China (BX20190273); and in part by the Science and Technology Program of Xi'an, China under Grant 21RGZN0017.

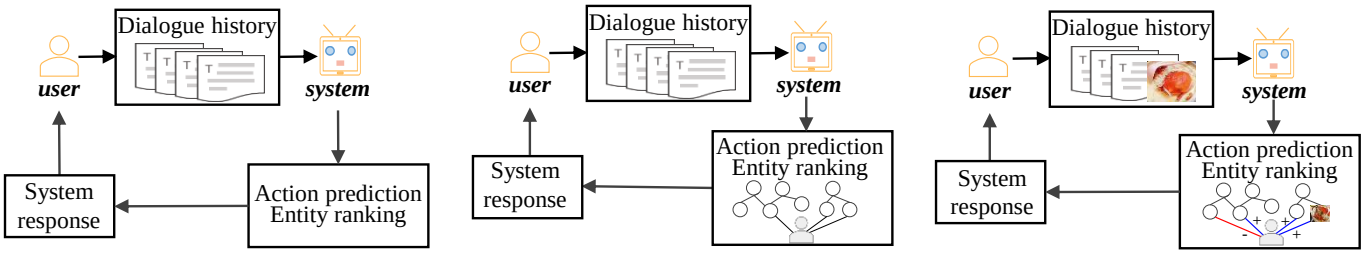
Yuxia Wu, Guoshuai Zhao and Xueming Qian are with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University (e-mail: wuyuxia@stu.xjtu.edu.cn, guoshuai.zhao@xjtu.edu.cn qianxm@mail.xjtu.edu.cn)

Lizi Liao is with the Singapore Management University (e-mail: liaolizi.llz@gmail.com).

Gangyi Zhang is with the University of Science and Technology of China (e-mail: gangyi.zhang@outlook.com)

Wenqiang Lei and Tat-Seng Chua are with the National University of Singapore (e-mail: wenqianglei@gmail.com, dcscts@nus.edu.sg)

* Lizi Liao and Xueming Qian are the corresponding authors.



(a) Corpus-based methods [14, 16] (b) Graph-based methods [7, 15, 23, 37] (c) Our method

Fig. 1: The workflow of the existing methods and ours.

modality's contribution in conversational recommendation.

To address the above-mentioned issues, we focus on a new scheme as shown in Fig. 1(c). We explore the complex relationship among slots and values in the back-end item database and represent the structured data into a multimodal knowledge graph. Then the graph is updated with user preferences represented in the dialogue states with the conversation goes on. Based on which, the actions to take, such as recommendation or further inquiry, can be generated via performing reasoning on the state graph. Specifically, the state graph is initiated as a signed graph containing positive and negative links to model the complicated relationships among items, slots and values, as well as the rich modalities of information in the back-end database. Then we gradually update the state graph to capture the dynamic user requirements based on the user intention harvested from the conversation history. The state graph is updated in an explicit way by adding, deleting or changing the links between the user and and other nodes based on the dynamic user preference. Basically, the state graph keeps track of the conversation progress and serves as a base for performing reasoning about entity ranking of the nodes. We then train an end-to-end graph reasoning model SGR to infer reasoning which explicitly differentiates between positive and negative user preferences. Since the prediction results inherit the graph structure, it caters for cold start venues and provides natural explanations.

We summarize our contributions as follows:

- We explicitly model users' dynamic preferences and integrate it with a multimodal knowledge graph for better state representation. The update of the graph reflects the real change of users' preferences based on both textual and visual modality.
- We design a state graph reasoning model to capture the various evidence in the multimodal knowledge graph, and generate more accurate agent behavior predictions.
- Extensive experiments demonstrate the effectiveness of our proposed method. Qualitative results also show that the proposed method not only handles zero-shot situations well but also offers good explainability.

II. Related Work

Our work is closely related to three lines of researches: conversational recommendation, graph reasoning and multimodal knowledge graph. Here, we will briefly discuss the connections between these lines of research and emphasize the research gap targeted by this work.

A. Conversational Recommendation

Conversational recommendation aims at providing interactive recommendations through dialogues. Compared with traditional static recommender systems, it has the advantage of capturing users' dynamic preferences from the multi-turn utterances [13].

The existing works on conversational recommender systems fall into three broad categories. One line of efforts was largely question driven. They focused on learning to ask attributes/topics/categories of items to reduce the search space [5, 38, 39, 42]. For example, Multi-Memory Network (MMN) [39] was a unified model integrating query/item representation learning and conversational search/recommendation. It learned user preference by asking questions. However, it did not contain any special policy network to decide when to ask or recommend. Another line of studies targeted at better strategy for making successful recommendations in less turns. For instance, Conversational Recommender Model (CRM) [32] applied reinforcement learning to decide when to ask or recommend items based on users' current preference learned by a belief tracker. However, they would only recommend one time, which would fail if the user gave negative feedback on recommendations. To solve these problems, Estimation–Action–Reflection (EAR) [14] was proposed to learn user preferences and rank the items and attributes. A policy network was applied to determine whether to ask attributes further or recommend items. The model would also be updated when the user rejects the recommendations. Later on, the authors further devised an interactive path reasoning mechanism to help ranking of item and attributes [15]. Beyond emphasizing on the recommendation part, the third line of efforts delved into in-depth dialogue understanding and response generation [4, 16, 21, 37, 41]. For example, knowledge graph was introduced in Knowledge-Based Recommender Dialog (KBRD) [4] to bridge the recommender system and the dialogue system in an end-to-end manners. The model linked the entities in dialogue history to the external knowledge graph to enhance the representation of users' preferences. Then the dialogue system can generate responses that are consistent with users' interest. Similarly, the model in [40] incorporated both word-based and entity-based Knowledge Graph (KG) to enhance the semantic representations in CRS. However, these models lack

dialogue state management. In our work, we incorporate the knowledge graph into our internal dialogue state representation and perform reasoning on it to yield better results.

B. Graph Reasoning

With knowledge graph structures, there are a lot of efforts trying to do reasoning over the graph to enhance conversational recommendation performance. Actually, graph reasoning has been successfully applied to many tasks such as social network analysis, question answering, recommendation, and so on. For conversational recommendation, Open-ended Dialog KG (OpenDialKG) [23] learns the optimal path of dialogue context within a large common-sense KG for open-ended dialogue system. The system would ask questions about the attributes of items and also chat with the users. It applies attention-based graph decoder to rank the candidate entities from the KG. During path walking, it prunes unattended paths to effectively reduce the search space. A Simple Conversational Path Reasoning (CPR) [15] is an extension of EAR which utilizes user's attribute feedback explicitly and converts conversational recommendation as an interactive path reasoning problem over graph. It relies on users' historical interaction records to learn user and item representations by offline training. Besides, it only considers the one-hop neighbor entities of the attributes or items while ignoring different slots and values of the items.

