Multimodal Dialog System: Relational Graph-based Context-aware Question Understanding

Haoyu Zhang
Shandong University
zhang.hy.2019@gmail.com

Meng Liu*
Shandong Jianzhu University
mengliu.sdu@gmail.com

Zan Gao
Shandong Artificial Intelligence Institute
zangaonsh4522@gmail.com

Xiaoqiang Lei
Kuaishou Technology
daimeng@kuaishou.com

Yinglong Wang
Shandong Artificial Intelligence Institute
wangyl@sdas.org

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

ABSTRACT

Multimodal dialog system has attracted increasing attention from both academia and industry over recent years. Although existing methods have achieved some progress, they are still confronted with challenges in the aspect of question understanding (*i.e.*, user intention comprehension). In this paper, we present a relational graph-based context-aware question understanding scheme, which enhances the user intention comprehension from local to global. Specifically, we first utilize multiple attribute matrices as the guidance information to fully exploit the product-related keywords from each textual sentence, strengthening the local representation of user intentions. Afterwards, we design a sparse graph attention network to adaptively aggregate effective context information for each utterance, completely understanding the user intentions from a global perspective. Moreover, extensive experiments over a benchmark dataset show the superiority of our model compared with several state-of-the-art baselines.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Multimodal Dialog System; Attribute-enhanced Text Representation; Sparse Relational Context Modeling

ACM Reference Format:
Haoyu Zhang, Meng Liu, Zan Gao, Xiaoqiang Lei, Yinglong Wang, and Liqiang Nie. 2021. Multimodal Dialog System: Relational Graph-based Context-aware Question Understanding. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475234

*Corresponding author: Meng Liu (mengliu.sdu@gmail.com) and Liqiang Nie (nieliqiang@gmail.com)

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475234

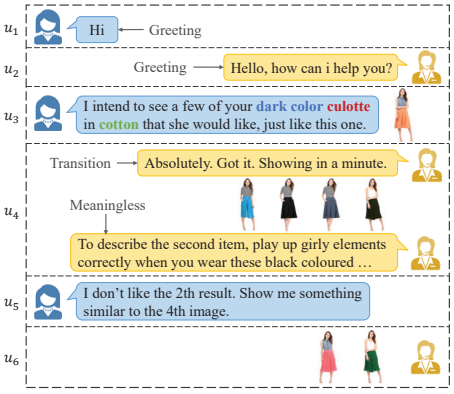

Figure 1: Illustration of a multimodal dialog system between a user and an agent, where u_i represents the i-th utterance.

1 INTRODUCTION

With the popularization of intelligent service robots, multimodal dialog systems have attracted increasing research interest, due to their significance in retail, travel, and other domains. Compared to the traditional dialog systems that purely focus on the textual conversation between users and agents [9, 37], the multimodal dialog systems allow users to express their intentions with complement images. This not only greatly improves user experience but also facilitates the agents to better understand the user intentions. As illustrated in Figure 1, the user can easily express her preferred "culottes" through a product image. Confronted with the diverse and complex conversations, how to completely comprehend users' questions and hence correspondingly give accurate system responses becomes a crucial task in multimodal dialog systems, especially in the retail domain.

Inspired by the astonishing success of deep learning techniques in various multimedia analysis tasks [10, 20, 21, 27, 35, 36], several deep learning based multimodal dialog systems have been presented and demonstrated their advanced abilities. For instance, as the pioneering study, Saha *et al.* [25] released a multimodal dialog dataset (MMD) in the retail domain, and designed a basic multimodal dialog system via multimodal hierarchical encoder and decoder. In the same year, Liao *et al.* [18] introduced a multimodal dialog system,

