MSCRS: Multi-modal Semantic Graph Prompt Learning Framework for Conversational Recommender Systems

Yibiao Wei
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
weiyibiao12138@gmail.com

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

Weikang Guo
Southwestern University of Finance and Economics
Chengdu, Sichuan, China
guowk@swufe.edu.cn

Guoqing Wang
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
gqwang0420@hotmail.com

Xing Xu
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
xing.xu@uestc.edu.cn

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

Abstract

Conversational Recommender Systems (CRSs) aim to provide personalized recommendations by interacting with users through conversations. Most existing studies of CRS focus on extracting user preferences from conversational contexts. However, due to the short and sparse nature of conversational contexts, it is difficult to fully capture user preferences by conversational contexts only. We argue that multi-modal semantic information can enrich user preference expressions from diverse dimensions (e.g., a user preference for a certain movie may stem from its magnificent visual effects and compelling storyline). In this paper, we propose a multi-modal semantic graph prompt learning framework for CRS, named MSCRS. First, we extract textual and image features of items mentioned in the conversational contexts. Second, we capture higher-order semantic associations within different semantic modalities (collaborative, textual, and image) by constructing modality-specific graph structures. Finally, we propose an innovative integration of multi-modal semantic graphs with prompt learning, harnessing the power of large language models to comprehensively explore high-dimensional semantic relationships. Experimental results demonstrate that our proposed method significantly improves accuracy in item recommendation, as well as generates more natural and contextually relevant content in response generation. Code and extended multi-modal CRS datasets are available at https://github.com/BIAOBIAO12138/MSCRS-main.

CCS Concepts

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

*Corresponding author.

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730040

Keywords

Multi-modal, Conversational Recommendation, Prompt Learning

ACM Reference Format:
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. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730040

1 INTRODUCTION

Conversational Recommender Systems (CRSs) [5, 19, 35], as an emerging research direction that integrates natural language processing and recommendation technologies, aim to precisely capture users' preferences [14] through multi-turn conversations and thus provide personalized recommendations. Early CRS [13, 21, 28, 29, 41] primarily focused on analyzing and modeling user preferences from conversational contexts. However, due to the limited nature of conversational contexts (e.g., short and sparse), relying solely on extracting user preferences from conversations makes it challenging to achieve personalized modeling. This results in recommendations being confined to common options and failing to capture more granular user needs, thus affecting accuracy.

Accordingly, to overcome this shortcoming, some approaches attempt to introduce external knowledge, including structured knowledge graph [2, 34], hypergraph [48], unstructured item reviews [26, 51], metadata [43], and entities appearing in similar conversations [6] to enhance the user representation for improving CRS. Although these methods have made significant progress in the field of CRS, they mainly focus on a single textual modality and fail to fully utilize the rich multi-modal semantic information of items. As shown in Figure 1, users' descriptions of their preferences within a single conversation are often based on their rich multi-modal experiences involving visual and textual information in reality. We argue that leveraging multi-modal semantic information is highly helpful for modeling user preferences comprehensively. First, the multi-modal features of items (e.g., posters, trailers, reviews) enrich user preference expressions from different perspectives, capturing users' multi-modal preferences that cannot be fully expressed solely through conversational contexts. Second, collaborative information,

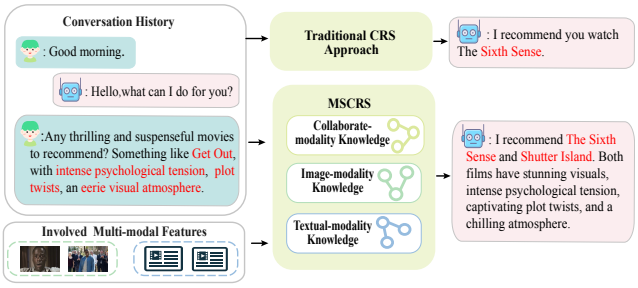


Figure 1: Comparison between traditional CRS models and our MSCRS model.

as a type of multi-modal semantic information, provides an extra perspective on user relationships (e.g., analyzing the same entities mentioned in different user conversations can uncover latent preference connections). The combination of collaborative information with multi-modal features not only enhances the modeling of explicit user preferences but also enables better identification of implicit needs.

Although we highlight the potential of multi-modal semantic information, integrating these different modalities during the conversation remains a challenge. Many existing methods enhance entity representations by incorporating external information (e.g., reviews [26] and metadata [51]). However, it is well known that a semantic gap exists between conversational context and these external data, due to the inconsistencies in expression form and information structure between them, complicating the direct integration of different data sources. For instance, users expressed preferences in conversations are often subjective sentiments described through natural language, whereas reviews and metadata tend to contain more objective item characteristics and user evaluations. This makes it difficult to align semantics among different data sources, eventually affecting the performance of recommendations. While there have been some research attempts to directly fuse different data sources (e.g., contrastive learning [51]), the quest to align different modality data can be counterproductive because of the unique semantic associations within each modality. To address these challenges, in this paper, we propose a novel **M**ulti-modal **S**emantic graph prompt learning framework for **CRS** (**MSCRS**). Specifically, first, we extract textual descriptions of items by employing large language models (LLMs) and extract images of items from an external database (i.e., IMDb[1] in this work). Afterward, we utilize pre-trained models to extract textual and image features of the items. Second, as direct alignment of different modality features may destroy the intra-modal semantic associations, we construct a modal-specific semantic graph for the semantic features of each modality. For collaborative modality, we extract entities (including items and item-related entities) from conversational contexts and construct a collaborative semantic graph based on the co-mention frequency of these entities. For the textual and image modalities, we construct a textual semantic graph and an image semantic graph by exploiting intra-modal feature similarity based on the extracted textual and image features. By sharing the initial embeddings of all semantic graphs, we achieve an effective fusion of the three semantic graphs. This approach avoids direct fusion and alignment of different modality features while effectively preserving the semantic relationships within each modality. Third, we propose a novel approach that integrates multi-modal semantic graphs (textual semantic graph, image semantic graph, and collaborative semantic graph) with LLMs. This integration leverages the advantages of graph neural networks (GNNs) in aggregating neighborhood information, providing topological insights to LLMs. This also enables LLMs to fully exploit high-dimensional semantic associations, guiding the selection of relevant information from textual inputs and controlling the generation process. In this way, it not only improves the performance of the recommendation task but also generates more expressive responses for the conversation task. We summarize our contributions as follows:

- We propose a novel CRS model, MSCRS, which integrates multi-modal semantic information, including collaborative information and multi-modal features. To the best of our knowledge, this is the first effort to leverage both collaborative information and multi-modal item features for generation-based CRS.
- MSCRS constructs semantic graphs based on intra-modal relations and avoids the cross-modal semantic gap via shared embedding. Additionally, it proposes a novel framework to combine multi-modal semantic graphs with prompt learning, which leverages LLMs to explore higher-order semantic associations, enabling more accurate user preference modeling and more natural response generation.
- To support multi-modal CRS research, we supplemented the multi-modal features for two widely used CRS datasets. Experimental results on two widely used benchmark datasets demonstrate that our proposed MSCRS outperforms state-of-the-art baselines in both item recommendation and response generation.