To capture users' dynamic preference, researchers proposed user memory reasoning for conversational recommendation [37]. They constructed user memory graph by users' past preferences and the current requests during conversation. When updating the memory graph, they just considered the sentiment relations. Relational Graph Convolutional Network (R-GCN) was applied to learn the hidden state of each entity and predict dialogue action. More recently, an adaptive reinforcement learning framework, namely UNIfied COnversational RecommeNder (UNICORN) was proposed in [7]. The model integrated three separated decision-making processes in conversational recommender systems as a unified policy learning problem. A dynamic weighted graph was designed to capture the sequential information of the dialogue history which is beneficial to learn user's preference on items. However it also only uses weighted entropy to select the candidate slots and values similar to SCPR.

Our work is close to these works but has several key differences. First, the existing work ignores the complicated relations among items, slots and values. The inter-connections among them also provides evidence for reasoning. For instance, when one user prefers cheap venue, it signals negative tendencies towards the venues which are connected to other price categories. To this end, we propose the signed Graph Convolutional Network (GCN) based method to better model the polarity in user preferences. Second, we update the state graph to model the dynamic change of user preferences in an explicit way. Our model performs reasoning over the global state graph instead of local one as in [15].

C. Multimodal Knowledge Graph

As we incorporate multimodal information into knowledge graph to perform reasoning, our work is also closely related to the works on Multimodal Knowledge Graph (MMKG). MMKG integrates multimodal data (such as images and texts) into the knowledge graph and treats the image or text as an entity or an attribute of the entity [20, 33]. In general, MMKG representation learning can be divided into two categories: feature-based methods [24, 36] and entity-based methods [26]. The former kind treated the visual information as the features of entities. They modify TransE model [2] to integrate the visual features of entities. However, this kind of methods requires each entity to provide visual information, which was not suitable for many tasks. The later one [26] constructs the MMKG by adding extra relations on the original KG, such as *hasImage*, *hasDescription*. The multimodal information could be aggregated into its neighbor entity. Then GCN or R-GCN was applied to learn the representations of the entities. For instance, the researchers in [34] introduced Knowledge-driven Multimodal Graph Convolutional Network (KMGCN) to model the semantic representations of textual information, knowledge concepts and visual information for fake news detection. [31] was the first work that incorporated MMKG into recommender systems. In this work, multimodal knowledge graph attention network was proposed to learn the representations of entities. The experiments demonstrated that multimodal features outperform any single-modal features. However, the application of MMKG in conversational recommendation is currently under-explored. In this work, we aim to make use of the MMKG to boost conversational recommendation performance.

III. Methodology

The overall framework is illustrated in Fig. 2. The proposed **SGR** model starts from 1) constructing an MMKG. Then for each dialogue, it 2) updates the MMKG-based state graph turn by turn; and 3) reasons over the state graph for detailed decision making.

Specially, as each conversation begins and goes on, we update the state graph gradually to introduce user preferences expressed in both textual and visual modalities. The update module includes *add*, *change* and *negate* operations, which helps to capture user's dynamically changing requirements in a convenient and explicit way. Based on the up-to-date state graph, we conduct reasoning over it via signed graph convolutional neural networks. It integrates evidence from the interconnections of nodes and captures the preference polarities of user. Guided by the detailed intent actions predicted via pretrained GPT-2 model, the corresponding entities such as the slots, values or venues are then ranked via the learned node representations accordingly. In what follows, we introduce these modules in detail.

A. Constructing MMKG

Different from the current dialogue research using database query for the target, we aim at building graph structure to capture various information for reasoning the target. Hence, we

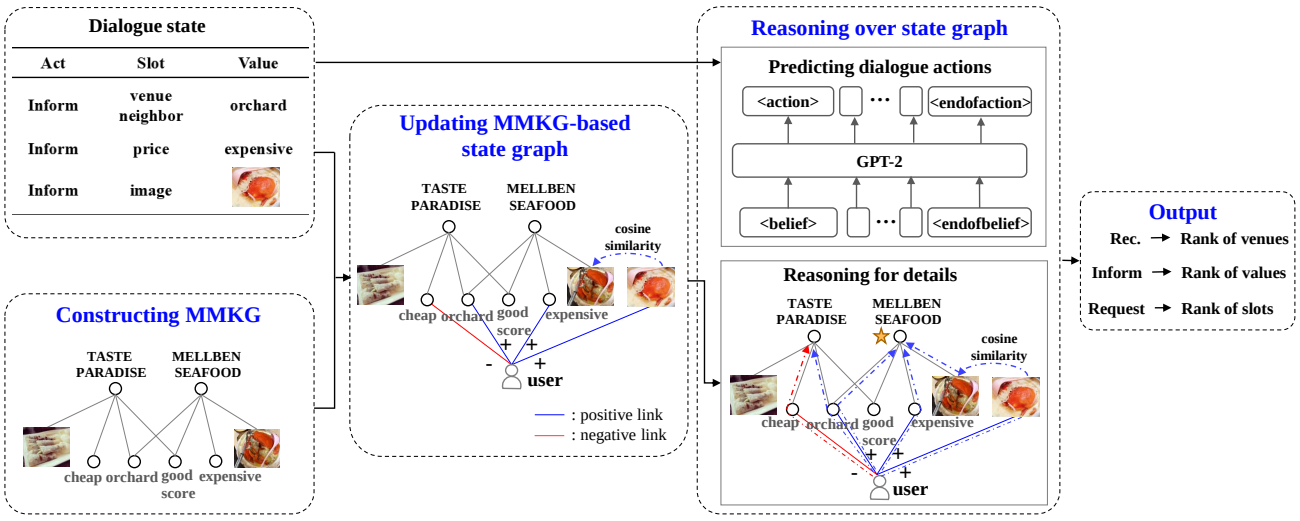


Fig. 2: The overall architecture of the proposed **SGR** model. It first builds the initial MMKG to capture back-end database knowledge. Then based on gradually updated state graph during the conversation process, it performs reasoning over the state graph to generate agent action decisions.

expect it to have the following characteristics: 1) it should have the ability to capture different modalities, such as texts and images; 2) it should be able to represent various relationships in the back-end database, such as exclusive relations among values of the same slot; and 3) we also expect it to be able to handle users' preference polarities such as like or dislike.

To achieve these, we construct the general MMKG to represent the back-end database. Formally, we represent the graph as $G = (\epsilon, \gamma)$, where ϵ denotes the set of nodes, and γ represents the links in the graph. Generally, the nodes represent the items and the attributes in the database. The attributes can be some textual terms or images belonging to the items. To cover different modalities, we follow the entity-based MMKG method [26] to treat the images as the nodes in G. The links connect the items and its attributes. They can be of different types indicating different relationships among attributes. The links can also represent different polarities such as positive or negative links. For example, for the scenario of conversational recommendation on the task of finding places, the nodes are the venues and the slot-value pairs of these venues (here the venues refer to the aforementioned items, and the slot-value pairs refer to the attributes such as *price-cheap*). There exists various relationships among the values for the same slot: 1) slots containing mutually exclusive values (such as price with value candidates cheap, moderate and expensive etc); 2) slots without mutually exclusive values (such as *has image*). To represent the exclusive relationships, we update these link combinations between the venue and the slot-values. For example, if the attributes of one venue contain *price-cheap*, then there will be positive link between the venue and *price-cheap* and negative link between the venue and the other price-value candidates.