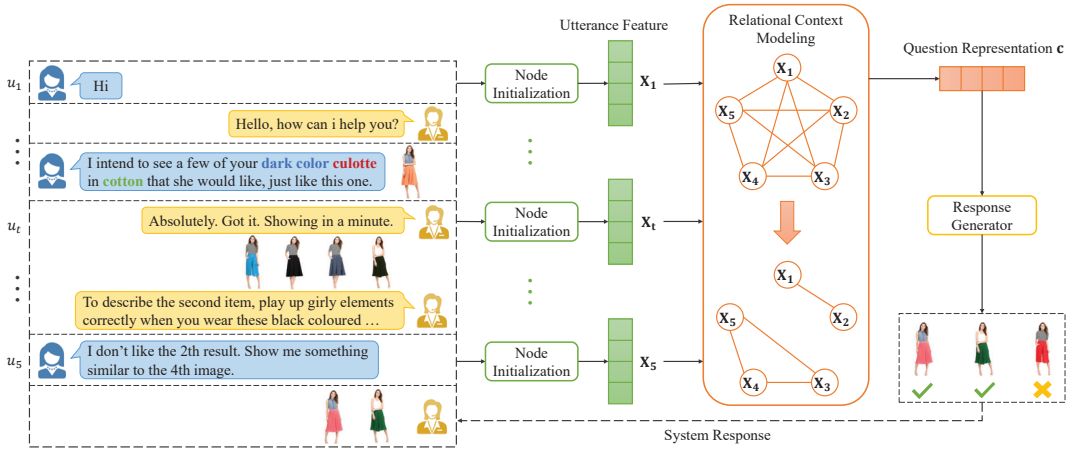

Figure 2: Schematic illustration of our TREASURE model.

which incorporates the style-tips knowledge into the neural model and adopts an Exclusive&Independent tree [17] to capture fine-grained semantics in images. Nevertheless, this system ignores the influence of the users' attention to the products. To this end, Cui *et al.* [6] proposed a user attention-guided multimodal dialog system. It learns attribute-level representations of taxonomy-guided product images through a hierarchical encoder. Regarding the diversity of domain knowledge and system responses, Nie *et al.* [22] developed a model that generates different types of system responses for user intentions in different modalities through adaptive decoders.

Although these aforementioned multimodal dialog systems have achieved promising performance, they still show certain limitations in the aspect of question understanding (*i.e.*, user intention comprehension). This is mainly for the following reasons: 1) **Textual sentence modeling**. These prior efforts [18, 22, 25] directly feed the whole textual sentence of each utterance into an online encoder (*e.g.*, Long Short-Term Memory Network, referred to as LSTM) to establish one feature vector for the entire sentence. However, simply treating a sentence holistically as one feature vector may overlook the keywords with rich semantic cues. As such, they may fail to emphasize the informative words, such as the attribute words "dark", "cotton", and "culotte" of u_3 in Figure 1, which are significant to comprehend the user intentions. Therefore, it is crucial to build a textual encoder to adaptively pay close attention to the intention-related information from textual sentences. And 2) **Relational context modeling**. Existing methods [6, 18, 22, 25] commonly utilize recurrent neural network to encode the utterance sequence, ignoring the complementary relationship between different utterances and the inconsistency of their contributions to the user intentions. As illustrated in Figure 1, each utterance has its own context. For u_5, compared with other utterances that contain no effective information, such as u_1 and u_2, u_3 and u_4 play an important role in understanding u_5 since both the textual information in u_3 and the visual information in u_4 clarify the user requirements. In this paper, the contextual information (e.g., u_3 and u_4) that is important to understanding the question (e.g., u_5) is defined as the relational context. In light of this, it is essential to consider the relational context of each utterance and adaptively reweigh their contributions for precise user intention modeling.

To solve the above issues, we propose a rela**T**ional g**R**aph-based cont**E**xt-**A**ware que**S**tion **U**nde**R**standing sch**E**me (**TREASURE**) for multimodal dialog system. As shown in Figure 2, it organizes the entire multimodal dialog into a full connected graph by treating each utterance as a node, to dynamically and adaptively capture relational context information for better question understanding. Particularly, we first design a novel node initialization module by considering the attribute matrices as the guidance information. This enables our model to focus on attribute-related textual words, enhancing the local representation of each node. Afterwards, to effectively capture the relational context for each node, we design a sparse graph attention network by simplifying the dense connections between nodes. It could adaptively aggregate few but crucial node information to strengthen the global representation of each node. Finally, based on the obtained powerful representation of the question node (*e.g.*, u_5 in Figure 2), we adopt a response generator to output the system responses. We have conducted extensive experiments over a well-known benchmark dataset and the results demonstrate the superiority of our proposed scheme. In addition, we release our code[1] to facilitate the research in this field.