2 RELATED WORK

2.1 Conversational Recommendation

As dialogue systems [33, 42, 45] have rapidly evolved, CRS [5, 14, 35] have become a thriving field of research. CRS aims to discern user preferences through multi-turn interactions and suggest items that users might find appealing. Current CRS can generally be divided into two categories: attribute-based CRS and generation-based CRS.

Attribute-based CRS [7, 17, 35] typically aims to enhance recommendation performance and reduce the number of dialogue turns required to complete recommendation tasks. They focus on asking clarifying questions [27, 52, 53] and gradually identifying the best candidate set based on user preferences. For example, many studies typically employed reinforcement learning [7, 16, 17] or bandit-based approaches [20], to optimize the long-term benefits of asking clarifying questions.

Generation-based CRS [1, 6, 30, 37, 48, 54, 55] emphasizes user interaction through natural language dialogue, intending to provide accurate recommendations and coherent responses, thereby enhancing its relevance to real-world application scenarios. For instance, Chen et al. [4] proposed a method that incorporates external knowledge to enhance recommendation and conversation effectiveness. Zhou et al. [49] proposed entity-based and word-based knowledge graphs to enrich entity modeling and generate high-quality responses. Zhou et al. [51] considered three types of data, i.e., reviews, knowledge graphs, and conversational contexts, and

[1]https://www.imdb.com

designed a coarse-to-fine contrastive learning approach to integrate these different data types. Besides knowledge graphs, Wang et al. [37] integrated recommendation and conversation tasks into an LLM through a unified prompt learning framework. Dao et al. [6] examined the application of semantically similar conversational contexts to enhance soft prompts in prompt learning. However, users' preferences are often based on their past multi-modal experiences, making the multi-modal features of items crucial for modeling user preferences. To this end, different from the aforementioned studies, we incorporate the multi-modal semantic information of items into CRS.

2.2 Multi-modal Recommendation

Multi-modal recommendation enhances performance by leveraging the multi-modal features of items. Early approaches [11, 40] typically incorporate multi-modal features of items as a complement to ID features within the collaborative filtering framework. Due to the development of GNNs, an increasing number of studies [38, 40] combine multi-modal features with graphs. For example, Wei et al. [40] proposed user and item representations in different modalities through specific modality graph structures. Zhang et al. [44] considered specific modality semantic graphs and integrated them with graph-based collaborative filtering methods. To address the semantic gap between different modalities, Zhou et al. [50] proposed a novel method and aligned these features through contrastive learning. Similarly, Wei et al. [38] employed adversarial learning to learn user and item representations across different modalities and fuse these representations through cross-modal contrastive learning. Besides contrastive learning, Wei et al. [39] fused user preferences from different modalities through ranking distillation [36]. Unlike the approaches mentioned above, we generate various semantic graphs and combine them with prompt learning in the CRS scenario. This allows LLMs to understand the topological structure of GNNs, thereby guiding the generation of recommendation and conversation tasks for CRSs.

3 PRELIMINARIES

We denote the set of items by $\mathcal{I} = \{i_1, i_2, \ldots, i_N\}$ and the set of conversational contexts by $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$. In the conversations $\mathcal{S}$, we extract all entities involved into the set $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$, with $\mathcal{I} \subseteq \mathcal{E}$. N is the number of items, M is the number of conversational contexts, and K is the number of all entities. Additionally, we collect the multi-modal features of items, denoted as $x_i^m \in \mathbb{R}^{d_m}$, where d_m denotes the dimension of the feature, and $m \in \{t, v\}$, where t denotes textual modality and v denotes image modality. A conversational context $s \in \mathcal{S}$ is represented as a collection of utterances c, expressed as $s = \{c_b\}_{b=1}^{n}$. In the b-th turn of the conversation, each utterance c_b consists of a sequence of words w, expressed as $c_b = \{w_j\}_{j=1}^{m}$. The set of words is denoted by $\mathcal{W}$. As the conversation progresses, utterances are aggregated into a conversation history. CRS uses this history to infer user preferences and generate conversation responses. During the b-th turn, the recommender component recommends a set of candidate items from the complete item set $\mathcal{I}$ based on the modeled user preferences. Meanwhile, the conversation component generates the next utterance c_b in response to the preceding conversation.

4 METHODOLOGY

Our approach consists of four main parts, as shown in Figure 2. First, we extract the corresponding multi-modal data for items and then encode them. Second, we introduce the multi-modal semantic graph modeling component, which primarily integrates the proposed specific semantic graphs of multiple modalities. Finally, we elaborate on our methods for recommendation and conversation tasks through multi-modal semantic graph prompt learning.

4.1 Multi-modal Feature Encoding

As shown in Figure 2, our model primarily considers three types of data: conversation history, textual descriptions of items, and image features of items. Next, we will introduce the feature extraction and encoding methods for each of these three types of data.

Encoding Conversation History. Like previous work [49, 51], we map the items (e.g., movies) and related entities (e.g., actors) in the conversation history to the knowledge graph DBpedia [2] to capture the intricate interconnections between entities. By incorporating the knowledge graph DBpedia, we enhance the semantic information of these entities. The knowledge graph $\mathcal{G}_{kg}$ consists of a set of entities $\mathcal{E}$ and a set of edges $\mathcal{R}$. It uses triples $\langle e_1, r, e_2 \rangle$ to store semantic facts, where $e_1, e_2 \in \mathcal{E}$ represent items or item-related entities, and $r \in R$ denotes the relationship between e_1 and e_2. We apply R-GCN [32] for encoding $\mathcal{G}_{kg}$. Specifically, the representation of node e at the $(l+1)$-th layer is computed as:

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

where $\mathbf{n}_e^{(l)}$ represents the embedding of node e at the l-th layer, and $\mathcal{E}_e^r$ refers to the set of neighboring nodes for e associated with relation r. The matrix $\mathbf{W}^{(l)}$ applies a learnable transformation to the node embeddings at the l-th layer, and $\mathbf{W}_r^{(l)}$ transforms the embeddings of neighboring nodes connected by relation r using a relation-specific matrix. The factor $Z_{e,r}$ normalizes the contribution of each neighboring node.

Encoding Textual Descriptions of Items. Based on the extensive general knowledge of existing LLMs, we employ the powerful GPT language model (GPT-4o) to extract textual descriptions v_i^t for an item i:

$$v_i^t = \mathbf{F}_{\text{GPT}}(i; \theta_p), \tag{2}$$

where θ_p is the prompt template, v_i^t represents the text description generated by $\mathbf{F}_{\text{GPT}}$. The specific prompt template is provided in Figure 3. After generating the textual description v_i^t, we utilize the pre-trained model RoBERTa [25] to encode it:

$$x_i^t = \text{AvgPool}\left(\mathbf{F}_{\text{RoBERTa}}(v_i^t; \theta_r)\right), \tag{3}$$

where AvgPool represents the average pooling operation, θ_r denotes all the parameters of RoBERTa. Finally, we obtain $X^t = \{x_1^t, x_2^t, x_3^t, \ldots, x_N^t\}$, which denotes the textual features of all items.