B. Updating MMKG-based State Graph

As the conversation begins and goes on, it is essential for the agent to understand users' intention with textual or visual modalities. We thus transform the dialogue states (such as *inform-price-expensive* which represents the action, slot and value, respectively) into state graph initiated by the MMKG and update it gradually as the conversation goes on. We add a node to represent the current user and use the signed links to denote the user preference polarity towards the entities in the MMKG. Note that our constructed MMKG contains information in both modalities, and the user dialogues are also flexible in modality usage. To give a clear view of how to update the state graph, we illustrate the process according to information modality separately.

1) Update Textual Slots: The textual information conveys users' requirements and preferences. It is natural for users to change their requirements as the conversation goes on. Therefore, we update the state graph dynamically including: *add, change, negate*. When the user provides new requirements (see turn 2 in Fig. 3), we add signed links between the user and the corresponding slot-value nodes. When the user changes requirements, we can also change the links for further update. It can be seen that the state graph can effectively reflect users' dynamic preferences in an explicit way. And it is convenient to change the links based on the dialogue state. By adding the signed edges between the user and the attributes, we can show users' like and dislike clearly in the state graph.

2) Update Intention via Image: In our multimodal conversational recommendation task, users usually offer images to express their intention conveniently. To understand users' intention based on images, we apply a layer-by-layer taxonomy-based ResNet [10] classifier to learn the visual features of the images [17]. To update users' intention into the state graph, we compute the cosine similarities between the user provided images and the images in the MMKG. If there exist images in the MMKG with similarity scores exceeding a pre-defined threshold, we will update the state graph by adding a positive link between the user and the image in the MMKG (see turn 3 in Fig. 3). In this way, the user will be quickly connected to a specific venue closely. Also, if the user negates an image of an venue provided by the agent, we will also add an negative

link between the user and the very similar images in MMKG. Then the related venue node would easily receive negative tendencies from the user via the links.

C. Reasoning over State Graph

To facilitate the description of the following modules, we define $B_t = \{b_1, b_2, \cdots, b_m\}$ as the belief states of turn *t*. b_t summarizes the dialogue history up to the current turn *t*. Each state consists of tuples like $\{a, s, v\}$, where *a* is action, *s* is slot and *v* is the value.

For turn *t*, given the historical dialogue state B_t and the previous state graph G_{t-1} of turn *t-1*, the target of our model is to predict a series of tuples $Y = \{y_1, y_2, \cdots, y_n\}$, where $y_i = \{a_i, s_i, v_i\}$. If $a_i = request$, the v_i will be set to *null*. If $a_i = recommend$, the s_i will be set to *null*. We first predict the actions and then get the detail arguments of the corresponding dialogue actions. We introduce the details of each step in the following sections.

1) Predicting Dialogue Actions: The dialogue actions of the agent have a strong dependence on the B_t. Many of the existing works apply classification model to predict the dialogue act for each turn. However, this oversimplifies the conversational recommendation scenarios as demonstrated in [18]. Human agents often perform more than one action in a single turn. As it is unrealistic to pre-define the number of actions to perform in each turn, we cast the action prediction as a sequence generation task, in which the model can automatically decide how many actions in one turn to perform based on the context.

With the development of the NLP techniques, various pre-training models are emerging such as Bert, Transformer, GPT-2 [27]. Inspired by SimpleTOD [12], we recast the action prediction task as a sequence-to-sequence generation problem where B_t is treated as the input and the target actions are treated as the output. The model will learn to automatically generate the end token if it feels confident enough. The pre-trained GPT-2 model is leveraged.

To adapt our input B_t to the GPT-2 model, we first transfer the dialogue states B_t into a sequence containing a list of triplets: $x =$ " $< |belief| > a_1, s_1, v_1; \cdots, a_m, s_m, v_m, < |endofbelief| >$ ". The output is also a sequence: $y =$ " $< |action| > \ a_1, \ \cdots, \ a_n \ < |endofaction| >$ ". A training sequence is the concatenation $[x; y]$ of the *x* and y. We denote it as $g = (g_1, g_2, \cdots, g_{|g|})$. Given a training instance like this, the joint probability of the sequence is calculated as:

$$p(g) = \prod_{i=1}^{|g|} p(g_i | g_{<i}).$$

Given a dataset with $|K|$ training instances, the loss for training the generator (a neural network with parameters θ) is the negative log-likelihood over the whole training data. We aim to minimize the loss as follows:

$$L_A = -\sum_{k=1}^{|\mathcal{K}|} \sum_{i=1}^{n_k} log p_\theta(g_i^k | g_{<i}^k),$$

[1]Here the user provided image is used for linking the similar image in MMKG. On the other side, the content of user provided image is captured in action prediction.

where n_k is the length of the instance g^k.

2) Reasoning for Details: After the action prediction module, the model manages to predict a set of actions $A = (a_1, a_2, \cdots, a_n)$. The next step is to enrich these predicted actions with detailed slots and values. In detail, if a_i is *inform*, then the top-1 slot-value pair (s_i, v_i) will be selected to yield a detailed tuple $(inform, s_i, v_i)$. When a_i is *request*, then the tuple will be $(request, s_i, null)$ where s_i is the top ranked slot. Similarly, when a_i is *recommend*, the tuple will be $(recommend, null, v_i)$ where the v_i is the top ranked venue.

To do reasoning, we first learn the node representations of the state graph. Considering that the graph has both positive and negative links, it is not suitable to apply the traditional GCN. Actually, to deal with this problem, researchers have carried out extensive explorations and proposed signed GCN [8]. To properly integrate the positive and negative tendencies during the aggregation process, we leverage multiple layers of signed GCN [8] over the graph and obtain the hidden representations of all nodes.