The contributions of our work are three-fold:

- We devise a novel relational graph-based context-aware question understanding model, *i.e.*, TREASURE, for multimodal dialog system. It jointly integrates textual sentence and relational context modeling into a unified framework.
- To enhance the local representation of each utterance, we design an attribute-enhanced textual encoder, which enforces the model to adaptively focus on attribute-related keywords.
- To comprehend the user intentions completely, we build a sparse graph attention network, which could strengthen the global representation of the question utterance by aggregating few but crucial relational context information.

2 RELATED WORK

2.1 Unimodal Dialog Systems

Traditional dialog systems merely involve textual modality data, which can be roughly divided into two categories: open-domain [5, 14, 26, 40, 43] and task-oriented dialog systems [28, 29]. The former

[1] https://acmmmtreasure.wixsite.com/treasure.

commonly adopts retrieval-based or generation-based methods to realize a wide range of conversations with users on a variety of topics. Specifically, existing retrieval-based methods [39, 41–43] select the optimal response of the current conversation from the repository via building different response selection algorithms. Despite the promising performance, they could only return responses from the predefined corpus. To overcome this restraint, generation-based methods [26, 34] are proposed. They can automatically generate responses for questions based on the historical context, even if these responses never appear in the corpus.

Different from the open-domain dialog systems, task-oriented ones [5] are introduced to complete specific tasks in certain vertical domains, such as navigation and ticket booking. Moreover, most of them employ a typical pipeline [9, 11]. Concretely, they first utilize a natural language understanding module to classify user intentions. And then they adopt the dialog state tracker to determine the user intentions and fill in the predefined slots. Afterwards, a policy learning module is leveraged to generate the next action of the system based on the state representation. Finally, the natural language generation module would deliver system responses through predefined templates or generation methods. Despite their effectiveness, these methods suffer from several serious problems [18, 44], such as error propagation and heavy dependence on components.

With the remarkable success of deep neural networks, several end-to-end task-oriented dialog systems have been proposed recently [16, 33]. Particularly, some dialog systems consider domain knowledge to improve their performance [4, 37], and some introduce deep reinforcement learning to strengthen the generative dialog systems [7, 15, 19]. Nevertheless, all these methods only consider the single modality information, ignoring the importance of other modalities.

2.2 Multimodal Dialog Systems

With the increasing prevalence of portable computing devices and promotion from social media platforms, massive amounts of multimedia data (*e.g.*, images) are generated daily. The textual dialog systems have been insufficient to satisfy the diverse user intentions, especially in the online shopping platforms. Thereby, multimodal dialog systems have attracted extensive attention, which could flexibly express the user intentions in different modalities [3]. However, due to the lack of large-scale multimodal dialog datasets, researches in this domain have been limited.