Encoding Image Features of Items. The visual features of items contain rich semantic information. We collect multiple still images $\{m_1^v, m_2^v, \ldots, m_l^v\}$ of item i using web scraping from IMDb. Then, we extract the image representations corresponding to the item i using the pre-trained model ViT [9]:

$$Z_i = \{z_1^v, z_2^v, \ldots, z_l^v\} = \mathbf{F}_{\text{ViT}}(m_1^v, m_2^v, \ldots, m_l^v; \theta_v). \tag{4}$$

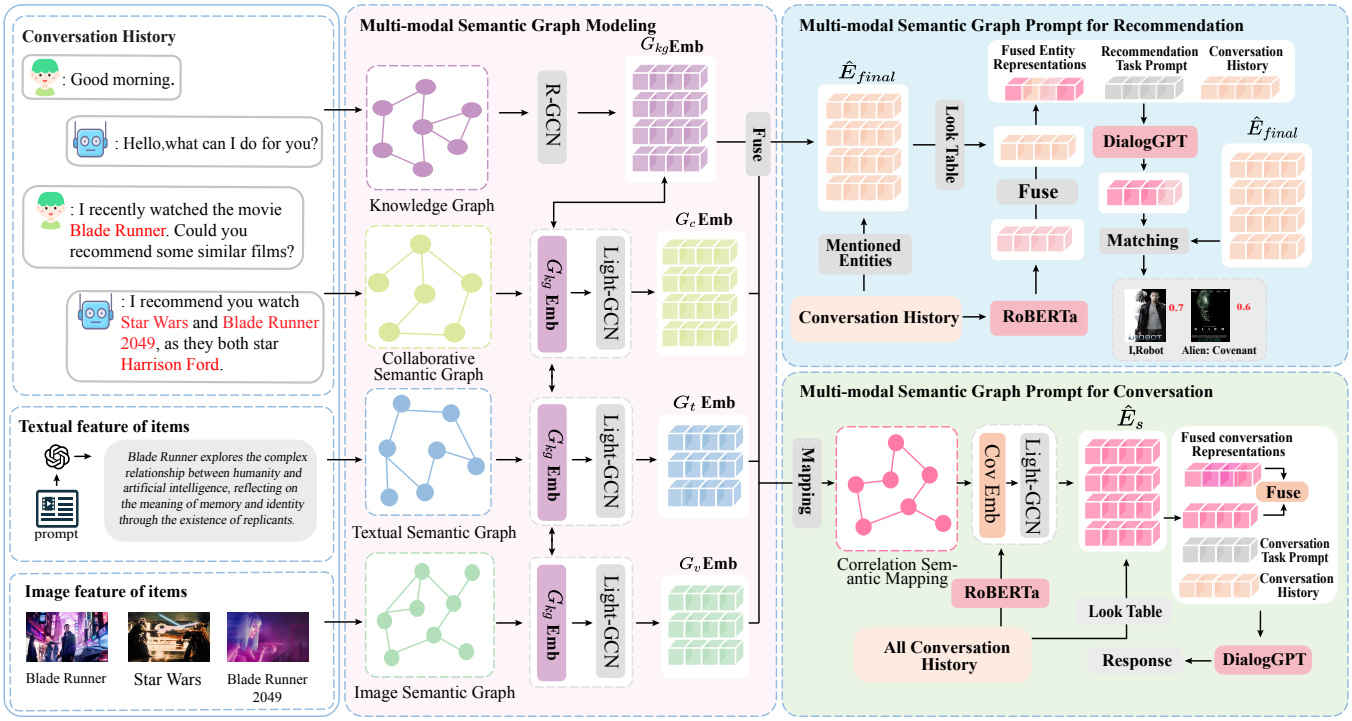


Figure 2: The framework of MSCRS.

Please provide a comprehensive and detailed description of the movie '{movie_name}', including the following aspects, with each description limited to 30 tokens:
Theme: The core idea or main themes explored in the movie, such as survival, isolation, or moral choices.
Emotion: The primary emotional impact of the film on viewers, such as fear, hope, tension, or joy.
Cultural Significance: Explain the cultural and societal impact of the movie. How does it contribute to or influence the genre, and what discussions or reflections does it provoke?
Plot Overview: Provide a brief summary of the film's plot, focusing on key elements and the setting.

Figure 3: The prompt template.

Then, we calculate the average of these embeddings to obtain the image feature representation for each item i:

$$x_i^v = \frac{1}{l}\sum_{j=1}^{l} z_j^v, \tag{5}$$

where l denotes the number of images of the item i, x_i^v represents the image feature representation of item i. Finally, we obtain the image features of all items as $X^v = \{x_1^v, x_2^v, x_3^v, \ldots, x_N^v\}$.

After encoding, we can generate representations for the textual features of items, image features of items, and knowledge graph features of all entities. Next, we explore how to model and fuse these multi-modal features to obtain a unified representation.

4.2 Multi-modal Semantic Graph Modelling

Collaborative Semantic Graph. Although the knowledge graph models the complex real-world knowledge among global entities to some extent, it still faces issues such as noise, errors, inconsistent data, and "data silos". These problems can impact the accuracy and reliability of downstream recommendation tasks. To address these challenges, we introduce the collaborative semantic graph. The collaborative semantic graph models the relationships between entities from a co-mention perspective, thereby enhancing the structural information of the entities. The collaborative semantic graph $\mathcal{G}_c = (\mathcal{E}, \mathcal{R}_c)$, where $\mathcal{R}_c$ is the set of edges. The matrix $\mathbf{C} \in \mathbb{R}^{K\times K}$ is a sparse matrix representing the co-mention counts between entities, where K is the total number of entities. Specifically, the elements of the matrix are defined as:

$$\mathbf{C}_{i,j} = \sum_{y=1}^{Y} \text{count}(e_i, e_j \mid Q_y), \tag{6}$$

where $Q_y = [e_1, e_2, \ldots, e_K](e \in \mathcal{E})$ denote the entities that appear in a single conversation, and Y is the total number of conversational contexts in the train data. Each element $\mathbf{C}_{i,j}$ quantifies the frequency with which entities e_i and e_j are co-mentioned across all conversations, thereby revealing their potential associations. This matrix $\mathbf{C}$ serves as the foundation for constructing the collaborative semantic graph $\mathcal{G}_c$, where an edge is established between entities e_i and e_j if their co-mention count exceeds a predefined threshold, reflecting their semantic relationships.

