Beyond Whole Dialogue Modeling: Contextual Disentanglement for Conversational Recommendation

Guojia An
University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
anguojia.1999@gmail.com

Jie Zou*
University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
jie.zou@uestc.edu.cn

Jiwei Wei
University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
mathematic6@gmail.com

Chaoning Zhang
University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
chaoningzhang1990@gmail.com

Fuming Sun
Dalian Minzu University
Dalian, Liaoning, China
sunfuming@dlnu.edu.cn

Yang Yang
University of Electronic Science and
Technology of China
Chengdu, Sichuan, China
yang.yang@uestc.edu.cn

Abstract

Conversational recommender systems aim to provide personalized recommendations by analyzing and utilizing contextual information related to dialogue. However, existing methods typically model the dialogue context as a whole, neglecting the inherent complexity and entanglement within the dialogue. Specifically, a dialogue comprises both focus information and background information, which mutually influence each other. Current methods tend to model these two types of information mixedly, leading to misinterpretation of users' actual needs, thereby lowering the accuracy of recommendations. To address this issue, this paper proposes a novel model to introduce contextual disentanglement for improving conversational recommender systems, named DisenCRS. The proposed model DisenCRS employs a dual disentanglement framework, including self-supervised contrastive disentanglement and counterfactual inference disentanglement, to effectively distinguish focus information and background information from the dialogue context under unsupervised conditions. Moreover, we design an adaptive prompt learning module to automatically select the most suitable prompt based on the specific dialogue context, fully leveraging the power of large language models. Experimental results on two widely used public datasets demonstrate that DisenCRS significantly outperforms existing conversational recommendation models, achieving superior performance on both item recommendation and response generation tasks.

CCS Concepts

• **Information systems** → **Recommender systems**.

*Corresponding author.

Keywords

Conversational Recommendation, Prompt Learning, Disentanglement Learning

1 INTRODUCTION

Traditional recommender systems provide personalized recommendations by mining user behavior. Although they have achieved some success, they still face the limitation in capturing users' real-time intentions [1, 36, 50]. To address this problem, conversational recommender systems (CRS) have emerged, which can dynamically understand user needs through natural language interactions to provide personalized and effective recommendations. With the rapid advancements of CRS in fields such as e-commerce, movie recommendation, and travel planning, it has become a prominent research direction in both academia and industry.

In early studies, researchers find that users usually express their preferences through specific entities, prompting some studies to focus on modeling entity-related information [2, 19, 46, 52]. Although this method may be effective in certain cases, focusing only on specific entities while ignoring the context of the conversation may lead to the loss of information, thereby misunderstanding the user's true intent. Fortunately, with the advancement of natural language processing technologies, the model's ability to understand the semantics of the conversation context has become increasingly powerful. Therefore, recent studies have shifted focus toward modeling the conversation context as a whole [4, 34, 39, 47]. For example, C^2-CRS [48] employs transformers to model the entire conversation context. Then, UniCRS [34] and DCRS [4] leverage RoBERTa [21] to capture the semantic information throughout the dialogue.

Motivation: Although existing methods of modeling the conversation context as a whole have shown promising results in practice, they often neglect the complexity and entanglement of information within dialogue contexts. Specifically, as shown in Figure 1(a),
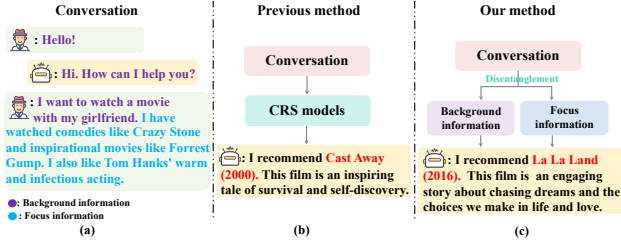
Figure 1: Previous CRS method vs. our method.

the information within a conversation contains focus information (contextual information related to entities) and background information (contextual information unrelated to entities). Previous methods treat the entire dialogue context uniformly and model these two types of information mixedly, which prevents the model from effectively distinguishing the importance of different pieces of information, thereby hindering its ability to accurately capture the user's actual intent. For example, in the case illustrated in Figure 1(b), the previous methods primarily concentrate on the focus information while ignoring the background information that the user wants to watch a movie with his/her girlfriend, leading to wrongly recommending the movie 'Cast Away'. In contrast, our approach first separates these two types of information and then dynamically balances them based on the dialogue context. As illustrated in Figure 1(c), our approach recommends 'La La Land', which is the result of dynamically integrating focus information and background information (i.e., go watch a movie with his/her girlfriend). The advantage of this disentangled modeling approach lies in its ability to effectively distinguish the importance of different types of information, reducing interference from irrelevant details [14, 20]. Also, it flexibly utilizes background information as supplementary input when necessary, thereby enhancing the accuracy of identifying the user's primary intent. Furthermore, from the experimental results, as shown in Figure 2, we compare the performance differences between the dialogue disentanglement approach and the existing holistic modeling method across three state-of-the-art conversational recommendation models. The results show that the disentanglement approach can consistently improve performance, further confirming the necessity and effectiveness of disentanglement for conversational context.

Challenges: Although it is essential to separate the focus information and background information from the conversation context, there are two major challenges to utilize these two types of information: (1) In the absence of supervisory signals, how to effectively disentangle focus information and background information from the conversation context has not been fully explored. (2) In different conversation scenarios, how to utilize dynamic adjustment mechanisms to balance the focus information and background information to achieve accurate identification of the user's true intentions remains challenging.

Our Work: To tackle the aforementioned challenges, we propose a novel model of Contextual **Disen**tanglement for **C**onversational **R**ecommender **S**ystem (**DisenCRS**). Specifically, we design a *dual contextual disentanglement module*, which effectively decomposes complex dialogue context information into focus information and
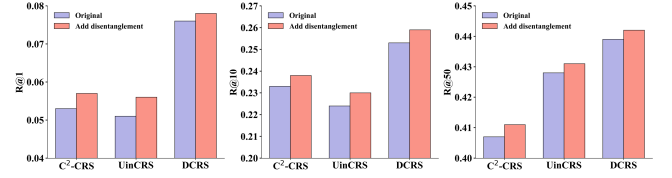

Figure 2: Performance comparison of the original method and the method with added disentanglement.

background information under unsupervised conditions[1]. This module comprises two components: (1) contrastive disentanglement, which establishes a proxy mechanism to guide the disentanglement of the context, and (2) counterfactual inference disentanglement, which aims to capture the impact of the remaining factor by eliminating the influence of either the background information or the focus information factor. Further, we design an *adaptive prompt learning module* that can accurately identify and utilize key information in different conversation scenarios through a dynamic adjustment mechanism to address the second challenge mentioned above. In particular, we treat focus and background information as two types of prompts and construct a prompt pool through weighted fusion. A learnable prompt selector is then introduced to dynamically select the most appropriate prompts from the prompt pool according to the specific dialogue context. In summary, the main contributions of this paper are as follows:

- We emphasize and analyze the importance of disentangling focus and background information from complex conversation contexts. To the best of our knowledge, this paper is the first effort to introduce contextual disentanglement into conversational recommendation, providing a new perspective for subsequent research on conversational recommendation.
- We propose a new CRS model, called DisenCRS, which is equipped with contextual disentanglement and adaptive prompt learning, to effectively take advantage of focus information and background information for improving conversational recommendation.
- Extensive experiment results on two benchmark datasets demonstrate that the proposed model outperforms state-of-the-art methods, achieving statistically significant performance improvements in both recommendation and response generation tasks.