In the signed state graph, for each node ϵ_i there are two kinds of neighbors: positive-linked neighbors $\mathbb{N}_i^+$ and negative-linked neighbors $\mathbb{N}_i^-$. It is not sufficient to learn one single representation for each node as in traditional unsigned GCN. Thus we maintain two kinds of representations of each node. We define h_i^P and h_i^Q as the positive and negative representations of ϵ_i, respectively. The representations are aggregated layer by layer. To incorporate the signed information during aggregation, we follow the balanced theory mentioned in [8]. The aggregation of the neighbor information comes from two parts: the information from the neighbors $\mathbb{N}_i^+$ and the information from the neighbors $\mathbb{N}_i^-$. The detail aggregation process is shown as follows:

$$h_i^{P(l)} = \sigma(W^{P(l)}[\sum_{j \in \mathbb{N}_i^+} \frac{h_j^{P(l-1)}}{|\mathbb{N}_i^+|}, \sum_{k \in \mathbb{N}_i^-} \frac{h_k^{Q(l-1)}}{|\mathbb{N}_i^-|}, h_i^{P(l-1)}]),$$

$$h_i^{Q(l)} = \sigma(W^{Q(l)}[\sum_{j \in \mathbb{N}_i^+} \frac{h_j^{Q(l-1)}}{|\mathbb{N}_i^+|}, \sum_{k \in \mathbb{N}_i^-} \frac{h_k^{P(l-1)}}{|\mathbb{N}_i^-|}, h_i^{Q(l-1)}]),$$

where $h_i^{P(l)}$ and $h_i^{Q(l)}$ are the positive and negative representations at layer l respectively; and $W^{P(1)}$ and $W^{Q(1)}$ are the linear transformation matrices. The aggregation starts from h_i^0 which is the initial representation of ϵ_i.

The final representation of the node ϵ_i is the concatenation of the positive and negative representations:

$$h_i^l = [h_i^{P(l)}, \ h_i^{Q(l)}].$$

Then we calculate the loss L_H of node representation learning like [8]. The loss is designed to capture the relationships among the nodes. We construct a set *M* containing triplets $(\epsilon_i, \epsilon_j, z)$ where $\{+, -, ?\}$ denotes positive, negative and no link, respectively. For each pair of linked nodes (ϵ_i, ϵ_j), we sample a non-linked node ϵ_k. The first term is a weighted multinomial logistic regression (MLG) classifier to classify the relationship $z \in \{+, -, ?\}$ of two nodes. The second term is to guarantee the distance of positive-linked nodes is closer

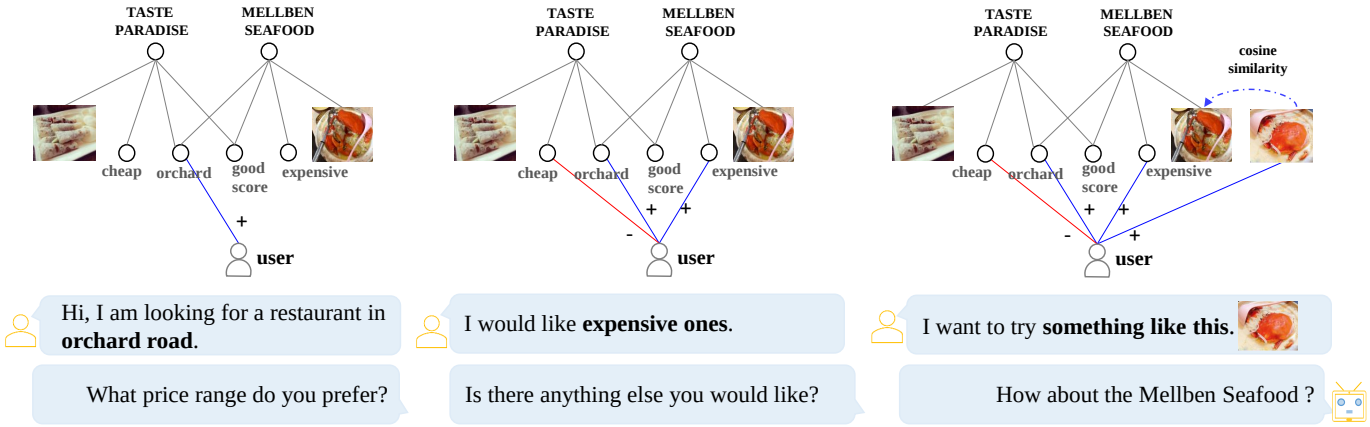


Fig. 3: The update of KG-based state graph[1]

than that of the non-linked nodes, and the distance of the non-linked nodes is closer than that of the negative-linked nodes.

$$L_H = -\frac{1}{M}\sum_{(\epsilon_i,\epsilon_j,z)\in M} w_z log\frac{exp([h_i;h_j]\theta_z^{MLG})}{\sum_{q\in\{+,-,?\}} exp([h_i;h_j]\theta_q^{MLG})} + \lambda[\frac{1}{|M_{(+,?)}|}\sum_{f_{ijk}\in M_{(+,?)}} max(0,(\|h_i-h_j\|_2^2-\|h_i-h_k\|_2^2)) + \frac{1}{|M_{(-,?)}|}\sum_{f_{ijk}\in M_{(-,?)}} max(0,(\|h_i-h_j\|_2^2-\|h_i-h_k\|_2^2))] + Reg(\theta_W,\theta_{MLG}),$$

where w_z denotes the weights of class z. θ_W and θ_{MLG} are the parameters of the signed GCN and MLG. f_{ijk} denotes the nodes (ϵ_i, ϵ_j, ϵ_k) in $M_{(+,?)}$ and $M_{(-,?)}$, which are the set of paired nodes including the linked nodes (ϵ_i, ϵ_j) and non-liked nodes (ϵ_i, ϵ_k). Reg stands for the regularization of the parameters.

After obtaining the node representations, we compute the ranking score of the corresponding slots and values. For the ranking of slots *S*, considering that there are only venues and slot-values pairs in our graph, we compute the scores of slots by aggregating the scores of the slot-value pairs belonging to the same slot and use *softmax* to get the normalized score of all the slots. For the ranking of slot-values and venues, We apply multi-layer perception (MLP) based on the concatenation of the representations of the user and the corresponding nodes in MMKG. We denote the ranking scores of slot, slot-values and venue names as y_S, y_V and y_C, respectively, which are computed as follows:

$$y_S = Softmax(\sum_{v\in S_i} y_v),$$
$$y_V = Sigmoid(MLP([h_u,h_V]),$$
$$y_C = Sigmoid(MLP([h_u,h_C]),$$

where MLP is the multi-layer perception. h_u, h_V and h_C are the representations of the user, slot-values and venues.

Here we apply cross entropy to calculate the loss functions for ranking:

$$L = CrossEntropyLoss(y,y^*), \quad (1)$$

where y denotes the ranking result (e.g. y_S, y_V, y_C) and y^* is the ground truth result accordingly. Finally we obtain the total loss of ranking as : $L_R = \delta_S L_S + \delta_V L_V + \delta_C L_C$. δ_S, δ_V and δ_C denote the balancing coefficients. L_S, L_V and L_C are the loss of the slots, slot-values and venues.

3) Joint Training: For joint training, the total loss of the node presentation learning and graph reasoning is as follows:

$$\mathbb{L} = \alpha L_H + \beta L_R, \quad (2)$$

where L_H and L_R represent the loss of the node representation learning and ranking, respectively. α and β are the coefficients.

IV. Experiments

In this section, we will evaluate our proposed model. For better explain and analyze our model, the following questions are used to guide the analysis of the experiment.