As shown in Eq. (1), we adopt an embedding table $\mathbf{N}^1 \in \mathbb{R}^{K\times d}$ generated by a layer of R-GCN as the initial embedding table $\mathbf{E}_c^{(0)} \in \mathbb{R}^{K\times d}$ for the collaborative semantic graph $\mathcal{G}_c$. Then, we utilize LightGCN [12] for encoding $\mathcal{G}_c$. LightGCN streamlines the graph convolution operations by omitting feature transformation and nonlinear activation components, enhancing recommendation effectiveness while also facilitating the optimization of the model. Specifically, the representations for items at the l-th layer of graph

convolution in $\mathcal{G}_c$ are derived as follows:

$$\mathbf{E}_c^{(l)} = (\mathbf{D}_c^{-\frac{1}{2}}\mathbf{C}\mathbf{D}_c^{-\frac{1}{2}})\mathbf{E}_c^{(l-1)}, \tag{7}$$

where $\mathbf{D}_c \in \mathbb{R}^{K\times K}$ is the degree matrix. We obtain l-layer representations from the l-layer collaborative semantic graph, and then generate the average entity embedding table $\hat{E}_c$ using average pooling:

$$\hat{\mathbf{E}}_c = \text{AvgPool}([\mathbf{E}_c^{(0)}, \mathbf{E}_c^{(1)}, \ldots, \mathbf{E}_c^{(l)}]). \tag{8}$$

Textual and Image Semantic Graph. User experience with actual items often stems from multi-modal perception (e.g., when watching a movie, users not only focus on the plot and dialogue but are also influenced by visual effects and the soundtrack, which together shape their viewing experience). Meanwhile, the multi-modal features of items provide rich and valuable information. In this section, we propose modality-specific semantic graphs to comprehensively model the multi-modal features of items. Grounded in the idea that similar items are more inclined to interact than dissimilar items [44], we evaluate the semantic relationship between two items based on their similarity. In Section 4.2, we obtain the text features $\mathbf{X}^t \in \mathbb{R}^{N\times d_t}$ and image features $\mathbf{X}^v \in \mathbb{R}^{N\times d_v}$ of the items. We calculate the semantic relevance between modality-specific features using cosine similarity:

$$\mathbf{A}_{ij}^m = \frac{(\boldsymbol{x}_i^m)^\top \boldsymbol{x}_j^m}{\|\boldsymbol{x}_i^m\|\|\boldsymbol{x}_j^m\|}, \tag{9}$$

where $m \in \{t, v\}$, $\mathbf{A}^m \in \mathbb{R}^{N\times N}$. A higher value of $\mathbf{A}_{ij}^m$ indicates a stronger semantic correlation between items i and j within modality m. Typically, the adjacency matrix of a graph is expected to be nonnegative; however, $\mathbf{A}_{ij}$ spans the interval $[-1, 1]$. Consequently, we set the negative values to zero. Furthermore, common graph structures tend to be much sparser than fully connected graphs, which not only incurs higher computational costs but may also introduce extraneous and insignificant edges. We perform kNN sparsification [3] on the dense graph A^m. For each item i, we retain only the top-k edges with the highest confidence scores:

$$\tilde{\mathbf{A}}_{ij}^m = \begin{cases} 1 & \text{if } \mathbf{A}_{ij}^m \in \text{top-}k(\mathbf{A}_i^m), \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $\tilde{\mathbf{A}}_{ij}^m$ is the sparsified adjacency matrix. Similar to Eq. (7), we adapt the LightGCN to encode the modality-specific semantic graph:

$$\mathbf{E}_m^{(l+1)} = (\mathbf{D}_m^{-\frac{1}{2}}\tilde{\mathbf{A}}^m\mathbf{D}_m^{-\frac{1}{2}})\mathbf{E}_m^{(l)}, \tag{11}$$

where $\mathbf{D}_m \in \mathbb{R}^{N\times N}$ denotes the degree matrix of the modality-specific semantic graph. Consistent with the method used to initialize $\mathcal{G}_c$, we initialize the modality-specific semantic graph embedding table $\mathbf{E}_m^{(0)} \in \mathbb{R}^{N\times d}$ using the embedding table $\mathbf{N}^1 \in \mathbb{R}^{K\times d}$ enhanced by a layer of R-GCN. We obtain l-layer representations from the l-layer modality-specific semantic graph, and then generate the entity embedding table using average pooling:

$$\hat{\mathbf{E}}_t = \text{AvgPool}([\mathbf{E}_t^{(0)}, \mathbf{E}_t^{(1)}, \ldots, \mathbf{E}_t^{(l)}]), \tag{12}$$

$$\hat{\mathbf{E}}_v = \text{AvgPool}([\mathbf{E}_v^{(0)}, \mathbf{E}_v^{(1)}, \ldots, \mathbf{E}_v^{(l)}]), \tag{13}$$

where $\hat{\mathbf{E}}_t$ is the average embedding table of textual semantic graph, while $\hat{\mathbf{E}}_v$ is the average embedding table of the image semantic graph. $\hat{\mathbf{E}}_t$ and $\hat{\mathbf{E}}_v$ capture semantic information from multiple layers of modality-specific semantic graphs, providing a more comprehensive representation of the items.

While the textual and image semantic graphs specifically enhance items, we fused the two modality-specific semantic graphs using a weighting function:

$$\hat{\mathbf{E}}_m = \lambda\hat{\mathbf{E}}_t + (1-\lambda)\hat{\mathbf{E}}_v, \tag{14}$$

where $\lambda \in (0, 1)$ is the hyperparameter controling the fusion ratio.

Next, we fuse the original knowledge graph with the collaborative semantic graph:

$$\hat{\mathbf{E}}_\alpha = \text{AvgPool}([\hat{\mathbf{E}}_c, \mathbf{N}^{(1)}]). \tag{15}$$

Then, we fuse the multi-modal embeddings $\hat{E}_m$ and $\hat{\mathbf{E}}_\alpha$:

$$\hat{\mathbf{E}}_{final} = \hat{\mathbf{E}}_\alpha[\mathcal{I}] + \hat{\mathbf{E}}_m, \tag{16}$$

where $\mathcal{I}$ represents the indices of common items between $\hat{\mathbf{E}}_\alpha$ and $\hat{\mathbf{E}}_m$.

4.3 Multi-modal Semantic Graph Prompt Learning For Recommendation

In Section 4.2, we generated the fused embeddings $\hat{\mathbf{E}}_{final}$ from $\hat{\mathbf{E}}_m$ and $\hat{\mathbf{E}}_\alpha$. For a conversation $s \in S$, we can query the embedding $\mathbf{V}_s \in \mathbb{R}^{q\times d}$ of q entities involved in the conversation s from $\hat{\mathbf{E}}_{final}$. We use RoBERTa to extract the embedding $\mathbf{T}_s \in \mathbb{R}^{p\times d_c}$ of the current conversation s, where p represents the number of tokens in the conversational context, and d_c represents the dimensionality of the token embeddings. Next, we map $\mathbf{V}_s$ to the same dimension as $\mathbf{T}_s$ using a bilinear function:

$$\tilde{\mathbf{V}}_s = \mathbf{W}_1\mathbf{V}_s\mathbf{W}_2, \tag{17}$$

where $\mathbf{W}_1 \in \mathbb{R}^{p\times q}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times d_c}$ are weight matrix.