2 RELATED WORK

2.1 Conversational Recommender Systems

Existing CRS research can be categorized into two major directions: attribute-based CRSs [3, 8, 25, 31, 38] and generation-based CRSs [34, 35, 39, 46, 49]. The former focuses on accurately capturing user preferences in as few interaction rounds as possible by asking the user clarifying questions [23, 51] about item attributes. In contrast, the latter not only emphasizes the accuracy of recommendation results but also highlights the importance of the fluency and quality of the generated natural language responses.

---

[1]We conducted an experiment using PCA to visualize the embeddings of focus/background information and target items. The results show that the focus information embeddings are closer to the target items than those of the background information, validating the effectiveness of disentangling focus and background information.
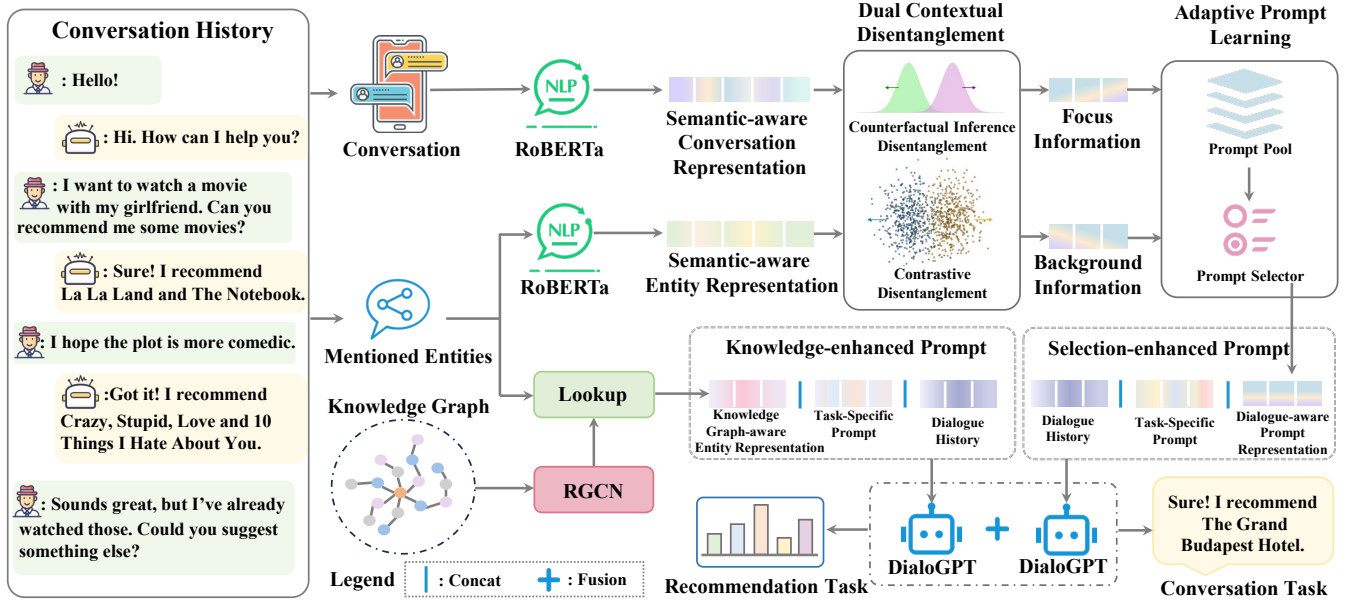
Figure 3: The overall architecture of DisenCRS, which consists of four main components: (1) the representation learning module, which learns the representations essential for our task; (2) the dual contextual disentanglement module, which separates focus information and background information from the conversation context; (3) the adaptive prompt learning module, which dynamically adjusts prompts based on the input dialogue context; (4) the downstream task adaption module, which utilizes the outputs from the previous modules to perform specific tasks.

Attribute-based conversational recommender methods rely on predefined templates to interact with users (i.e., asking users about their preferences for item attributes or making recommendations) [32, 33, 37]. Such CRSs aim to achieve high accuracy of recommendations while with a focus on designing effective strategies to accurately understand user intentions within the minimum number of interaction rounds. They usually adopt reinforcement learning [5, 11, 12, 28] or bandit-based solutions [16] in such systems to perform decision-making for whether to ask a question or recommend items. By leveraging such decision-making methods, CRSs can continuously adjust recommendation strategies based on user feedback to maximize long-term returns.

Generation-based CRSs have attracted widespread attention due to their ability to generate more natural and fluent dialogue content. In early research of generation-based CRSs, many scholars recognized that entity information often effectively represents user intentions and plays a crucial role in improving the accuracy of recommendations. For instance, Chen et al. [2] firstly introduce knowledge graphs to CRSs to model entity information in dialogues. Building on this, KGSF [46] incorporate both word-oriented and entity-oriented knowledge graphs to enhance data representations in CRSs. Subsequently, Zou et al. [53] further models the sequential relationships between entities to better utilize entity information. Kim et al. [9] attempt to enrich the semantic representations of items by leveraging user reviews of the items. With the remarkable advancements in natural language understanding technologies, leveraging complete dialogue context information has become a popular trend for capturing users' complex interest preferences. Specifically, C^2-CRS [48] employs a Transformer architecture to model dialogue context. They integrate entity information and

review data to generate the final user representation through a coarse-to-fine contrastive learning approach. Then, Wang et al. [34] utilize the pre-trained language model RoBERTa to represent dialogue context, emphasizing the modeling of consistency between dialogue tasks and recommendation tasks, thereby enhancing the task synergy performance in CRS. On this basis, DCRS [4] also utilizes RoBERTa to model dialogue context and further develops a demonstration-augmented prompt learning approach for conversational recommendation.

Our work belongs to the field of generation-based CRSs, but has significant differences compared to existing approaches. Although we also utilize conversational context information, we place special emphasis on the complexity and entanglement of context information. Instead of directly using the entire conversation context, we propose a novel method that explicitly distinguishes between focus and background information in the dialogue context and adaptively balances the use of these two types of information, thereby enhancing both recommendation accuracy and response quality.

2.2 Disentanglement for Recommendation

Disentanglement in recommendation refers to effectively separating different features and factors from the latent embedding space, enabling a clearer understanding and representation of user interests [18, 24]. Recent disentanglement methods aimed to differentiate various reasons in the context of recommendation, such as user interest and consistency [45], long-term and short-term interests [44], user intent [27, 43], ID and modality [41]. Unlike the aforementioned disentangled recommendation methods, this study attempts to separate the entangled information inherent in dialogue context within the conversational recommendation scenario.

3 METHODOLOGY

In this section, we first define the task and provide a detailed introduction to our proposed DisenCRS model, the architecture of which is shown in Figure 3.