To this end, Saha *et al.* [25] constructed a benchmark dataset MMD in the retail domain, which contains more than 150k conversation sessions and a variety of domain knowledge. Along with the dataset, they also proposed two basic tasks (*i.e.*, text response generation and image response selection) and a basic multimodal hierarchical encoder-decoder model (MHRED), a precedent in the domain of multimodal dialog. Later, considering the understanding of fine-grained visual semantics and the application of domain knowledge, Liao *et al.* [18] designed a knowledge-aware multimodal dialog system (KMD). To be specific, they built an Exclusive&Independent tree [17] to capture fine-grained semantics in images. Meanwhile, they introduced style-tips knowledge into the model through the memory network [38] and adopted deep reinforcement learning to maximize the expected future reward. However, they only considered one type of domain knowledge and ignored the users' attention to the products. Therefore, Cui *et al.* [6] designed a user attention-guided multimodal dialog system (UMD), which learns attribute-level representations of taxonomy-guided product images through a hierarchical encoder. Chauhan *et al.* [2] introduced an ordinal and attribute aware multimodal dialog system (OAM), which employs a novel position and attribute aware attention mechanism to learn enhanced image representation in the text response generation task. Considering the diversity of external knowledge and system responses, Nie *et al.* [22] proposed a multimodal dialog system with adaptive decoders (MAGIC). It can incorporate different forms of domain knowledge for different intents through intention classification, and generate general responses, knowledge-aware responses, as well as multimodal responses through adaptive decoders. Moreover, combining with transformer [30], He *et al.* [13] advanced a multimodal dialog system via capturing context-aware dependencies of semantic elements (MATE). This model uses relevant images and ordinal information in the dialog history to generate context-aware responses in the text response generation task.

Most existing multimodal dialog systems merely focus on enhancing the representation of the image by considering product attribute information, thoroughly overlooking the importance of strengthening the comprehension of the textual information. Although UMD utilizes Convolutional Neural Network (CNN) to aggregate multiple word embeddings in the sentence, both its results and interpretability are far from practicability. Furthermore, previous studies have not considered the relationship between different utterances in the context of each utterance, and ignore the importance of the complementary information for the user intention comprehension.

3 METHODOLOGY

This section details our proposed model TREASURE, which comprises three components: the node initialization module (Section 3.1), the relational context modeling module (Section 3.2), and the response generator (Section 3.3), as shown in Figure 2. In this paper, given a multimodal context $\mathcal{U} = \{u_1, ..., u_t, ..., u_N\}$, where each utterance u_t consists of textual sentences or images with sentences[2], our TREASURE organizes $\mathcal{U}$ into a graph $\mathcal{G}$ by setting each utterance as a node. In the node initialization module, we first encode the visual and textual information of each utterance u_t, and then fuse them to obtain the initial node representation $\mathbf{X}_t$. Afterwards, we leverage the relational context modeling module to construct the relationship between nodes, outputting the enhanced question representation vector $\mathbf{c}$. Finally, we feed the vector $\mathbf{c}$ into the response generator to generate the correct system responses.

3.1 Node Initialization Module

As each node u_t consists of textual sentences or both textual sentences and images, we introduce a novel node initialization module to process the utterance information. As shown in Figure 3, it is composed of three components: a textual encoder, a visual encoder, and a fusion module. To be specific, the textual encoder and visual encoder are respectively leveraged to encode the sentence and image information in each utterance. As to the fusion module, it is

[2]Note that some utterances may not contain images. In such case, we only take the textual information as the input of our node initialization module.

utilized to fuse multimodal features, obtaining the utterance representation. To get a better understanding of our node initialization module, in what follows, we will take the t-th node as an example to successively elaborate the above three components.

3.1.1 Textual Encoder. For the textual information in the given utterance u_t, we utilize pre-trained GloVe [24] to extract word embeddings $\mathbf{W} = [\mathbf{w}_1; \ldots; \mathbf{w}_K] \in \mathbb{R}^{K \times D_w}$, where K denotes the number of words in the textual sentence and D_w represents the dimension of word embeddings. As illustrated in Figure 1, attribute information in the textual sentences is vital for understanding user intentions and generating system responses. Therefore, how to enforce the model adaptively pay more attention to the attribute-related words becomes a significant issue.