- **RQ1.** Compared with the state-of-the-art conversational recommendation methods, how does our method perform?
- **RQ2.** Is our model robust to different settings, and which design of our model has more significant effects?
- **RQ3.** Whether our model can handle zero-shot scenarios and provide explanations for decision making?

A. Experiments Setup

1) Dataset: There are several multimodal dialogue datasets contributed. However, to the best of our knowledge, most of them are not suitable for our conversational recommendation with MMKG scenario. For example, MMD [28] comes with no dialogue state or dialogue act annotation. MDMMD++ [9] looks promising but the dataset is not publicly available yet. SIMMC [22] comes with state and action annotations but it is not a recommendation setting. Fortunately, Liao et al [18] propose a fully annotated task-oriented Multimodal Multi-domain Conversational dataset (MMConv[2]) which provides realistic conversational recommendation scenarios. The dataset contains large-scale multi-turn dialogues covering five domains: food, hotel, nightlife, shopping mall and sightseeing.

[2]https://github.com/liziliao/MMConv

It also contains a structured venue database and annotated images. During the conversation, both the agent and the user can provide images to each other. The statistics of the dataset is shown in Tab. I. The conversations between the user and the agent are designed based on real user settings. The goal of the agent is to recommend the target venues to the user. The dialogues are fully annotated with dialogue belief states of the user and tuples of the agent such as *inform-price-cheap*.

TABLE I: Statistics of the Dataset

# Dials	# of Turns	# of venues	# of reviews	# of images
5,106	39,759	1,772	39,772	103,773

TABLE II: Statistics of the Dataset Split

Dataset	# of Dials	# of Turns	Avg. # of Turns
Train	3,500	26,869	7.677
Val	606	4,931	8.137
Test	1,000	7,959	7.959

2) Training Details: We split the dataset for training, validation and testing. To facilitate the investigation of the zero-shot situation, we split the dataset by different goals of the dialogues. We ensure that there are no overlapping goals in training, validation and testing datasets. The statistics is shown in Tab. II. The input to the action prediction model is tokenized with pretrained BPE codes [30] associated with DistilGPT2 [29]. We use default hyper parameters for GPT-2 and DistilGPT2 in Huggingface Transformers [35]. Text sequences longer than 1024 tokens are truncated. For reasoning part, the layer number *L* of the signed GCN is set to 2. The dimension of the node features in signed GCN is 128 and the batch size is set to 64. The learning rate is set to 0.001. The initial representations of the entities in the MMKG and the user in state graph are set to random vectors. The maximum number of training epoch is 100. The maximum number of turns of online evaluation is set to 15. To define the image similarity threshold when updating MMKG via images, we apply simple greedy search method to validate the performance based on the candidate thresholds $\{0.5, 0.7, 0.9\}$. All the parameters are tuned on the validation set.

3) Evaluation Metrics: Our goal is to predict the dialogue act of the agent and also provide the detailed content of act to the user such as inform slot value, request slot or recommend a venue. We measure the performance of our model by offline and online evaluation similar to [8].

For offline evaluation, the **Act Accuracy** is the proportion of the correct predicted samples to the total samples. Considering that the result contains more than one actions, we regard it correct when all the predicted actions are strictly equal to the ground truth omitting the order of the actions. **EMR** stands for turn-level entity matching rate, which compares predicted entities (slots, values, venues) against annotated ones when the dialogue act is predicted correctly. Given the testing set with *R* turns, the **EMR** is calculated as follows:

$$EMR@k = \frac{1}{R}\sum_{i=1}^{R} EMR_i,$$

$$EMR_i = \begin{cases} \frac{\left|Z_{gt}^i \cap Z_{pre}^i\right|}{\left|Z_{gt}^i\right|}, & \text{if the predicted actions are correct} \\ 0, & \text{otherwise} \end{cases}$$

where EMR_i is the EMR score of the *i*-th sample. The Z_{gt}^i and Z_{pre}^i is the ground truth entity set and the top-k predicted entity set for all actions, respectively.

IMR stands for dialogue-level item matching rate, which evaluates the predicted venues against the ground-truth across all turns in a dialogue. For each dialogue, we maintain a venue set J_{pre}^i which stores the top-1 predicted venue of the turn whose predicted actions contain "recommendation", the **IMR** is calculated as:

$$IMR = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|J_{gt}^i \cap J_{pre}^i\right|}{\left|J_{gt}^i\right|}$$

where J_{gt}^i and J_{pre}^i is the ground truth item set and the top-1 predicted item set, respectively.

For online evaluation, we use user simulator to evaluate the performance of recommendation like [8]. We simulate the interaction process between the user and the agent. The user will randomly *inform* a slot-value at the first turn, and then the agent provides the response to the user based on the output of our model. After the multi-turn interactions, the dialogue will finish when the agent successfully recommends the target venues or the dialogue reaches to the predefined number of turns. We use **SR@**t to measure the cumulative ratio of conversation completion by turn t of dialogue. SR is mainly used to evaluate whether the user's ground truth items can be found quickly in an interactive scenario.

4) Baselines: Several baselines on conversational recommendation are used for comparison.

Max Entropy. This method designs the rules to perform the actions of the dialogue. When generating questions, it always chooses the attribute that has the maximum entropy among the candidate attributes in each turn. The method makes a recommendation based on the number of candidate item sets with a certain probability.

Abs Greedy [6]. This method only performs the recommendation and it recommends item in each turn until it makes successful recommendation. It updates the model and takes the user rejected items as negative samples. The method achieves equal or better performance than bandit algorithm such as Upper Confidence Bounds [1] and Thompson Sampling [3].

SCPR [15]. It is a graph-based path reasoning method to model the multi-turn conversational recommendation. It starts from the user vertex and then walks through the attribute vertices on the graph based on user feedbacks.

UMGR [37]. This method represents user preference by user memory graph and then applies graph reasoning to model multi-turn conversational recommendation. The dialogue acts are predicted based on the hidden state of the user memory graph.

UNICORN [7]. This method is a unified conversational recommendation policy learning method. The authors leverage a dynamic weighted graph based RL method to capture dynamic user preferences and learn the action selection strategies at each conversation turn. They apply preference-based

TABLE III: Performance Comparison With Baselines

	Offline Evaluation					Online Evaluation		
Methods	Act Accuracy	EMR			IMR	SR		
		EMR@1	EMR@3	EMR@5		SR@5	SR@10	SR@15
Max Entropy	27.76	0.97	12.41	15.08	4.88	8	28.22	44.56
Abs Greedy	21.85	-	-	-	2.66	18.89	29.56	38.33
SCPR	26.56	2.98	20.08	**29.03**	5.28	11.78	37.78	49.44
UMGR	23.96	10.01	11.80	15.91	9.58	24.73	38.04	45.52
UNICORN	24.49	12.25	16.38	19.19	10.08	12	24.11	40.33
SGR	**37.2**	**17.11**	**23.49**	25.28	**11.7**	**38.21**	**50.04**	**54.60**

item selection and weighted entropy-based attribute selection strategies to obtain the detailed actions.