3.1 Problem Formalization

Given a dialogue C in CRS, it is composed of a sequence of utterances, which can be expressed as $C = \{\mathcal{U}_i\}_{i=1}^{n_u}$ (n_u represents the number of utterances in the dialogue). And, $\mathcal{U}_i = \{\mathcal{W}_j\}_{j=1}^{n_w}$ (n_w represents the number of words in the utterance) represents a sequence of words from the vocabulary $\mathcal{V}$ in the j-th utterance.

Formally, the tasks in CRS can be expressed as follows: For the recommendation task, given a dialogue C, the goal is to select an appropriate subset of items $\mathcal{I}_{\text{rec}}$ that align with the user's intent from the candidate item set $\mathcal{I}$ ($\mathcal{I}_{\text{rec}} \subseteq \mathcal{I}$), and then recommend them to the user. For the response generation task, the goal is to generate an accurate response $\mathcal{R} = \{\mathcal{U}_i\}$ based on the dialogue C.

3.2 Representation Learning Module

In this section, we present our representation learning module, which encodes three types of representations: knowledge graph-aware entity representation, semantic-aware entity representation, and semantic-aware conversation representation.

3.2.1 Knowledge Graph-aware Entity Representation. To accurately model the representation of entities in the latent space, we follow existing research [34, 46] and adopt a knowledge graph-based approach. Specifically, we construct a knowledge graph $\mathcal{G}$ using an entity set $\mathcal{N}$ and a relation set $\mathcal{R}$, where semantic facts are represented in the form of triples $\langle e_1, r, e_2 \rangle$, with $e_1, e_2 \in \mathcal{E}$, and $r \in R$. Subsequently, we employ a relational graph convolutional network (RGCN) [29] to encode the knowledge graph $\mathcal{G}$. Formally, this process can be expressed as:

$$\mathbf{h}_e^{(l+1)} = \sigma \left(\sum_{r \in \mathcal{R}} \sum_{e' \in \mathcal{E}_e^r} \frac{1}{Z_{e,r}} \mathbf{W}_r^{(l)} \mathbf{h}_{e'}^{(l)} + \mathbf{W}_e^{(l)} \mathbf{h}_e^{(l)} \right), \quad (1)$$

where $\mathbf{h}_e^{(l+1)} \in \mathbb{R}^{d_e}$ represents the embedding of entity e in the latent space at layer $l+1$. d_e defines the dimensions of entity representations. σ is the ReLU activation function, which enhances the non-linear expressiveness of the feature representation. $\mathcal{E}_e^r$ denotes the set of all neighboring nodes connected to node e through relation r. $\mathbf{W}_r^{(l)}$ and $\mathbf{W}_e^{(l)}$ are learnable weight matrices. $Z_{e,r}$ is the normalization factor, representing the number of neighbors.

As the dialogue context may involve multiple entities, we plan to use a self-attention mechanism to combine all entity representations, thereby obtaining a unified knowledge graph-aware conversation representation $\mathbf{h}_{\mathcal{G}}$. The formula is shown as follows:

$$\boldsymbol{\alpha}_1 = \text{softmax} \left(\mathbf{b}_1^{\text{T}} \tanh \left(\mathbf{W}_a \mathbf{H}_e^g \right) \right), \quad (2)$$

$$\mathbf{h}_{\mathcal{G}} = \mathbf{H}_e^g \boldsymbol{\alpha}_1^{\text{T}}, \quad (3)$$

where $\boldsymbol{\alpha}_1$ represents the weights of all mentioned entities in the context, $\mathbf{H}_e^g \in \mathbb{R}^{d_e \times n_e}$ is the entity embedding matrix, n_e is the number of entities mentioned in the context C, $\mathbf{W}_a \in \mathbb{R}^{d_e \times d_e}$, $\mathbf{b}_1 \in \mathbb{R}^{d_e}$ are learnable transformation matrix.

3.2.2 Semantic-aware Entity Representation. In addition to modeling entities through the knowledge graph, we also use RoBERTa to enhance semantic-aware embeddings of entities. Specifically, we extract an entity set from the dialogue context and encode them with RoBERTa to generate semantic-aware entity representation $\mathbf{H}_e^s \in \mathbb{R}^{d_e \times n_e}$.

$$\mathbf{H}_e^s = \text{RoBERTa}(\mathcal{N}_{C_k}), \quad (4)$$

where RoBERTa [21] is a pre-trained language model, $\mathcal{N}_{C_k}$ denotes the set of entities in conversation C_k. Subsequently, similar to eqs. (2) and (3), a comprehensive semantic representation of entities $\mathbf{h}_p$ is obtained.

3.2.3 Semantic-aware Conversation Representation. In addition to entity mentions, a conversation C_k also encompasses other rich contextual features that are essential for accurately capturing and representing user interests. Therefore, we also utilize RoBERTa to encode the dialogue context, generating each token representation $\mathbf{h}_c$. The formulation is as follows:

$$\mathbf{h}_c = \text{RoBERTa}(C_k). \quad (5)$$

Additionally, to obtain the global semantic representation of the entire dialogue context, we use the embedding vector of the [CLS] token in RoBERTa. This allows RoBERTa to effectively model the complex semantic relationships within the context and produce semantic-aware conversation representation $\mathbf{h}_{cls}$.

3.3 Dual Contextual Disentanglement Module

In this section, we propose the dual contextual disentanglement module, which includes two self-supervised approaches, i.e., contrastive disentanglement and counterfactual inference disentanglement, to disentangle focus information and background information within the dialogue context.

3.3.1 Contrastive Disentanglement. We propose a contrastive disentanglement strategy to ensure that the embedding vectors of focus information and background information are significantly distinguishable in the representation space. Since explicit labels for focus and background information are not provided in the dialogue context, this presents a challenge for disentangling them. Therefore, we extract entity-related information from the dialogue context as proxy signals to guide the disentanglement of focus and background information. Specifically, we design a contrastive learning task where the similarity between focus information and the proxy information is significantly higher than the similarity between background information and the proxy information. similarly, the similarity between focus information and the proxy information should also be significantly higher than the similarity between focus information and background information. The contrastive process is formulated as follows:

$$\text{sim} \left(\mathbf{h}_f, \mathbf{h}_p \right) > \text{sim} \left(\mathbf{h}_f, \mathbf{h}_b \right), \quad (6)$$

$$\text{sim} \left(\mathbf{h}_f, \mathbf{h}_p \right) > \text{sim} \left(\mathbf{h}_b, \mathbf{h}_p \right), \quad (7)$$

where $\mathbf{h}_f$ and $\mathbf{h}_b$ represent the focus information and background information, which are disentangled from the dialogue semantic representation $\mathbf{h}_{cls}$. We use semantic-aware entity representation $\mathbf{h}_p$ as the proxy information and use cosine similarity $\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$

as the similarity function. Finally, we implement loss functions based on triplet loss to accomplish the contrastive process in eqs. (6) and (7). Formally, the loss functions based on Euclidean distance are computed as follows:

$$
\begin{aligned}
\mathcal{L}_{cd} = \ &\max\left(0, \text{sim}\left(\mathbf{h}_f, \mathbf{h}_b\right) - \text{sim}\left(\mathbf{h}_f, \mathbf{h}_p\right) + m\right) + \\
&\max\left(0, \text{sim}\left(\mathbf{h}_b, \mathbf{h}_p\right) - \text{sim}\left(\mathbf{h}_f, \mathbf{h}_p\right) + m\right),
\end{aligned}
\tag{8}
$$

where m represents a margin value, used to control the minimum gap between similarities.