To tackle this problem, we design a multi-attribute attention mechanism. Concretely, for each attribute in the attribute set[3] $\mathcal{A}$, we first build an attribute matrix $\mathbf{M}^a = [\mathbf{m}^a_1; \ldots; \mathbf{m}^a_{K_a}] \in \mathbb{R}^{K_a \times D_w}$ ($a \in \{1, 2, \ldots, A\}$) to store word embeddings of its attribute values (*e.g.*, red and black in color attribute), where A denotes the number of attributes, K_a represents the number of attribute values for the a-th attribute, and $\mathbf{m}^a_j \in \mathbb{R}^{D_w}$ is the word embedding of the j-th attribute value of the a-th attribute[4]. Afterwards, for each word in u_t, we calculate its relevance score with respect to each attribute value of all attributes as follows,

$$\alpha^a_{i,j} = \mathbf{m}^a_j \mathbf{w}^T_i, \tag{1}$$

where $\alpha^a_{i,j}$ ($i \in \{1, \ldots, K\}, j \in \{1, \ldots, K_a\}$) denotes the relevance score of the i-th utterance word in relation to the j-th value of the a-th attribute.

Based on the above-mentioned attention weights, a weighted combination of all the attribute-values is created, with correlated attribute-values to the word of high attention. Intuitively, a word-value pair should have a high similarity score if the word embedding has similar semantic to the attribute-value embedding. Then the word-specific attribute-value representation $\hat{\mathbf{w}}^a_i \in \mathbb{R}^{D_w}$ with respect to the a-th attribute is defined as follows,

$$\begin{cases} \hat{\mathbf{w}}^a_i = \sum_{j=1}^{K_a} \tilde{\alpha}^a_{i,j} \mathbf{m}^a_j, \\ \tilde{\alpha}^a_{i,j} = \frac{\exp(\alpha^a_{i,j})}{\sum_{j=1}^{K_a} \exp(\alpha^a_{i,j})}. \end{cases} \tag{2}$$

Therefore, for each word, we could obtain its word-specific attribute-value representations in regard to A attributes, *i.e.*, $\{\hat{\mathbf{w}}^1_i, \ldots, \hat{\mathbf{w}}^A_i\}_{i=1}^K$. To aggregate these representations, we adopt an attention network as follows,

$$\begin{cases} \bar{\mathbf{w}}_i = \sum_{a=1}^{A} \tilde{\beta}^a_i \hat{\mathbf{w}}^a_i, \\ \tilde{\beta}^a_i = \frac{\exp(\beta^a_i)}{\sum_{a=1}^{A} \exp(\beta^a_i)}, \\ \beta^a_i = \hat{\mathbf{w}}^a_i \mathbf{w}^T_i, \end{cases} \tag{3}$$

where $\tilde{\beta}^a_i$ ($i \in \{1, \ldots, K\}, a \in \{1, \ldots, A\}$) represents the relevance score of the i-th word in relation to the a-th attribute and $\bar{\mathbf{w}}_i$ is the output representation that incorporates all attribute information with respect to the i-th utterance word.

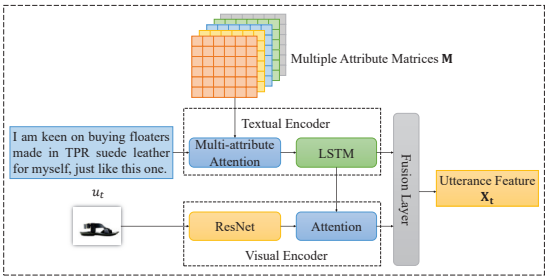

Figure 3: Illustration of our proposed node initialization module.

Thereafter, we utilize the short-cut mechanism to fuse $\mathbf{w}_i$ and $\bar{\mathbf{w}}_i$, obtaining the attribute-enhanced word representation. Finally, these powerful word representations are fed into a LSTM network, and the final hidden state is set as the textual representation for the utterance u_t, denoted as $\mathbf{h}_{u_t}$.