B. Quantitative Results

*1) **Main Results**:* The performance of all methods on the dataset is presented in Tab. III. We can observe that our method outperforms the baselines on most of the metrics. From the results of the offline evaluations, we have the following discoveries: First of all, it is important to see that our model shows higher Act Accuracy compared to other baselines. Our model is designed to be more suitable for natural dialogue scenarios. In each turn, our model is able to generate several actions at the same time. However, the baselines can only consider one action at one turn. Moreover, our model also obtains better EMR and IMR, especially EMR@1 and EMR@3, which is a significant improvement over all the comparison algorithms. This effectively demonstrates the superiority of our method in offline prediction. Specifically, it can be found that our model achieves a higher improvement under EMR@1. This indicates that our algorithm can effectively select the most appropriate slot, slot-value pair or venues while maintaining a high act accuracy. This means that our model is not only closer to the natural form of dialogue, but can also ask the most effective questions for the user.

Compared with our algorithm, other algorithms have lower offline evaluation results. Max Entropy uses simple rule-based protocol to select the actions of the dialogue with probability. Such a policy has high randomness and does not maintain the coherence of the dialogue. SCPR takes advantage of the structural information of the graph to effectively filter out some useless slot-values, so its EMR and IMR are higher than that of Max Entropy and Abs Greedy. However, compared with our model, SCPR has lower EMR@1, EMR@3, IMR That's because our model has better performance on slot and venue ranking. SCPR uses an algorithm similar to max entropy to sort the entities. It cannot well select the most appropriate slots and venues according to the current state of the dialogue. For UMGR, the ranking of entities is only related with the hidden state of entities without considering the user information, which makes it less sensitive to the exact dialogue situation. The EMR@1 and IMR of UNICORN is higher than other baselines. That's because the UNICORN applied dynamic weighted graph to capture the sequential information of the dialogue history which is beneficial to learn user's preference on items. However it still uses weighted entropy to select the candidate slots and values similar to SCPR. It also does not consider the complex relationship among slots and values. Therefore, the performance of UNICORN is worse than that of our model.

From the online evaluation, it can be seen that SCPR has superior results than Max Entropy and Abs Greedy because SCPR uses reinforcement learning to learn a well-designed policy that is responsible for performing the appropriate act to interact with the user. Intuitively, reinforcement learning can make better use of feedback in the process of user interaction. The performance of UNICORN is relatively lower than that of other baselines. The reason is that the UNICORN applies weighted entropy to obtain the requested slots which is more difficult to select the candidate items in less turns. Our model has a significant improvement compared with the baselines. It indicates that our model can identify the ground truth item in a shorter number of dialogue turns, which shows that our design can adapt to more flexible interaction scenarios. We suspect that the multi-action prediction also contributes to the better performance than other baselines.

*2) **Ablation Studies**:* (I) Analysis of signed links. To explore the complex relationship among the slots and values of the venues, we design positive links and negative links in the database MMKG and user state graph. During reasoning, we leverage signed GCN to learn the node features which can better capture the signed relationship among different nodes. To demonstrate the effective of signed links, we perform experiment on a variant model which transfers the signed links into unsigned links. We delete the negative links and construct unsigned graph for database and user state graph. Then we apply a widely used GCN-based method named LightGCN [11] to learn the node representations. To be fair, the rest of our model remains the same. The result is shown in Fig. 4. The SGR_LightGCN represents the variant model using LightGCN to learn the node features. We can observe that although there's no obvious difference under EMR between the two method, our SGR performs much better than SGR_LightGCN under IMR. It indicates that the signed links can better recommend proper venues to the users. That's because the slots with exclusive values increase the number of paths between the user and venues by positive and negative links. Thus with the help of signed links, the relationship between the user and venues are enhanced.

(II) Effectiveness of the ratio of negative links. We conduct experiments under different proportion of negative links of each slots. The proportion ranges from 0 to 1, where 0 means there is no negative links in the MMKG and 1 means that we leverage all the negative links of each slot in the MMKG. As shown on Fig. 5, with the proportion increasing, the performance improved and then degraded with a larger proportion. We suspect that the negative links provide more evidence about users' preferences. Thus the introduction of negative links with proper proportion will help the model do

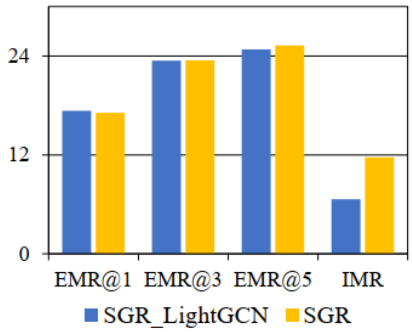


Fig. 4: The effectiveness of signed links.

better reasoning. However, when there are too many negative links, the MMKG becomes larger and more complicated. The node will aggregate more information from the neighbors which may bring noise into the learning process.

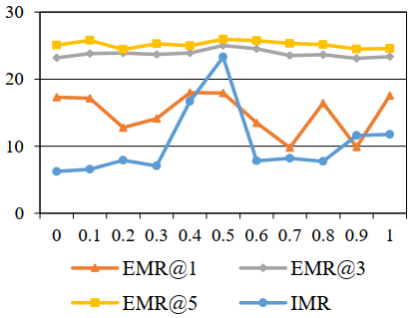


Fig. 5: The effectiveness of the ratio of negative links.

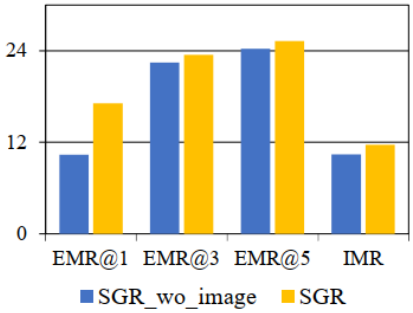


Fig. 6: The effectiveness of images.

(III) Effectiveness of images. Compared with traditional textual conversation, users and agent are allowed to communicate with images in our task of multimodal conversational recommendation. As the saying goes, "A picture is worth a thousand words". The images convey more information about users' preferences which help the agent recommend proper venues to them. To demonstrate the effectiveness of images, we perform experiments with a variant mode named SGR_wo_image. It denotes the variant model by omitting the image information in dialogues and only considering the textual conversation.

The performance on EMR and IMR is shown in Fig. 6. We can observe that the performance is decreased when we only consider the textual conversation. It indicates that the images can help better represent users' preference in some cases. This is because when user provides an image, we can search for similar images in the database and add a positive edge between user and the similar image. Note that the image is also linked with the venue it belongs to in the graph. In this way, we link the user and candidate venue in the state graph. By leveraging signed GCN, we integrate the information of the venue into the user node, which is beneficial for the model to better rank the entities as well as the venues.

*3) **The performance of different domains**:* We conduct experiment about the performance of different domains in the dataset used in our paper. The result is shown in Tab. IV.