3.3.2 Counterfactual Inference Disentanglement.

In addition to contrastive disentanglement, we introduce a counterfactual inference strategy to enhance the disentanglement of focus and background information. Counterfactual inference is a process of exploring the consequences of alternative actions in hypothetical situations, which are usually contrary to the actual observed results under specific conditions [41].

In this paper, the idea behind counterfactual inference is that, when either background or focus information is missing, the recommendation system's decision will rely on the remaining information source. This provides an opportunity to precisely identify and disentangle focus and background information. Specifically, for a dialogue C_k, we first determine whether the user's choice in the dialogue is primarily influenced by focus information or background information. To achieve this, we compute the similarity between the focal information and background information with the ground truth item respectively, thus generating pseudo labels for each conversation leader. Then Based on the idea of counterfactual inference, the similarity between the dominant factor and the target item is expected to be significantly greater than the similarity between the other factor and the target item. By identifying the dominant factor influencing the user's choice, this process facilitates the disentanglement of focus and background information.

$$
\mathcal{L}_{ci} = \begin{cases}
-\frac{\text{sim}(\mathbf{h}_f, \mathbf{h}_t)}{\text{sim}(\mathbf{h}_f, \mathbf{h}_t) + \text{sim}(\mathbf{h}_b, \mathbf{h}_t)}, & \mathbf{h}_t \in D_f \\
-\frac{\text{sim}(\mathbf{h}_b, \mathbf{h}_t)}{\text{sim}(\mathbf{h}_b, \mathbf{h}_t) + \text{sim}(\mathbf{h}_f, \mathbf{h}_t)}, & \mathbf{h}_t \notin D_f
\end{cases}
\tag{9}
$$

where $\mathbf{h}_t$ is the ground truth item. D_f denotes the set of $\mathbf{h}_t$ dominated by focus information. If $\mathbf{h}_t \in D_f$, then the target item is mainly dominated by the focus information, otherwise it means that the target item is dominated by the background information.

3.4 Adaptive Prompt Learning Module

After successfully disentangling the focus information and background information, we design an adaptive prompt learning module to efficiently utilize these two types of information for accurate recommendations. This module first constructs a prompt pool and then uses a prompt selector to dynamically select the appropriate prompt from the prompt pool based on the current conversation.

3.4.1 Prompt Pool.

To provide a variety of options, we construct the initial prompt pool $\mathbf{T}_{pool}$, which is defined as follows:

$$
\mathbf{T}_{pool} = \left[\mathbf{T}^1; \ldots; \mathbf{T}^i; \ldots; \mathbf{T}^\eta\right],
\tag{10}
$$

$$
\mathbf{T}_f^i = \left[\mathbf{h}_f; \mathbf{h}_{cls}\right],
\tag{11}
$$

$$
\mathbf{T}_b^i = \left[\mathbf{h}_b; \mathbf{h}_{cls}\right],
\tag{12}
$$

where $\mathbf{T}^i$ consists of two components: $\mathbf{T}_f^i$ and $\mathbf{T}_b^i$. Then $\mathbf{T}_f^i$ and $\mathbf{T}_b^i$ respectively represent the fusion of the focus information $\mathbf{h}_f$ and background information $\mathbf{h}_b$ with the dialogue context $\mathbf{h}_{cls}$.

3.4.2 Prompt Selector.

After constructing the prompt pool, we design a prompt selector to choose the most pertinent prompt from the prompt pool, on the basis of the ongoing dialogue context. The overall process is as follows:

$$
\mathbf{T}_c^i = \mathbf{T}_f^i \mathbf{W}_f^i + \mathbf{T}_b^i \mathbf{W}_b^i,
\tag{13}
$$

$$
\mathbf{T}_c = \left[\mathbf{T}_c^1; \ldots; \mathbf{T}_c^i; \ldots; \mathbf{T}_c^\eta\right],
\tag{14}
$$

$$
\mathbf{T}_{select} = \mathbf{T}_c\left[\mathbf{F}_{select}(\mathbf{T}_c)\right],
\tag{15}
$$

where $\mathbf{W}_f^i$ and $\mathbf{W}_b^i$ controls the importance of the focus information $\mathbf{T}_f^i$ and the background information $\mathbf{T}_b^i$. And these weights is predefined and satisfy the condition $\mathbf{W}_b^i + \mathbf{W}_f^i = 1, \mathbf{W}_b^i, \mathbf{W}_f^i \in [0, 1]$. $\mathbf{T}_c$ refers to the final prompt pool obtained after performing the weight-based fusion of the focus information and background information from the initial prompt pool. $\mathbf{F}_{select}$ is a learnable neural network, e.g., a multilayer perceptron or an attention network, to select the index of best prompt based on $\mathbf{T}_c$. In this work, for simplicity, we use a multilayer perceptron. $\mathbf{T}_{select}$ represents the embedding corresponding to the best index in the prompt pool.

3.5 Downstream Tasks

We further design task-specific prompts tailored for downstream recommendation and conversation tasks, which fully leverage the advantages of prompt learning to enhance the model's overall performance.

3.5.1 Item Recommendation Task.

The recommendation task focuses on predicting items that align with user interests. Next, we introduce the item recommendation prompt and parameter optimization for the recommendation task.

Item Recommendation Prompt. For accurately capturing user interests in item recommendations, we construct a recommendation-oriented prompt. Specifically, the recommendation-oriented prompt consists of the selected context representation $\mathbf{T}_{select}$, recommendation-specific soft tokens $\mathbf{T}_{rec}^{init}$, and the original dialogue history C, formally represented as follows:

$$
\mathbf{T}_{rec}^{select} = \left[\mathbf{T}_{select}; \mathbf{T}_{rec}^{init}; C\right],
\tag{16}
$$

$$
\mathbf{T}_{rec}^{kg} = \left[\mathbf{h}_{\mathcal{G}}; \mathbf{T}_{rec}^{init}; C\right].
\tag{17}
$$

Then, $\mathbf{T}_{rec}^{select}$ and $\mathbf{T}_{rec}^{kg}$ is separately fed into a frozen DialoGPT to produce the vector representation $\mathbf{h}_{rec}^{select}$ and $\mathbf{h}_{rec}^{kg}$, respectively. Finally, we adaptively fuse them into the final user representation $\mathbf{h}_{rec}$.

$$
\mathbf{h}_{rec} = \mathbf{W}_s \mathbf{h}_{rec}^{select} + \mathbf{W}_{kg} \mathbf{h}_{rec}^{kg},
\tag{18}
$$

where $\mathbf{W}_s$ and $\mathbf{W}_{kg}$ are learnable weight matrices.