We fuse entities $\tilde{\mathbf{V}}_s$ and conversation $\mathbf{T}_s$ using a contrastive learning:

$$L_{fuse} = -(log\frac{exp(\mathbf{T}_s \cdot \tilde{\mathbf{V}}_s/\tau)}{\sum_\gamma^\Omega exp(\mathbf{T}_s \cdot \mathbf{T}_\gamma/\tau)} + log\frac{exp(\tilde{\mathbf{V}}_s \cdot \mathbf{T}_s/\tau)}{\sum_\gamma^\Omega exp(\tilde{\mathbf{V}}_s \cdot \tilde{\mathbf{V}}_\gamma/\tau)}), \tag{18}$$

where Ω denotes the number of negative examples of contrastive learning and τ denotes the temperature coefficient. The final fused entity $\hat{\mathbf{V}}_s = \tilde{\mathbf{V}}_s + \mathbf{T}_s$.

We adopt prompt learning [23] to make use of LLM in a simple and flexible way. The final prompt $\mathbf{r}_s$ for the recommendation task consists of the following three parts:

$$\mathbf{r}_s = [\hat{\mathbf{V}}_s; \mathbf{O}_{rec}; s], \tag{19}$$

where $\mathbf{O}_{rec}$ denotes embeddings for the recommendation task-specific soft prompt (random initialization), and s denotes the conversational context (word tokens). We chose DialoGPT [47] as the base LLM, which uses a Transformer-based architecture and was pre-trained on a large-scale conversation corpus extracted from Reddit, as done in existing studies [6, 37]. We input $\mathbf{r}_s$ into DialoGPT and apply a pooling layer to derive the multi-modal semantic graph enhanced user preference embedding $\hat{\mathbf{r}}_s \in \mathbb{R}^d$:

$$\hat{\mathbf{r}}_s = \text{Pooling}(\mathbf{F}_{\text{DialoGPT}}(\mathbf{r}_s; \theta_{rec})), \tag{20}$$

where θ_{rec} denotes the trainable parameters, which consist of $\hat{\mathbf{V}}_s$ and $\mathbf{O}_{rec}$. We use the last token representation of DialoGPT for item recommendations or generation tasks.

Pre-training. Due to the semantic gap between the multi-modal semantic graph enhanced prompt $\hat{\mathbf{V}}_s$ and the conversational context s, we associate them through pre-training. Specifically, we employ multi-modal semantic graph enhanced user preference embedding $\hat{\mathbf{r}}_s$ to predict the entities contained in the current conversation s. The probability of entity i is calculated as follows:

$$\mathbf{P}_{entity}(i) = \text{Softmax}(\hat{\mathbf{r}}_s \hat{\mathbf{E}}_{final}^{\top}). \tag{21}$$

We combine the fuse loss with cross-entropy to optimize the model parameters.

$$L_{pre}(\theta_{rec}) = -\sum_{j \in S_{train}} \sum_{i=1}^{K} \log \mathbf{P}^{j}_{entity}(i \mid \hat{\mathbf{r}}_s, \theta_{rec}) + \delta L_{fuse}, \tag{22}$$

where K is the number of entities, S_{train} denotes the set of all conversations in the training set. δ is the hyper-parameter to control the fuse loss weight.

Recommendation. The recommendation task predicts the probabilities of all items. Similar to Eq. (21), we generate the probabilities $\mathbf{P}_{item}(i)$ of the recommended items:

$$\mathbf{P}_{item}(i) = \text{Softmax}(\hat{\mathbf{r}}_s \hat{\mathbf{E}}_{final}^{\top}[\mathcal{I}]). \tag{23}$$

Then we train the recommendation task using a cross-entropy loss and fuse loss:

$$L_{rec}(\theta_{rec}) = -\sum_{j \in S_{train}} \sum_{i=1}^{N} y_i^j \log \mathbf{P}^{j}_{item}(i \mid \hat{\mathbf{r}}_s, \theta_{rec}) + \delta L_{fuse}, \tag{24}$$

where y_i^j is the corresponding label of the item i in the conversation instance j, and N is the number of items.

4.4 Multi-modal Semantic Graph Prompt Learning For Conversation

The conversation task aims to provide appropriate responses based on the current user utterance. Previously, we modeled the multiple relationships between different entities. By leveraging the entities present in various conversational contexts, we designed a correlation semantic mapping that integrates the contexts of all conversations in the training data. This approach allows us to capture conversational contexts with semantic similarities to the current conversational context, thereby enhancing the semantic information of the current conversational context.

Specifically, for a conversational context s, let Q_s be the set of entities involved. We obtain the multi-modal semantic graph enhanced entity set $\hat{Q}_s$ through the first-order adjacency relationships of four types of semantic graphs (knowledge graph $\mathcal{G}_{kg}$, collaborative semantic graph $\mathcal{G}_c$, textual semantic graph $\mathcal{G}_t$, and image semantic graph $\mathcal{G}_v$). If there exists an edge connection between entities, we will also include the entities connected by these edges into the enhanced entity set $\hat{Q}_s$:

$$\hat{Q}_s = Q_s \cup \bigcup_{G \in \{\mathcal{G}_{kg}, \mathcal{G}_c, \mathcal{G}_t, \mathcal{G}_v\}} \{e_j \mid \exists e_i \in E(G), A_{ij}(G) = 1\}. \tag{25}$$

Table 1: Statistics of the used datasets in our experiments.

Dataset	# Conversations	# Utterances	# Entities/Items
ReDial	10,006	182,150	64,364/6,924
INSPIRED	1,001	35,811	17,321/1,123

After obtaining the multi-modal semantic graph enhanced entity set $\hat{Q}$ for all conversational contexts S, we represent the similarity between different conversational contexts by the number of common entities. Similar to Eqs. (9, 10, 11), we construct the correlation semantic mapping $\mathcal{G}_s$. The final enhanced representation of the conversational context based on the correlation semantic mapping is as follows:

$$\hat{\mathbf{E}}_s = \text{AvgPool}([\mathbf{E}_s^{(0)}, \mathbf{E}_s^{(1)}]), \tag{26}$$

where $\mathbf{E}_s^{(0)} \in \mathbb{R}^{(p \times d_c) \times M}$ is initialized by encoding all conversational contexts into embeddings with RoBERTa, where M is the number of all conversational contexts. We utilize MLP to simulate the neighbor aggregation of one layer of LightGCN and generate the enhanced representations $\hat{\mathbf{E}}_s$. For a conversational context $s \in S$, we fuse $\hat{\mathbf{T}}_s = \mathbf{T} + \tilde{\mathbf{V}}_s$ and $\hat{e}_s \in \hat{\mathbf{E}}_s$ to generate the enhanced context embedding $\tilde{\mathbf{T}}_s$:

$$\tilde{\mathbf{T}}_s = \text{AvgPool}([\text{Reshape}(\hat{e}_s), \hat{\mathbf{T}}_s]). \tag{27}$$

The final prompt c_s for the conversation task consists of the following three parts:

$$\mathbf{c}_s = [\tilde{\mathbf{T}}_s; \mathbf{O}_{cov}; s], \tag{28}$$

where $\mathbf{O}_{cov}$ denotes the conversation task specific soft prompt embeddings (random initialization), s denotes the conversational context (word tokens). Then we input $\mathbf{c}_s$ into DialoGPT and apply a pooling layer to derive embedding $\hat{\mathbf{c}}_s$. We use $\hat{\mathbf{c}}_s$ to drive the loss for learn θ_{gen}. The optimization function for the conversation task is shown as follows:

$$L_{gen}(\theta_{gen}) = -\frac{1}{Y} \sum_{i \in S_{train}} \sum_{j=1}^{l_i} \log \mathbf{P}(w_{i,j} \mid \hat{\mathbf{c}}_s; \theta_{gen}; w_{<j}), \tag{29}$$

where Y is the number of training contexts, l_i represents the length of the label utterance, and $w_{<j}$ denotes the words preceding the j-th position.