TABLE IV: The performance of different domains.

domains	#samples	EMR@1	EMR@3	EMR@5	IMR
food	56.7%	18.42	22.87	24.39	7.73
hotel	11.1%	15.29	18.66	21.85	16.28
nightlife	10.2%	17.97	21.74	23.89	12.04
shopping mall	7.1%	21.94	27.31	28.60	36.15
sightseeing	14.9%	14.06	19.78	22.01	12.92

We can observe that the number of venues and the distinctiveness of attributes affect the performance. For example, the IMR score for the food domain is smaller than that of others, which is due to its largest number of venues. On the contrary, the EMR and IMR for the shopping mall domain is better than those of others. We suppose this is because the number of shopping malls is relatively small and the slot values for it are rather distinctive. It is easier for the model to learn the preferences of users and recommend the proper shopping mall. For example, the locations of shopping malls are relatively scattered, and several popular shopping malls are quite distinctive in certain aspects.

C. Qualitative Results

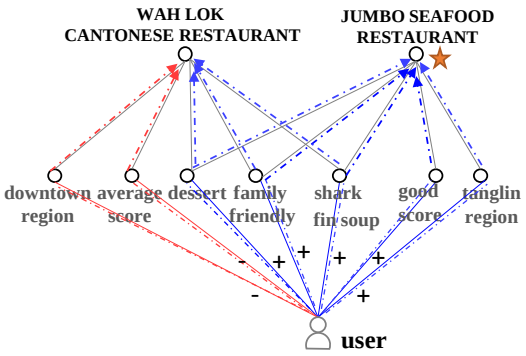


Fig. 7: The case study of the zero-shot situation.

*1) **Zero-shot Case Study**:* Our model has the ability to handle zero-shot items. That's because we can learn the inherent features of venues based on the MMKG of the database via signed GCN. When a user intends to look for a new venue, we can match user's preference by reasoning on the updated graph according to the dialogue state. The IMR for old venues and new venues are about 21.25 and 7.45, respectively. Although the performance of new venues is not as good as that of old venues, the graph in our model enables the prediction of new venues, while other non-graph based methods cannot handle new venues. We show a case study for the zero-shot situation in Fig. 7. In the example, the target venue *JUMBO SEAFOOD RESTAURANT* is a new

venue never seen by the model during the training procedure. The venue is linked with many attributes in MMKG. Here we just show part of the attributes related to user's preferences. Based on the dialogue states up to the current turn, the user is positively linked with (*dessert, family friendly, shark fin soup, good score, tanglin region*) and negatively linked with *downtown region, average score*. According to the information aggregation and propagation in signed GCN, the representation of the venue *JUMBO SEAFOOD RESTAURANT* integrates the information of the attributes and the user. The venue *WAH LOK CANTONESE RESTAURANT* integrates the negative information from the user. Therefore, the target venue *JUMBO SEAFOOD RESTAURANT* is more likely to be recommended to the user.

2) ***Explainability via State Graph****:* To show the explainability, we give a simple example in Fig. 8. The user wants to find a place with *eggs* to have breakfast. We show the state graph in dashed lines. In the first step, the user is positively linked with the attribute *breakfast* and *eggs* to illustrate the user's preferences. The venues with these corresponding attributes are activated (we just show parts of the candidates here) and receive the preference information from the user in the MMKG. After reasoning over the state graph via signed graph convolution, we obtain the scores of the candidate venues which are shown on the arrow lines linking the user and the venues. The top three venues are *FORTY HANDS, SRI KAMALA VILAS* and *SWEE CHOON TIM SUM* with the scores 0.42, 0.41 and 0.40, respectively. As the conversation proceeds, the agent learns more about the user's preferences. The user further gives his/her preference on region *little India* in the following step. Then the representations of the venues in *little India* are enhanced. During model reasoning, the user's new preferences are integrated and propagated in the updated MMKG and the scores of the candidate venues are changed. We can observe that the venue *SRI KAMALA VILAS* better matches the representation of the user with a higher score 0.48. Therefore, the agent is more likely to recommend the *SRI KAMALA VILAS* to the user.

V. Conclusion

In this work, we explored a new type of internal state representation for multimodal dialogues and proposed a signed-graph based reasoning model over it to guide the decision making process. Specifically, inspired from studies on multimodal knowledge graph, we construct multimodal state graph and gradually update it to keep track of dynamic user preferences. Based on the state graph, an end-to-end graph convolutional neural network integrates preference tendencies from the graph and reasons about the next action to perform. As the prediction results inherit the graph structure, it caters for cold start venues and naturally provides explanations for predictions. We conducted experiments on a public multmodal conversational recommendation dataset. Both quantitative and qualitative results demonstrate the effectiveness of our proposed method while also show its ability in handling zero-shot cases and enhancing explainability.

In future work, there are several problems can be further explored. For example, how to make better integration of the images in multimodal conversational recommendation. Besides, when user provides less preference at the first several turns, how to conduct proper dialogue policy is also a challenge to be considered.

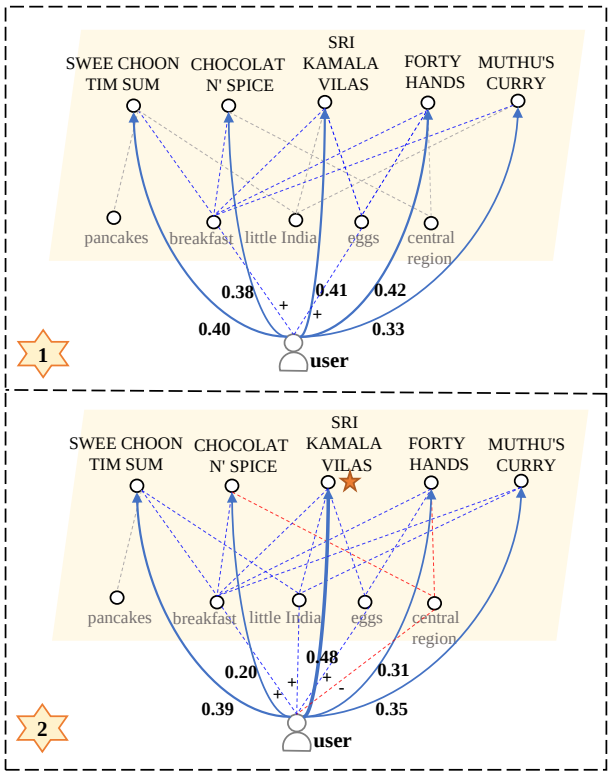


Fig. 8: The sample for explainability of our model.

References

[1] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[2] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 2013, pp. 2787–2795.

[3] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2011, pp. 2249–2257.

[4] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang, "Towards knowledge-based recommender dialog system," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1803–1813.

[5] K. Christakopoulou, A. Beutel, R. Li, S. Jain, and E. H. Chi, "Q&r: A two-stage approach toward interactive recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 139–148.

[6] K. Christakopoulou, F. Radlinski, and K. Hofmann, "Towards conversational recommender systems," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 815–824.