Parameter Optimization for Recommendation. Following the training approach in Dao et al. [4], we first pre-train the item recommendation module to bridge the gap between prompts and contextual information. Formally, the prediction probability of an entity is formulated as follows:

$$
P_{entity} = \text{Softmax}\left(\mathbf{h}_{rec}^{\mathsf{T}} \mathbf{E}_{entity}\right),
\tag{19}
$$

where $\mathbf{E}_{entity} \in \mathbb{R}^{d_e \times N_{entity}}$ is the embedding matrix of all entities. In the pre-training stage, we employ cross-entropy as the loss function to optimize the model parameters:

$$\mathcal{L}_{pre} = - \sum_{i=1}^{N_{(c)}} \sum_{e=1}^{N_{(e)}} \left[y_{i,e} \log(P_{entity}^{i,e}) + \right.$$
$$\left. (1 - y_{i,e}) \log(1 - P_{entity}^{i,e}) \right] + \lambda (\mathcal{L}_{cd} + \mathcal{L}_{ci}), \quad (20)$$

where $N_{(c)}$ represents the total number of conversation instances, i denotes the index of a specific conversation, $N_{(e)}$ is the set of mentioned entities in the input context, and e represents the index of an entity. After completing the pre-training, we also use the cross-entropy loss function to optimize the model parameters. The difference lies in the optimization targets: the pre-training stage focuses on entities, while the training stage focuses on items. The specific optimization formula is as follows:

$$\mathcal{L}_{rec} = - \sum_{i=1}^{N_{(c)}} \sum_{j=1}^{M} \left[y_{i,j} \log(P_{item}^{i,j}) + \right.$$
$$\left. (1 - y_{i,j}) \log(1 - P_{item}^{i,j}) \right], \quad (21)$$

where M denotes the total number of items, and j serves as the index for an individual item. The variable $y_{i,j}$ represents the ground-truth label, indicating that item i is the correct label for the j-th training instance. The probability $P_{item}^{i,j}$ of recommending candidate items is calculated with similar to eq. (19). In addition, during the training process, the parameters of the DialoGPT model are frozen.

3.5.2 Response Generation Task. In conversational recommendation, the purpose of the response generation task is to produce natural, relevant, and informative text responses to enhance user interaction experience. Next, we describe the prompt construction and parameter optimization for the response generation task.

Response Generation Prompt. Similar to constructing prompts for recommendation tasks, we also design task-specific prompts for the response generation task. Specifically, we develop two types of prompts: a selection-enhanced prompt $\mathbf{T}_{gen}^{select}$ and a knowledge-enhanced prompt $\mathbf{T}_{gen}^{kg}$. These are formalized as follows:

$$\mathbf{T}_{gen}^{select} = \left[\mathbf{T}_{select}; \mathbf{T}_{gen}^{init}; C \right], \quad (22)$$

$$\mathbf{T}_{gen}^{kg} = \left[\mathbf{h}_{\mathcal{G}}; \mathbf{T}_{gen}^{init}; C \right], \quad (23)$$

where $\mathbf{T}_{gen}^{init}$ represents the initial conversation-specific soft tokens, C is represented as a sequence of word tokens, which denotes the original conversation history.

Parameter Optimization for Response Generation. After constructing the two prompts, we optimize the parameters Θ_{gen} of the generation model by minimizing the specified objective function:

$$\mathcal{L}_{gen} =$$
$$- \frac{1}{N_{(c)}} \sum_{i=1}^{N_{(c)}} \sum_{j=1}^{L_{gen}} \log P_{gen} \left(w_{i,j} \mid \mathbf{T}_{gen}^{select}, \mathbf{T}_{gen}^{kg}; \Theta_{gen}; w_{<j} \right), \quad (24)$$

Table 1: Statistics of datasets

Statistic	ReDial	INSPIRED
# users	956	1,482
# conversation	10,006	1,001
# utterance	182,150	35,811
# entities	64,364	17,321

where L_{gen} denotes the length of the generated response. $w_{<j}$ represents the sequence of words that appear before the j-th position. $P_{gen}(\cdot)$ indicates the probability of response generation.

4 EXPERIMENTS

To comprehensively demonstrate the effectiveness of our proposed DisenCRS model, we design and conduct extensive experiments to answer the following key questions: (1) **RQ1**: How does the performance of DisenCRS compare to existing baselines? (2) **RQ2**: What is the impact of the proposed dual contextual disentanglement module? (3) **RQ3**: Is our adaptive prompt learning module effective? (4) **RQ4**: How do different hyperparameters affect the performance of DisenCRS?

4.1 Experimental Setup

4.1.1 Datasets. We conduct extensive experiments on two widely used conversational recommendation datasets, i.e., ReDial [15] and INSPIRED [7], similar to previous work [4, 34]. The detailed statistics of these datasets are shown in Table 1. Both ReDial and INSPIRED are conversational recommendation datasets specifically designed for movie recommendations, aiming to simulate the interaction process between users and recommendation systems.

4.1.2 Evaluation Metrics. Following previous work [4, 34], we employ multiple evaluation metrics for the recommendation task, including Recall@k (k = 1, 10, 50), NDCG@k (k = 10, 50), and MRR@k (k = 10, 50). For the conversational task, both automatic and human evaluations are conducted. In the automatic evaluation, we report metrics such as BLEU-N (N = 2, 3), ROUGE-N (N = 2, L), and Distinct-N (N = 2, 3, 4). For the human evaluation, we randomly sample 20 responses and invite two annotators to score them. The scoring ranges from 1 to 3 and primarily focuses on two key aspects: informativeness and fluency.

4.1.3 Baseline Methods. We select the following competitive baselines for performance comparison based on Dao et al. [4]: (1) **Popularity**: It ranks the items based on their historical recommendation frequency. (2) **TextCNN** [10]: It applies convolutional neural networks to extract contextual features for ranking items. (3) **BERT** [6]: It is a pre-trained deep bidirectional transformer model that aims to improve natural language understanding through masked language modeling and next-sentence prediction tasks. (4) **GPT-2** [26]: It serves as a fundamental baseline for text generation, benefiting from large-scale pre-trained language models. (5) **DialoGPT** [42]: It is a large-scale generative pre-trained model optimized for conversational response generation, which leverages extensive dialogue data to produce contextually relevant and coherent responses. (6) **BART** [13]: It aims to improve natural language generation and understanding tasks by reconstructing corrupted text. (7) **ReDial** [15]: It proposes a dialogue module based on HRED [30] and a

Table 2: Overall performance on recommendation task. * indicates a statistically significant difference with a p-value < 0.05 compared to other baseline methods. Compared with the baselines, our proposed model achieves the highest results on all metrics.

Dataset	Metric	Popularity	TextCNN	BERT	ReDial	KBRD	KGSF	TREA	VRICR	UniCRS	DCRS	**DisenCRS**
ReDial	Recall@1	0.011	0.010	0.027	0.010	0.033	0.035	0.045	0.054	0.065	0.076	**0.081**[*]
	Recall@10	0.053	0.066	0.142	0.065	0.150	0.175	0.204	0.244	0.241	0.253	**0.268**[*]
	Recall@50	0.183	0.187	0.307	0.182	0.311	0.367	0.403	0.406	0.423	0.439	**0.451**[*]
	NDCG@10	0.029	0.033	0.075	0.034	0.083	0.094	0.114	0.138	0.143	0.154	**0.162**[*]
	NDCG@50	0.057	0.059	0.112	0.059	0.118	0.137	0.158	0.174	0.183	0.195	**0.210**[*]
	MRR@10	0.021	0.023	0.055	0.024	0.062	0.070	0.087	0.106	0.113	0.123	**0.130**[*]
	MRR@50	0.027	0.028	0.063	0.029	0.070	0.079	0.096	0.114	0.121	0.132	**0.138**[*]
INSPIRED	Recall@1	0.031	0.025	0.049	0.009	0.042	0.051	0.047	0.043	0.085	0.093	**0.094**[*]
	Recall@10	0.155	0.119	0.189	0.048	0.135	0.132	0.146	0.141	0.230	0.226	**0.252**[*]
	Recall@50	0.322	0.245	0.322	0.213	0.236	0.239	0.347	0.336	0.398	0.414	**0.423**[*]
	NDCG@10	0.085	0.066	0.112	0.023	0.088	0.092	0.095	0.091	0.149	0.153	**0.165**[*]
	NDCG@50	0.122	0.094	0.141	0.059	0.109	0.114	0.132	0.134	0.187	0.192	**0.200**[*]
	MRR@10	0.064	0.050	0.088	0.015	0.073	0.079	0.076	0.075	0.125	0.130	**0.139**[*]
	MRR@50	0.071	0.056	0.095	0.023	0.077	0.083	0.087	0.085	0.133	0.137	**0.146**[*]