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Datasets. We validate the effectiveness of our model on two widely used conversation recommendation datasets, **ReDial** [19] and **INSPIRED** [10], similar to previous work [6, 37]. Both datasets are specifically designed for conversational movie recommendation, consisting of realistic conversations between users and agents about movie recommendations. We obtain movie stills from IMDb. Detailed statistics are presented in Table 1.

5.1.2 Baselines. The CRS includes two tasks: recommendation and conversation. Consequently, we compare our method with the following representative methods:

- **Popularity**: A simple metric that ranks items based on their occurrence frequency in the dataset.

Table 2: Recommendation performance comparison on ReDial and INSPIRED datasets, with the best results in bold and * indicating significant improvements over the best baseline (p-value < 0.05). Unless otherwise stated, * marks significant improvements and bold values denote the best performances in the following paper.

Model	ReDial							INSPIRED						
	Recall			NDCG		MRR		Recall			NDCG		MRR	
	@1	@10	@50	@10	@50	@10	@50	@1	@10	@50	@10	@50	@10	@50
Popularity	0.011	0.053	0.183	0.029	0.057	0.021	0.027	0.031	0.155	0.322	0.085	0.122	0.064	0.071
TextCNN	0.010	0.066	0.187	0.033	0.059	0.023	0.028	0.025	0.119	0.245	0.066	0.094	0.050	0.056
BERT	0.027	0.142	0.307	0.075	0.112	0.055	0.063	0.049	0.189	0.322	0.112	0.141	0.088	0.095
ReDial	0.010	0.065	0.182	0.034	0.059	0.024	0.029	0.009	0.048	0.213	0.023	0.059	0.015	0.023
KBRD	0.033	0.150	0.311	0.083	0.118	0.062	0.070	0.042	0.135	0.236	0.088	0.109	0.073	0.077
KGSF	0.035	0.175	0.367	0.094	0.137	0.070	0.079	0.051	0.132	0.239	0.092	0.114	0.079	0.083
TREA	0.045	0.204	0.403	0.114	0.158	0.087	0.096	0.047	0.146	0.347	0.095	0.132	0.076	0.087
COLA	0.048	0.221	0.426	-	-	0.086	0.096	-	-	-	-	-	-	
VRICR	0.054	0.244	0.406	0.138	0.174	0.106	0.114	0.043	0.141	0.336	0.091	0.134	0.075	0.085
UNICRS	0.065	0.241	0.423	0.143	0.183	0.113	0.125	0.085	0.230	0.398	0.149	0.187	0.125	0.133
DCRS	0.076	0.253	0.439	0.154	0.195	0.123	0.132	0.093	0.226	0.414	0.153	0.192	0.130	0.137
MSCRS	**0.081***	**0.264***	**0.451***	**0.161***	**0.201***	**0.128***	**0.136***	**0.096***	**0.257***	**0.425***	**0.168***	**0.202***	**0.140***	**0.148***

- **BERT** [8]: An extensively utilized pre-trained model designed for text classification applications. We fine-tune BERT to forecast a selection of potential items.
- **DialogGPT** [47]: It is a large-scale generative pre-trained model trained on extensive dialogue data, specifically optimized for generating contextually relevant and fluent conversational responses
- **GPT-2** [31]: A powerful benchmark for text generation that benefits from extensive pre-training on language models.
- **BART** [18]: A denoising autoencoder pretraining model for generation tasks.
- **Redial** [19]: This model was introduced alongside the ReDial dataset, which includes an autoencoder for recommendations and a generation model based on hierarchical RNN.
- **KBRD** [4]: The method enhances recommendation and conversation tasks by introducing an entity-based knowledge graph.
- **KGSF** [49]: This method enhances the information of entities and words through entity-based and word-based knowledge graphs.
- **TREA** [22]: This method models recommendation and conversation tasks through a multi-layer inferable tree structure.
- **COLA** [24]: It enhances conversational recommendation systems by enriching item and user representations through an interactive user-item graph and retrieving similar conversations.
- **VRICR** [46]: This method enhances the original knowledge graph through dynamic inference using variational Bayes.
- **UNICRS** [37]: This method combines the recommendation and conversation sub-tasks into the same prompt learning paradigm.
- **DCRS** [6]: This method enhances recommendation and conversation tasks by retrieving conversations that are similar to the current conversation.

5.1.3 Evaluation Metrics. We evaluate the recommendation and conversation tasks using different metrics. For recommendation, we follow [6, 37] and adopt **Recall@k** (k=1, 10, 50), **NDCG@k** (k=10, 50), and **MRR@k** (k=10, 50). For conversation, we apply both automatic and human evaluations. Automatic metrics include **BLEU-N** (N=2, 3), **ROUGE-N** (N=2, L) and **Distinct-N** (N=2, 3, 4). For human evaluation, we randomly select 100 responses from each model and ask three annotators to score them on **Fluency** and **Informativeness** (0–2 scale), then we compute the average score for all test samples.

5.1.4 Implementation Details. We trained our proposed model on a 32GB V100 GPU. In our model, we use RoBERTa [25] to encode the textual features of items and input tokens. We employ ViT [9] to extract the image features of movies, and use DialoGPT-small [47] as the base LLM. The extracted textual features have a dimensionality of 768, while the image features have a dimensionality of 1024. We then map the features of different modalities to the same dimension. We set the number of R-GCN [32] layers to 1. For the collaborative semantic graph, text semantic graph, and image semantic graph, we use a 3-layer of LightGCN [12] for encoding. We tune our soft prompt between 10 and 50 for the recommendation and conversation tasks. The batch size is set to 64 for the recommendation task and 8 for the conversation task. We use Adam [15] as the optimizer for our model and adjust our learning rate between 1e-5 and 1e-3.

5.2 Evaluation on Recommendation Task

5.2.1 Automatic Evaluation. Table 2 shows the experimental results for the recommendation task. Our MSCRS model achieves state-of-the-art performance, ranking first across all metrics on ReDial and INSPIRED. Compared to the strongest baseline DCRS, MSCRS achieves 0.081 (+6.5% for ReDial) and 0.096 (+3.2% for INSPIRED) in Recall@1, indicating higher accuracy in satisfying users immediate needs. For Recall@10, MSCRS achieves 0.264 (+4.3% for ReDial) and 0.257 (+13.7% for INSPIRED), demonstrating superior top-10 coverage. Additionally, Recall@50 scores of 0.451 (+2.7% for ReDial) and 0.425 (+2.6% for INSPIRED) show that MSCRS maintains high accuracy in longer recommendation lists.