[7] Y. Deng, Y. Li, F. Sun, B. Ding, and W. Lam, "Unified conversational recommendation policy learning via graph-based

reinforcement learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1431–1441.

[8] T. Derr, Y. Ma, and J. Tang, "Signed graph convolutional networks," in *2018 IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 929–934.

[9] M. Firdaus, N. Thakur, and A. Ekbal, "Aspect-aware response generation for multimodal dialogue system," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 2, pp. 1–33, 2021.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639–648.

[12] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 20 179–20 191.

[13] W. Lei, X. He, M. de Rijke, and T.-S. Chua, "Conversational recommendation: Formulation, methods, and evaluation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2425–2428.

[14] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua, "Estimation-action-reflection: Towards deep interaction between conversational and recommender systems," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 304–312.

[15] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T.-S. Chua, "Interactive path reasoning on graph for conversational recommendation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2073–2083.

[16] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal, "Towards deep conversational recommendations," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 9748–9758.

[17] L. Liao, L. Kennedy, L. Wilcox, and T.-S. Chua, "Crowd knowledge enhanced multimodal conversational assistant in travel domain," in *International Conference on Multimedia Modeling*, 2020, pp. 405–418.

[18] L. Liao, L. H. Long, Z. Zhang, M. Huang, and T.-S. Chua, "Mmconv: An environment for multimodal conversational search across multiple domains," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 675–684.

[19] L. Liao, Y. Ma, X. He, R. Hong, and T.-s. Chua, "Knowledge-aware multimodal dialogue systems," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 801–809.

[20] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum, "Mmkg: multi-modal knowledge graphs," in *European Semantic Web Conference*, 2019, pp. 459–474.

[21] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, and T. Liu, "Towards conversational recommendation over multi-type dialogs," in *Proceedings of the 58th International Conference on Computational Linguistics*, 2020, pp. 1036–1049.

[22] S. Moon, S. Kottur, P. A. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho *et al.*, "Situated and interactive multimodal conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1103–1121.

[23] S. Moon, P. Shah, A. Kumar, and R. Subba, "Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 845–854.

[24] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 225–234.

[25] L. Nie, W. Wang, R. Hong, M. Wang, and Q. Tian, "Multimodal dialog system: Generating responses via adaptive decoders," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1098–1106.

[26] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3208–3218.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 1–24, 2019.

[28] A. Saha, M. Khapra, and K. Sankaranarayanan, "Towards building large scale multimodal domain-aware conversation systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 696–704.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," in *the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019, pp. 1–5.

[30] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[31] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1405–1414.

[32] Y. Sun and Y. Zhang, "Conversational recommender system," in *The 41st international acm sigir conference on research & development in information retrieval*, 2018, pp. 235–244.

[33] M. Wang, H. Wang, G. Qi, and Q. Zheng, "Richpedia: a large-scale, comprehensive multi-modal knowledge graph," *Big Data Research*, vol. 22, p. 100159, 2020.

[34] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 540–547.

[35] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[36] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3140–3146.

[37] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, and S. Y. Philip, "User memory reasoning for conversational recommendation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5288–5308.

[38] Y. Zhang, Z. Ou, and Z. Yu, "Task-oriented dialog systems that consider multiple appropriate responses under the same context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 9604–9611.

[39] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, "Towards conversational search and recommendation: System ask, user respond," in *Proceedings of the 27th acm international conference on information and knowledge management*, 2018, pp. 177–186.

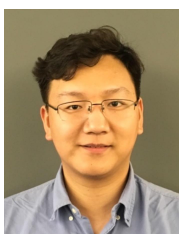
[40] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu, "Improving conversational recommender systems via knowledge graph based semantic fusion," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1006–1014.
[41] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen, "Towards topic-guided conversational recommender system," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4128–4139.
[42] J. Zou, Y. Chen, and E. Kanoulas, "Towards question-based recommender systems," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 881–890.

Yuxia Wu received the B.S. degree from Zhengzhou University, Henan, China, in 2014, the M.S. degree from the Fourth Military Medical University, Xi'an, China, in 2017, and is currently working toward the Ph.D. degree at Xi'an Jiaotong University, Xi'an, China. She is now a visiting student at the National University of Singapore. Her research interests include social multimedia mining, recommender systems and natural language processing.

Lizi Liao is an assistant professor with Singapore Management University. She received the Ph.D. degree in 2019 from NUS Graduate School for Integrative Sciences and Engineering at the National University of Singapore. Her research interests include conversational system, multimedia analysis and recommendation. Her works have appeared in top-tier conferences such as MM, WWW, ICDE, ACL, IJCAI and AAAI, and top-tier journals such as TKDE. She received the Best Paper Award Honorable Mention of ACM MM 2018. Moreover, she has served as the PC member for international conferences including SIGIR, WSDM, ACL, and the invited reviewer for journals including TKDE, TMM and KBS.

Gangyi Zhang received the bachelor's degree at University of Electronic Science and Technology of China, Chengdu, China, in 2020. He is currently pursuing the master's degree with the School of Data Science, University of Science and Technology of China, Hefei, China. His main research interests include conversational recommendation systems, machine learning and data mining techniques.

Wenqiang Lei is a Postdoc Research Fellow in School of Computing, National University of Singapore. He received his Ph.D. degree from National University of Singapore in 2019. His research interests focus on conversational AI, inclusive of conversational recommendation, dialogue and QA system, user feedback modeling. He has published relevant papers at top venues such as KDD, WSDM, TOIS, ACL, EMNLP and the winner of ACM MM 2020 best paper award. He has also actively give tutorials on the topic of conversational recommendation at multiple conferences: RecSys 2021, SIGIR 2020, CCL 2020, CCIR 2020.



Guoshuai Zhao received the BE degree from Heilongjiang University, Harbin, China, in 2012, and the MS and PhD degrees from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2019, respectively. He was an intern with the Social Computing Group, Microsoft Research Asia from January 2017 to July 2017, and was a visiting scholar with Northeastern University, from October 2017 to October 2018 and with MIT, from June 2019 to December 2019. Now, he is an associate professor with Xi'an Jiaotong University. His research interests include social media big data analysis, recommender systems, and natural language generation.

Xueming Qian (M'10) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. He was previously an Assistant Professor at Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi'an Jiaotong University. His rsearch interests include social media big data mining and search.

Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School from 1998-2000. Dr Chua's main research interest is in multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval and question-answering (QA) of text and rich media arising from the Web and multiple social networks. He is the co-Director of NExT, a joint Center between NUS and Tsinghua University to develop technologies for live social media search.

Dr Chua is the 2015 winner of the prestigious ACM SIGMM award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications. He is the Chair of steering committee of ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. Dr Chua is also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves in the editorial boards of four international journals. Dr. Chua is the co-Founder of two technology startup companies in Singapore. He holds a PhD from the University of Leeds, UK.