recommendation module based on an autoencoder. (8) **KBRD** [2]: It brings knowledge graphs to conversational recommendation, connecting recommendation and dialogue systems through knowledge propagation. (9) **KGSF** [46]: It integrates word-oriented and entity-oriented knowledge graphs to enhance data representation in conversational recommendation. (10) **UniCRS** [34]: It pioneers the unification of recommendation and dialogue tasks within the prompt learning paradigm and leverages pre-trained language models to accomplish both subtasks. (11)**TREA** [17]: It introduces a multi-hierarchical tree structure for reasoning in CRSs, enhancing causal understanding of entities and improving response generation with historical context. (12) **VRICR** [40]: It focuses on enhancing incomplete knowledge graphs using large-scale dialogue corpora and perform dynamic knowledge reasoning based on the dialogue context. (13) **DCRS**[4]: It proposes a demonstration-augmented conversational recommendation system that improves the understanding of dialogue context by retrieving and learning demonstrations.

4.1.4 Implementation Details. In our experiments, we fine-tune the parameters of DisenCRS using the AdamW [22], with the learning rate set to 0.0005 for the recommendation task and 0.0001 for the dialogue task. The batch size is set to 24. We use the DialoGPT-small model, pre-trained on 147M Reddit dialogues, as the base language model and freeze all its parameters to ensure stability during training. The embedding dimension and the number of Transformer layers are 768 and 12, respectively. Additionally, for the recommendation subtask, we follow DCRS [4] and use a fixed RoBERTa-base model to encode the input text semantically.

4.2 Overall Performance (RQ1)

In this section, we conduct extensive experiments and report the average metrics to verify the effectiveness of our model in both recommendation and response generation tasks.

4.2.1 Evaluation on Recommendation Task. Table 2 shows the recommendation performance of various models on the ReDial and

INSPIRED datasets in terms of Recall@k, NDCG@k, and MRR@k metrics. The experimental results demonstrate that our model DisenCRS significantly outperforms other models across all metrics. Analysis of this table reveals the following phenomena:

(1) Popularity and TextCNN, as traditional recommendation approaches, exhibit mediocre performance on both datasets, indicating that these methods struggle to effectively capture contextual information in conversational recommendation scenarios. BERT demonstrates superior performance compared to traditional methods, yet it remains inferior to most conversational recommendation models. This suggests that although BERT can understand the semantic information of conversational context, it still falls short in modeling user interests effectively.

(2) In conversational recommendation models, ReDial serves as a classic baseline. However, due to its limitations in entity and contextual modeling, its performance is suboptimal. KBRD is the first to introduce an external knowledge graph for entity modeling, resulting in improved recommendation performance. Subsequently, KGSF further integrates word and entity knowledge graphs based on KBRD, optimizing data representation and thus improving the model's overall performance. Compared with KBRD and KGSF, TREA improved results by better mining entity-related information, e.g., introducing tree-structure reasoning to uncover the causal relationships between entities and retrieving similar conversations through entity information. VRICR further improved recommendation performance by reconstructing missing entity relationships in incomplete knowledge graphs. Notably, UniCRS and DCRS exhibit outstanding performance compared to other baselines. Both of them combine entity representation and semantic information of conversational contexts, and incorporate well-designed prompts into DialoGPT for downstream tasks. In comparison with UniCRS, DCRS achieves relatively better performance, mainly because it employs a dialogue retrieval approach, fully leveraging richer contextual information to enhance the accuracy of recommendations.

(3) Finally, DisenCRS achieves the best performance on both datasets compared to the aforementioned baseline models. This achievement is attributed to its effective extraction of focus and

Table 3: Automatic evaluation results on the response generation task. Our model achieves the highest performance for the response generation task.

Model	ReDial							INSPIRED						
	BLEU		ROUGE		DIST			BLEU		ROUGE		DIST		
	-2	-3	-2	-L	-2	-3	-4	-2	-3	-2	-L	-2	-3	-4
DialoGPT	0.041	0.021	0.054	0.258	0.436	0.632	0.771	0.031	0.014	0.041	0.207	1.954	2.750	3.235
GPT-2	0.031	0.013	0.041	0.244	0.405	0.603	0.757	0.026	0.011	0.034	0.212	2.119	3.084	3.643
BART	0.024	0.011	0.031	0.229	0.432	0.615	0.705	0.018	0.008	0.025	0.208	1.920	2.501	2.670
ReDial	0.004	0.001	0.021	0.187	0.058	0.204	0.442	0.001	0.000	0.004	0.168	0.359	1.043	1.760
KBRD	0.038	0.018	0.047	0.237	0.070	0.288	0.488	0.021	0.007	0.029	0.218	0.416	1.375	2.320
KGSF	0.030	0.012	0.039	0.244	0.061	0.278	0.515	0.023	0.007	0.031	0.228	0.418	1.496	2.790
VRICR	0.021	0.008	0.027	0.137	0.107	0.286	0.471	0.011	0.001	0.025	0.187	0.853	1.801	2.827
TREA	0.022	0.008	0.039	0.175	0.242	0.615	1.176	0.013	0.002	0.027	0.195	0.958	2.565	3.411
UniCRS	0.045	0.021	0.058	0.285	0.433	0.748	1.003	0.022	0.009	0.029	0.212	2.686	4.343	5.520
DCRS	0.048	0.024	0.063	0.285	0.779	1.173	1.386	0.033	0.014	0.045	0.229	3.950	5.729	6.233
DisenCRS	**0.050***	**0.027***	**0.065***	**0.286***	**0.784***	**1.212***	**1.411***	**0.034***	**0.017***	**0.048***	**0.232***	**4.014***	**6.318***	**6.623***

Table 4: Human evaluation results on the response generation task.

Model	Fluency	Informativeness
KBRD	2.32	1.97
KGSF	2.46	2.05
VRICR	2.37	2.17
UniCRS	2.71	2.53
DCRS	2.80	2.65
DisenCRS	**2.85***	**2.72***

Figure 4: Ablation comparison of the dual contextual disentanglement module.

background information in the conversation through contrastive disentanglement and counterfactual inference disentanglement, as well as its dynamic utilization of this information through the adaptive prompt learning module. Unlike existing methods that process the entire conversational context, DisenCRS more accurately utilizes the key information in the context, avoids the interference of minor information, and thus improves the accuracy of recommendations.