Compared to knowledge-enhanced CRS models (KBRD, KGSF, COLA, VRICR), MSCRS significantly improves recommendation accuracy by leveraging multi-modal semantic relationships. Against LLM-based CRS models (UNICRS, DCRS), MSCRS outperforms UNICRS and DCRS on all metrics. The superior performance of the MSCRS model can be attributed to its ability to learn rich knowledge

Table 3: Automatic evaluation for the conversation task on Redial and INSPIRED datasets.

Model	ReDial							INSPIRED						
	BLEU		ROUGE		DIST			BLEU		ROUGE		DIST		
	-2	-3	-2	-L	-2	-3	-4	-2	-3	-2	-L	-2	-3	-4
DialogGPT	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
COLA	0.026	0.012	-	-	0.387	0.528	0.625	-	-	-	-	-	-	
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
MSCRS	**0.054***	**0.027***	**0.070***	**0.294***	**0.784***	**1.332***	**1.553***	**0.040***	**0.019***	**0.052***	**0.235***	**4.197***	**5.983***	**6.556***

from three different semantic structures: the collaborative semantic graph, the textual semantic graph, and the image semantic graph. By combining these graph structures with higher-order semantic relationships and prompt learning for LLMs, the MSCRS model achieves significantly better results.

5.2.2 Ablation Study. Our recommendation method mainly enhances embeddings of items and item-related entities through the collaborative semantic graph, text semantic graph, and image semantic graph. To explore their impact on model performance, we designed four ablation variants: (1) **MSCRS w/o -co**, removing the collaborative semantic graph; (2) **MSCRS w/o -t**, removing the text semantic graph; (3) **MSCRS w/o -i**, removing the image semantic graph; and (4) **MSCRS w/o -r**, removing all three graph structures. Recall@10 is used as the evaluation metric due to its simplicity and consistent trends with Recall@1 and Recall@50.

From Figure 4 (a) and (b), it can be seen that removing any of the three graphs results in drops in performance. Removing the collaborative semantic graph (MSCRS w/o -co) resulted in a decrease in Recall@10, indicating its crucial role in capturing relationships between entities. Removing the textual semantic graph (MSCRS w/o -t) and the image semantic graph (MSCRS w/o -i) also results in a decline in performance, emphasizing the importance of textual and image information for recommendation quality. Finally, the variant that removes all enhanced graph structures (MSCRS w/o -r) exhibits the lowest Recall@10 value, further demonstrating the necessity of combining multiple graph structures to improve model performance. In summary, the ablation study shows that all the collaborative semantic graph, textual semantic graph, and image semantic graph improve the effectiveness of the recommendation.

5.3 Evaluation on Conversation Task

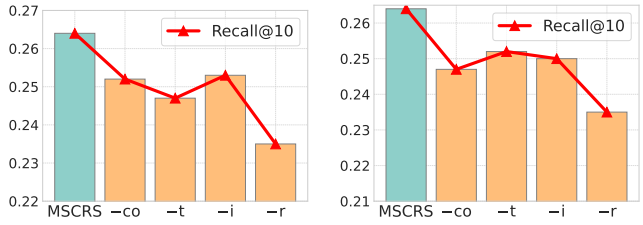
5.3.1 Automatic Evaluation. Table 3 presents the comparison of BLEU, ROUGE, and DIST scores on the ReDial and INSPIRED datasets. On the all dataset, MSCRS achieves the best performance across all three metrics, showcasing its superior text generation capability. These results highlight the outstanding conversation generation performance of MSCRS across both datasets, which can be attributed to its ability to establish complex high-order associations between enhanced entity representations and text during pre-training. Furthermore, the proposed correlation semantic mapping effectively enriches the semantic context, enabling MSCRS to generate more informative and fluent responses.

Table 4: Human evaluation for the conversation task on Redial dataset.

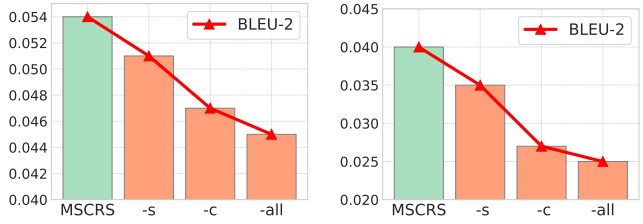
Models	Fluency	Informativeness
ReDial	1.31	0.98
KGSF	1.21	1.16
GPT-2	1.56	1.52
BART	1.48	1.43
UNICRS	1.68	1.56
DCRS	1.74	1.62
MSCRS	**1.79***	**1.67***

5.3.2 Human Evaluation. Table 4 indicates that MSCRS excels in fluency and informativeness, showcasing its robust capability to generate high-quality conversations. This is likely due to our multi-modal semantic awareness, which enhances conversation generation quality through more complex relationship modeling. ReDial achieves the lowest scores in both metrics. KGSF shows an improvement in informativeness but performs slightly worse in fluency, suggesting progress in content richness but a need for further improvement in language naturalness. GPT-2 and BART perform well in fluency and informativeness, validating the effectiveness of pre-training techniques in natural language generation tasks. UNICRS and DCRS outperform GPT-2 and BART in both metrics, with both providing optimization approaches for the conversation generation task, thereby demonstrating their capabilities in generating high-quality conversations.

5.3.3 Ablation Study. Our proposed model enhances response generation primarily through multi-modal semantic graph enhanced entity and correlation semantic mapping. To verify the effectiveness of these two modules, we designed different variants for ablation

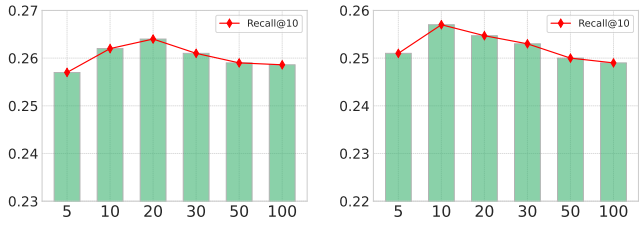


(a) ReDial (Rec.) (b) INSPIRED (Rec.)

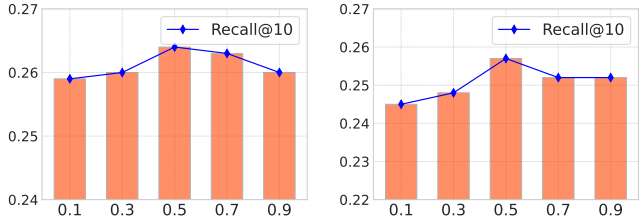


(c) ReDial (Cov.) (d) INSPIRED (Cov.)

Figure 4: Ablation studies for the recommendation and conversation tasks on the ReDial and INSPIRED datasets.