3.1.2 Visual Encoder. As the old saying goes, "there are a thousand Hamlets in a thousand people's eyes". Thereby, for the same product image, users may focus on different aspects. To capture useful visual information for response generation, we design a preference-aware attention network. To be specific, we select the pre-trained ResNet-18 [12] network without the final fully connected layer as the backbone of our visual encoder. It takes the image I_t from the utterance u_t as input and outputs the $R \times 512$ dimensional feature map, where R denotes the number of pixels in the feature map. Thereafter, based on the user's preference representation $\mathbf{h}_{u_t}$ obtained from the textual encoder, we calculate the alignment score between each visual region and the user's preferences as follows,

$$\begin{cases} s_i = \frac{exp(e_i)}{\sum_{j=1}^{R} exp(e_j)}, \\ e_i = f_{att}(\mathbf{h}_{u_t}, \mathbf{v}_i), \end{cases} \tag{4}$$

where $\mathbf{v}_i \in \mathbb{R}^{512}$ denotes the i-th pixel of the visual feature map and f_{att} denotes the attention network implemented by a 1-layer perception. After obtaining these preference-aware attention scores, the final visual representation $\mathbf{h}_{v_t}$ of the utterance u_t could be obtained as follows,

$$\mathbf{h}_{v_t} = \sum_{i=1}^{R} s_i \mathbf{v}_i. \tag{5}$$

3.1.3 Fusion Layer. Thus far, we have obtained the textual embedding $\mathbf{h}_{u_t}$ and the visual embedding $\mathbf{h}_{v_t}$ of the current utterance u_t. We hence can derive a cross-modal representation for the current utterance by employing the concatenation operator as follows,

$$\mathbf{X}_t = \mathbf{h}_{u_t} \oplus \mathbf{h}_{v_t}, \tag{6}$$

where $\oplus$ represents the concatenation operation and $\mathbf{X}_t$ denotes the initial representation of the current node u_t.

3.2 Relational Context Modeling Module

As illustrated in Figure 1, different utterances contain inconsistent effective information, and their ability to describe user intentions is different. For instance, the initial utterance is usually on greetings.

[3]In this paper, we consider the top 5 frequent types of attributes, including "color", "gender", "material", "style", and "type".

[4]In our work, we merely select the attribute values that appear more than 100 times. Therefore, the number of attribute values K_a corresponding to the five attributes are 31, 4, 94, 31, and 111, respectively. For the convenience of computing, we pad the number of values for each attribute to 120 with zero.

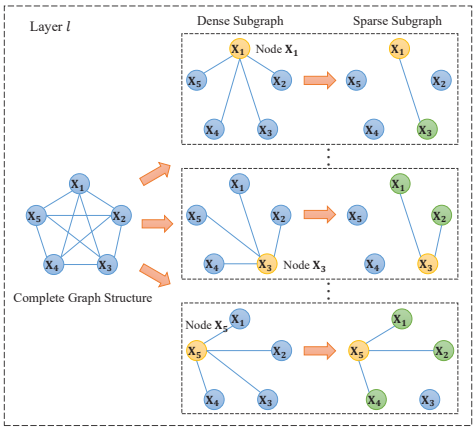

Figure 4: Schematic illustration of our sparse adjacency relationship update strategy (N=5).

Some middle utterances are transitional sentences, such as "Absolutely. Got it. Showing in a minute.", containing less information related to the user intentions. Besides, some utterances express the preferences of the user in detail, e.g., "... dark color culotte in cotton...". Therefore, in the context modeling, we should highlight the utterances with meaningful information, while restrain the meaningless ones. Moreover, the dialog context information is not entirely independent, therefore, how to adequately model the relationship between them for importance evaluation is also a concern.

To this end, we propose a sparse graph attention network by extending the graph attention network (GAT) [31]. GAT iteratively updates the representation of each node (*e.g.*, the utterance representation) by aggregating representations of the neighbor nodes via multi-head attention. Therefore, GAT could assign small even zero attention score to the irrelevant node. However, the normalization process implemented by Softmax would magnify these smaller values, which may lead to negative influence to the context encoding, further affecting the accuracy of response generation. To address this issue, we propose to extend the original GAT with sparse adjacency matrix update. We dynamically delete some connecting edges to ensure each node merely links to the closely related ones. Specifically, at the initialization, we define a adjacency matrix whose elements are all ones. In each iteration layer, the values in the adjacency matrix will be updated according to attention coefficients. To enforce each node to pay attention to the relevant nodes and avoid the interference of unimportant information, we only retain edges of the top-k neighbor nodes. More concretely, the value of k is calculated as follows,

$$k = [t/2] + 1, \tag{7}$$

where t is the index number of the t-th node, and $[\cdot]$ is the rounding function. Note that in each layer, multiple adjacency matrices obtained by different attention heads would be aggregated into one matrix through the majoritarian voting mechanism, as the adjacency matrix of the next iteration layer. In other words, at each location of the output adjacency matrix, its value is decided by a majority of values in the corresponding position of input matrices.