4.2.2 Evaluation on Response Generation Task. In this section, to comprehensively and accurately evaluate the effectiveness of the proposed model in the response generation task, we employ two evaluation methods: automatic evaluation and human evaluation.

Automatic Evaluation. In Table 3, DialoGPT, GPT-2, and BART demonstrate superior performance for response generation over most conversational recommendation models like ReDial, VRICR, and TREA, mainly because they are pre-trained on large-scale corpora and are able to generate informative and semantically consistent responses in CRS. UniCRS and DCRS are further optimized based on the DialoGPT model, thereby exhibiting superior performance. In contrast, DisenCRS achieves the best scores on BLEU, ROUGE, and DIST metrics, demonstrating its overall advantages in dialogue generation. This may be due to its effective disentanglement of contextual information, which allows the model to focus on relevant aspects of the dialogue while minimizing irrelevant noise, thereby producing more coherent and contextually appropriate responses.

Human Evaluation. To further validate the effectiveness of the DisenCRS model on the response generation task, we conduct a human evaluation, with the results shown in Table 4. As observed from Table 4, DisenCRS outperforms other baselines in terms of the
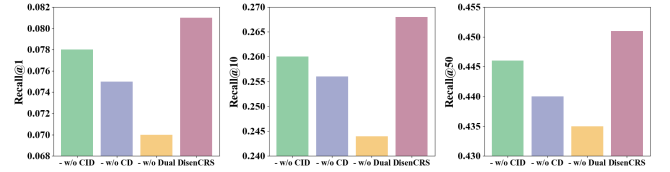
fluency and informativeness of the generated responses, demonstrating its advantage in producing coherent and information-rich responses. We believe that this advantage stems from the model's effective decoupling of focus and background information based on the dialogue history, which avoids the interference of extraneous information, thereby generating responses that are more contextually consistent and content-rich.

4.3 Impact of Dual Contextual Disentanglement Module (RQ2)

Here, to demonstrate the effectiveness of the proposed disentanglement method, we conduct ablation experiments. Specifically, we design the following three variants: - w/o CD, which refers to the removal of contrastive disentanglement; - w/o CID, which eliminates the counterfactual inference disentanglement; and - w/o Dual, which indicates the absence of the contextual disentanglement module.

The performance of the aforementioned variants on the ReDial dataset is illustrated in Figure 4. Upon examination, we can draw the following conclusions: (1) DisenCRS achieves better results than - w/o CD, which indicates that the contrastive disentanglement mechanism effectively helps the model distinguish between focus information and background information, thereby better understanding the semantic structure of contextual conversations. (2) DisenCRS outperforms - w/o CID, which shows that by introducing a counterfactual reasoning mechanism, the model can better simulate the information distribution under different contextual assumptions, thereby enhancing the ability to capture user intentions. (3) - w/o Dual has the worst effect, which further illustrates the effectiveness of the above two disentanglement strategies.

Table 5: Comparison of the adaptive prompt learning module with other methods.

Method	Recall@10	Recall@50	DIST@3	DIST@4
DisenCRS-fc	0.248	0.430	0.743	1.164
DisenCRS-bg	0.234	0.425	0.727	1.088
DisenCRS-fw	0.259	0.442	0.967	1.302
DisenCRS	**0.268**[*]	**0.451**[*]	**1.212**[*]	**1.411**[*]

4.4 Effectivenss of Adaptive Prompt Learning Module (RQ3)

To comprehensively evaluate the effectiveness of our adaptive prompt learning module, we conduct comparative experiments. We define the following variants: (1) DisenCRS-fc: represents using only the focus information from the dialogue context. (2) DisenCRS-bg: represents using only the background information from the dialogue context. (3) DisenCRS-fw: applies weighted fusion by manually setting fixed weights. We perform a grid search to select the best manual weights for the experiments.

The experimental results of the aforementioned variants are shown in Table 5. Both DisenCRS-fc and DisenCRS-bg perform worse than the full DisenCRS model, indicating that neither the background information nor the focus information can be discarded. They both play an important role in understanding user intentions. And DisenCRS-fc outperforms DisenCRS-bg, indicating that in most dialogues, focus information better reflects the user's true interests than background information. Furthermore, the DisenCRS-fw fusion strategy performs worse than the adaptive prompt learning module, suggesting that simply relying on fixed weights to fuse focus information and background information in dialogue makes it difficult to cope with diverse contextual requirements. They fail to fully leverage the complementary relationship between focus and background information. Overall, the adaptive prompt learning module is capable of dynamically selecting focus and background information based on the dialogue context, effectively filtering out noise while preserving key useful information.

4.5 Hyperparameter Study (RQ4)

4.5.1 The impact of the number of prompts η in the prompt pool.
To investigate the impact of the number of prompts in the prompt pool on model performance, we set different numbers of prompts η (such as 5, 10, 15, and 20), as shown in Figure 5. When the number of prompts in the prompt pool is small, the weight fusion granularity of the focus information and background information is coarse, making the model difficult to find the most suitable weight combination in different dialogue context scenarios. As the number of prompts in the prompt pool increases, the granularity of weight distribution becomes finer, enabling the model to select more appropriate weights for different dialogue scenarios, thereby improving performance. When the number of prompts in the prompt pool becomes excessive (e.g., $\eta>10$), the differences between weights may be too subtle, placing a higher challenge for the prompt selector and leading to performance degradation.

4.5.2 The impact of contextual disentanglement loss weights λ. To investigate the impact of the contextual disentanglement loss on model performance, we introduce a weighting factor λ before the
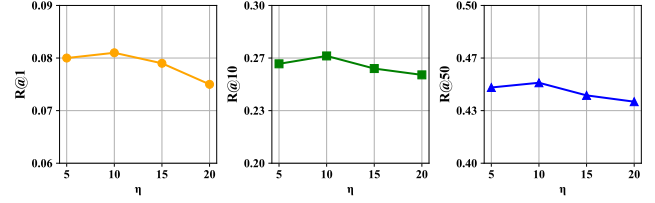
Figure 5: Comparison of the number of prompts η in the prompt pool.
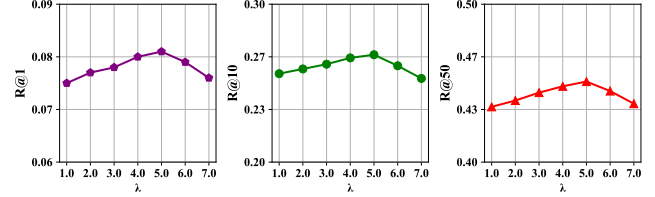
Figure 6: Comparison of the contextual disentanglement loss weights λ.

disentanglement loss. By adjusting this weighting factor (ranging from 1.0 to 7.0), we conduct multiple experiments to analyze its effect on the recommendation task. As shown in Figure 6, experimental results show that the weight of the contextual disentanglement loss has a significant impact on the overall performance of the model. When the weight is too small, the disentanglement effect is insufficient, negatively affecting the recommendation performance; when the weight is too large, excessive disentanglement weakens the performance of the recommendation task.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a contextual disentanglement approach for conversational recommendation, named DisenCRS. DisenCRS integrates contrastive disentanglement and counterfactual inference disentanglement techniques to effectively separate the focus information and the background information in the conversation context in an unsupervised manner. Moreover, we design an adaptive prompt learning module that can adaptively perceive the context information and dynamically select a suitable prompt from the prompt pool to support downstream recommendation and response generation tasks. Our extensive experiments on two benchmarking datasets verify the effectiveness of our DisenCRS model.