(a) ReDial (k) (b) INSPIRED (k)



(c) ReDial (λ) (d) INSPIRED (λ)

Figure 5: The impact of different k and λ values on Recall@10 for the ReDial and INSPIRED datasets.

experiments: (1) **MSCRS w/o -s** indicates that removing our proposed multi-modal semantic enhanced entity; (2) **MSCRS w/o -c** indicates removing the correlation semantic mapping enhancement component; (3) **MSCRS w/o -all** indicates the removal of both the multi-modal semantic graph enhanced entity component and the correlation semantic mapping component.

From Figure 4 (c) and (d), we observe that in both datasets, MSCRS w/o -s leads to a decline in the BLEU-2 score, indicating that multi-modal semantic graph enhanced entities improve the quality of response generation. Similarly, MSCRS w/o -c also results in a more significant performance decline, demonstrating the importance of correlation semantic mapping in enhancing conversational context and generating coherent and fluent conversation. The worst performance is observed with MSCRS w/o -all, further validating the indispensable role of these two modules in jointly enhancing the model in conversation generation. These experiments verify the independent effectiveness of the multi-modal semantic graph enhanced entity components and the correlation semantic mapping components.

5.4 Further Analysis.

5.4.1 Effect of k. In our proposed image and textual semantic graphs, we keep the top k items with the highest semantic relevance using the k-NN method [3]. We investigate the impact of different k values (Eq. (10)) on the model's performance. Specifically, we set the k values to [5, 10, 20, 30, 50, 100]. From Figure 5 (a) and (b), we can see that the ReDial and INSPIRED datasets exhibit similar trends in k value variations. The Recall@10 for the ReDial dataset reaches a higher peak at $k = 20$, while the INSPIRED dataset achieves the best performance at $k = 10$. Overall, selecting an appropriate k value is crucial. A small k value may fail to capture enough relevant items, while a large k value may introduce irrelevant noise, leading to a decline in performance. In our paper, we use the optimal k values $k = 20$ on ReDial and $k = 10$ on INSPIRED in other experiments.

5.4.2 Effect of λ. The parameter λ controls the fusion ratio between the textual semantic graph and the image semantic graph. We investigate the impact of different fusion ratios of λ on model performance. We adjust the value of λ within $[0.1, 0.3, 0.5, 0.7, 0.9]$. From Figure 5 (c) and (d), we observe that as λ varies, the Recall@10 metric exhibits similar trends on both the INSPIRED and ReDial datasets. On the ReDial dataset, the Recall@10 value peaks around $\lambda = 0.5$ before slightly declining. Similarly, on the INSPIRED dataset, the Recall@10 value reaches its peak around $\lambda = 0.5$ before starting to decline. These results indicate that the fusion ratio significantly influences model performance, and an optimal range of λ can balance the contributions of both textual and image modalities. We use the optimal values $\lambda = 0.5$ on ReDial and $\lambda = 0.5$ on INSPIRED in other experiments in our paper.

6 CONCLUSION

In this paper, we propose MSCRS, a novel multi-modal semantic graph prompt learning framework for CRS. Our approach integrates textual, image, and collaborative semantic information by constructing three semantic graphs to enhance entity representations and user preference modeling. In addition, by incorporating prompt learning with GNN-based neighborhood aggregation, MSCRS provides an LLM with topological cues, effectively guiding it to extract relevant information from text inputs and generate high-quality responses. Extensive experiments on recommendation and conversational tasks demonstrate that MSCRS improves performance in both item recommendation and response generation.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62402093) and, the Sichuan Science and Technology Program (2025ZNSFSC0479). This work was also supported in part by the National Natural Science Foundation of China under grants U20B2063 and 62220106008, and the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

References

[1] Guojia An, Jie Zou, Jiwei Wei, Chaoning Zhang, Fuming Sun, and Yang Yang. 2025. Beyond Whole Dialogue Modeling: Contextual Disentanglement for Conversational Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–11.

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics* 7 (2009), 154–165.

[3] Jie Chen, Haw ren Fang, and Yousef Saad. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research* 10, 9 (2009), 1989–2012.

[4] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. 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)*. 1803–1813.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 815–824.

[6] 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.

[7] 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.

[8] 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

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

[10] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *The Conference on Empirical Methods in Natural Language Processing*. 8142–8152.

[11] Ruining He and Julian McAuley. 2016. VBPR: visual Bayesian Personalized Ranking from implicit feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 144–150.

[12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. 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*. 639–648.

[13] Yeongseo Jung, Eunseo Jung, and Lei Chen. 2023. Towards a Unified Conversational Recommendation System: Multi-task Learning via Contextualized Knowledge Distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 13625–13637.

[14] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 1951–1961.

[15] Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014), 1–15.

[16] 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.

[17] 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 International Conference on Knowledge Discovery & Data Mining*. 2073–2083.

[18] 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.

[19] Raymond Li, Samira 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.

[20] 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. *ACM Trans. Inf. Syst.* (2021), 1–29.

[21] Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-Centric Conversational Recommendation with Multi-Aspect User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 223–233.

[22] 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.

[23] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 4582–4597.

[24] 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.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[26] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1161–1173.

[27] 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.

[28] Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1839–1851.

[29] 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.

[30] Zhangchi Qiu, Ye Tao, Shirui Pan, and Alan Wee-Chung Liew. 2025. Knowledge Graphs and Pretrained Language Models Enhanced Representation Learning for Conversational Recommender Systems. *IEEE Transactions on Neural Networks and Learning Systems* 36, 4 (2025), 6107–6121.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019), 0–9.

[32] 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 *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*. 593–607.

[33] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 3776–3783.

[34] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 4444–4451.

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

[36] Jiaxi Tang and Ke Wang. 2018. Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2289–2298.

[37] 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 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1929–1937.

[38] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[39] Wei Wei, Jiabin Tang, Lianghao Xia, Yangqin Jiang, and Chao Huang. 2024. PromptMM: Multi-Modal Knowledge Distillation for Recommendation with Prompt-Tuning. In *Proceedings of the ACM Web Conference 2024*. 3217–3228.

[40] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[41] Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2023. State Graph Reasoning for Multimodal Conversational Recommendation. *IEEE Transactions on Multimedia* 25 (2023), 3113–3124.

[42] Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural Response Generation with Meta-words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5416–5426.

[43] Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022. Improving Conversational Recommendation Systems' Quality with Context-Aware Item Meta-Information. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 38–48.

[44] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[45] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2204–2213.

[46] 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 Sixteenth ACM International Conference on Web Search and Data Mining*. 231–239.

[47] 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.

[48] Yongsen Zheng, Ruilin Xu, Ziliang Chen, Guohua Wang, Mingjie Qian, Jinghui Qin, and Liang Lin. 2024. HyCoRec: Hypergraph-Enhanced Multi-Preference Learning for Alleviating Matthew Effect in Conversational Recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2526–2537.

[49] 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 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1006–1014.

[50] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multi-modal Recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.

[51] 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 Fifteenth ACM International Conference on Web Search and Data Mining*. 1488–1496.

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

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

[54] 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.

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