As shown in Figure 4, the original connection relationship between nodes is relatively dense, while our sparse update strategy largely simplifies the connection relationship, where each node is only connected to the nodes that are closely related to it. This makes it easier for each node to aggregate effective information from other nodes. In this paper, we utilize $\bar{\mathcal{N}}_t$ to represent the top-k neighbor nodes of the node t. Thereby, the update process of our sparse GAT can be formulated as follows,

$$\begin{cases} \mathbf{X}_t^{l+1} = ||_{k=1}^{K_b}(\Phi(\sum_{s\in\bar{\mathcal{N}}_t} r_{t,s}^{lk}\mathbf{W}_k^l\mathbf{X}_s^l)), \\ r_{t,s}^{lk} = \dfrac{exp(\phi(\boldsymbol{\gamma}_l^T(\mathbf{W}_k^l\mathbf{X}_t^l \oplus \mathbf{W}_k^l\mathbf{X}_s^l)))}{\sum_{m\in\bar{\mathcal{N}}_t} exp(\phi(\boldsymbol{\gamma}_l^T(\mathbf{W}_k^l\mathbf{X}_t \oplus \mathbf{W}_k^l X_m)))}, \end{cases} \tag{8}$$

where $\boldsymbol{\gamma}_l$ and $\mathbf{W}_k^l$ are respectively the trainable parameter vector and matrix of the l-th layer, ϕ and Φ are respectively the LeakyReLU and Exponential Linear Unit (ELU) activation functions, K_b represents the number of attention heads, $||_{k=1}^{K}\mathbf{x}_k$ denotes the concatenation of vectors from $\mathbf{x}_1$ to $\mathbf{x}_K$, $r_{t,s}^{lk}$ is a normalized attention coefficient computed by the k-th attention head at layer l, and $\mathbf{X}_t^{l+1}$ denotes the updated t-th node representation at layer $l+1$.

After the L layers propagation[5], we obtain the final representation for each utterance, *i.e.*, $\mathbf{X}_t^L$. To obtain the final question representation $\mathbf{c}$, we first concatenate the original feature of the last node (*i.e.*, the question utterance) $\mathbf{X}_N$ and the updated one $\mathbf{X}_N^L$, and then feed it into a fully connected layer.

3.3 Response Generator

To verify the effectiveness of our question representation, we apply it to the image response selection task and design a response generator [22]. Note that our question representation can also be utilized for text response generation, which would be demonstrated in the section 4.4.2. Considering that attribute information could depict the characteristics of products, we integrate it with the visual information to enhance the representations of the candidate products. Specifically, we treat attributes of the candidates as their textual information[6], and then feed it along with their visual information into our node initialization module, outputting the representations of candidate products. Afterwards, we calculate the cosine similarity between the question vector (*i.e.*, the representation of user intentions) and the features of the candidates. Finally, the candidates with the higher similarity scores are returned as image responses.

In this work, we use the max-margin loss function to optimize the model, which is formulated as follows,

$$loss = max(0, 1 - cos(\mathbf{c}, \mathbf{y}_{pos}) + cos(\mathbf{c}, \mathbf{y}_{neg})), \tag{9}$$

where $\mathbf{y}_{pos}$ and $\mathbf{y}_{neg}$ denote the representations of the positive and negative products, respectively, and the function $cos(\mathbf{x}, \mathbf{y})$ refers to the cosine similarity between $\mathbf{x}$ and $\mathbf{y}$.