Although this paper introduces and highlights the potential of contextual disentanglement for advancing research on CRS, it represents a preliminary exploration and there is still room for improvement. In the future, we plan to explore large language models for disentanglement to enhance CRS, improving their ability to recognize different types of information in context.

References

[1] Guojia An, Jing Sun, Yuhan Yang, and Fuming Sun. 2024. Enhancing Collaborative Information with Contrastive Learning for Session-based Recommendation. *Information Processing and Management* 61, 4 (2024), 103738.

[2] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. *arXiv preprint arXiv:1908.05391* (2019).

[3] Zhendong Chu, Hongning Wang, Yun Xiao, Bo Long, and Lingfei Wu. 2023. Meta Policy Learning for Cold-Start Conversational recommendation. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining.* 222–230.

[4] Huy Dao, Yang Deng, Dung D Le, and Lizi Liao. 2024. Broadening the View: Demonstration-augmented Prompt Learning for Conversational Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 785–795.

[5] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. 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.* 1431–1441.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4171–4186.

[7] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 8142–8152.

[8] Dongxiao He, Jinghan Zhang, Xiaobao Wang, Meng Ge, Zhiyong Feng, Longbiao Wang, and Xiaoke Ma. 2024. TUT4CRS: Time-aware User-preference Tracking for Conversational Recommendation System. In *Proceedings of the 32nd ACM International Conference on Multimedia.* 5856–5864.

[9] Taeho Kim, Juwon Yu, Won-Yong Shin, Hyunyoung Lee, Ji-hui Im, and Sang-Wook Kim. 2023. LATTE: A Framework for Learning Item-features to Make a Domain-Expert for Effective Conversational Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1144–1153.

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing.* 1746–1751.

[11] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining.* 304–312.

[12] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive Path Reasoning on Graph for Conversational Recommendation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2073–2083.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 7871–7880.

[14] Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-seng Chua, and Donghong Ji. 2025. Revisiting Conversation Discourse for Dialogue Disentanglement. *Transactions on Information Systems* 43, 1 (2025), 1–34.

[15] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* 9748–9758.

[16] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2021. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *Transactions on Information Systems* 39, 4 (2021), 1–29.

[17] Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie, and Dangyang Chen. 2023. TREA: Tree-Structure Reasoning Schema for Conversational Recommendation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.* 2970–2982.

[18] Xinshu Li, Mingming Gong, and Lina Yao. 2024. Self-Distilled Disentangled Learning for Counterfactual Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1667–1678.

[19] Dongding Lin, Jian Wang, and Wenjie Li. 2023. COLA: Improving Conversational Recommender Systems by Collaborative Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 4462–4470.

[20] Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021. Unsupervised conversation disentanglement through co-training. *arXiv preprint arXiv:2109.03199* (2021).

[21] Yinhan Liu. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* 364 (2019).

[22] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations.*

[23] Heli Ma, Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Yi Bin, and Yang Yang. 2024. Ask or Recommend: An Empirical Study on Conversational Product Search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management.* 3927–3931.

[24] Kong Menglin, Jia Wang, Yushan Pan, Haiyang Zhang, and Muzhou Hou. 2024. C²DR: Robust Cross-Domain Recommendation based on Causal Disentanglement. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining.* 341–349.

[25] Mingjie Qian, Yongsen Zheng, Jinghui Qin, and Liang Lin. 2023. HutCRS: Hierarchical User-Interest Tracking for Conversational Recommender System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.* 10281–10290.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9.

[27] Xubin Ren, Lianghao Xia, Jiashu Zhao, Dawei Yin, and Chao Huang. 2023. Disentangled Contrastive Collaborative Filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1137–1146.

[28] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to Ask Appropriate Questions in Conversational Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 808–817.

[29] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Proceedings of the 15th International Conference on Semantic Web.* 593–607.

[30] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[31] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval.* 235–244.

[32] Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. 2022. Conversational Recommendation via Hierarchical Information Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2201–2205.

[33] Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023. Improving Conversational Recommendation Systems via Counterfactual Data Simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2398–2408.

[34] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of the 28th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1929–1937.

[35] Yibiao Wei, Jie Zou, Weikang Guo, Guoqing Wang, Xing Xu, and Yang Yang. 2025. MSCRS: Multi-modal Semantic Graph Prompt Learning Framework for Conversational Recommender Systems. *arXiv preprint arXiv:2504.10921* (2025).

[36] Mingshi Yan, Fan Liu, Jing Sun, Fuming Sun, Zhiyong Cheng, and Yahong Han. 2024. Behavior-Contextualized Item Preference Modeling for Multi-Behavior Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 946–955.

[37] Lu Zhang, Chen Li, Yu Lei, Zhu Sun, and Guanfeng Liu. 2024. An Empirical Analysis on Multi-turn Conversational Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 841–851.

[38] Xiaotong Zhang, Xuefang Jia, Han Liu, Xinyue Liu, and Xianchao Zhang. 2024. A Goal Interaction Graph Planning Framework for Conversational Recommendation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vol. 38. 19578–19587.

[39] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards Empathetic Conversational Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems.* 84–93.

[40] Xiaoyu Zhang, Xin Xin, Dongdong Li, Wenxuan Liu, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023. Variational Reasoning over Incomplete Knowledge Graphs for Conversational Recommendation. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining.* 231–239.

[41] Xiaokun Zhang, Bo Xu, Zhaochun Ren, Xiaochen Wang, Hongfei Lin, and Fenglong Ma. 2024. Disentangling ID and Modality Effects for Session-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1883–1892.

[42] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT : Large-Scale

Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

[43] Sen Zhao, Wei Wei, Ding Zou, and Xianling Mao. 2022. Multi-view Intent Disentangle Graph Networks for Bundle Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4379–4387.

[44] Yu Zheng, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2022. Disentangling Long and Short-Term Interests for Recommendation. In *Proceedings of the ACM Web Conference*. 2256–2267.

[45] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference*. 2980–2991.

[46] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *Proceedings of the 26th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1006–1014.

[47] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4128–4139.

[48] Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-CRS: Coarse-to-Fine Contrastive Learning for Conversational

Recommender System. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 1488–1496.

[49] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM on Web Conference*. 1726–1732.

[50] Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards Question-based Recommender System. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 881–890.

[51] Jie Zou, Jimmy Huang, Zhaochun Ren, and Evangelos Kanoulas. 2022. Learning to ask: Conversational product search via representation learning. *Transactions on Information Systems* 41, 2 (2022), 1–27.

[52] Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving Conversational Recommender Systems via Transformer-based Sequential Modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2319–2324.

[53] Jie Zou, Aixin Sun, Cheng Long, and Evangelos Kanoulas. 2024. Knowledge-Enhanced Conversational Recommendation via Transformer-Based Sequential Modeling. *Transactions on Information Systems* 42, 6 (2024), 1–27.