4 EXPERIMENTS

4.1 Dataset

In this paper, we conducted experiments on the widely-used benchmark dataset MMD constructed by Saha *et al.* [25], to evaluate our proposed model with several state-of-the-art baselines. The MMD

[5] In this paper, we follow the settings in [31], *i.e.*, the last layer uses single-head attention and removes the ELU activation function.

[6] We first arrange the attribute in the alphabetical order, and then concatenate each attribute and its values sequentially, constructing the input textual information.

Table 1: Performance comparison between our proposed model and several state-of-the-art baselines on the MMD dataset in the image response selection task. The best performance is highlighted in bold.

Methods	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
MHRED	0.1623	0.1787	0.2286	0.1240	0.2582	0.2766	0.0922	0.4583	0.3315
UMD	0.3431	0.3999	0.4019	0.1982	0.4629	0.4297	0.1169	0.5492	0.4596
MAGIC	0.5446	0.6589	0.6639	0.2990	0.7127	0.6841	0.1580	0.7549	0.6979
TREASURE	**0.5987**	**0.7139**	**0.7124**	**0.3134**	**0.7485**	**0.7272**	**0.1633**	**0.7817**	**0.7387**

Table 2: Performance comparison among the variants of our proposed model in the image response selection task. The best results are highlighted in bold.

Methods	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
TREASURE	**0.5987**	**0.7139**	**0.7124**	**0.3134**	**0.7485**	**0.7272**	**0.1633**	**0.7817**	**0.7387**
w/o Attribute	0.5799	0.6918	0.6817	0.3064	0.7318	0.6989	0.1606	0.7687	0.7118
w/o Graph	0.5823	0.6944	0.6839	0.3062	0.7308	0.6998	0.1602	0.7662	0.7121
w/o Sparsity	0.5899	0.7042	0.7030	0.3115	0.7445	0.7204	0.1629	0.7800	0.7327

dataset contains more than 150k conversations between users and agents in the retail domain, where each conversation describes a complete online shopping process with approximately 40 utterances. During the conversation, the user proposes his/her intentions in multimodal utterances and the agent introduces different products step by step until they make a deal. Moreover, more than 1 million fashion products with a variety of domain knowledge are crawled from several well-known online retailing websites, such as Amazon[7], Jabong[8], and Abof[9]. Meanwhile, Saha *et al.* [25] proposed two research tasks: the text response generation and the image response selection. The former is designed to generate text responses based on the question representation, while the latter aims to retrieve and sort candidate images based on the relevance between the question representation and the product vectors. In this paper, we mainly evaluated our model in the image response selection task.

4.2 Experimental Settings

4.2.1 Implementation Details. We optimized our proposed model on 1 GeForce RTX 2080 Ti GPU using PyTorch library. The Adam optimizer [1] is employed with a mini-batch size 64 and 12 epochs. The learning rate is set as 0.0001. Moreover, the dimension of the word embedding D_w is 300, the number of visual regions R is 49, and the dimension of question representation $\mathbf{c}$ is set to 2048. Besides, the utterance number N in the context is 10, the layer number L of our sparse GAT is 2, and the number of attention heads K_b is 3. As for the ratio of positive and negative products[10], we set it to 1:4 and 5:1000 for training and testing, respectively.

4.2.2 Evaluation Metrics. Following the existing baseline [25], we adopted Recall@k, Precision@k, and NDCG@k (k= 5, 10, and 20), as the evaluation metrics in the image response selection task. To be specific, Recall@k is the proportion of relevant products found in the top-k results. Precision@k is the proportion of selected products in the top-k set that are relevant. NDCG@k is an evaluation criterion to measure the ranking results, which is the ratio of the corresponding Discounted Cumulative Gain (DCG) to Ideal Discounted Cumulative Gain (IDCG). And in the text response generation task, we utilized BLEU-m [23] (m varies from 1 to 4) and NIST [8] to measure the similarity between the predicted and target responses.

4.3 Performance Comparison

To justify the effectiveness of our proposed TREASURE model, we compared it with the following state-of-the-art baselines of releasing the codes in the image response selection task.

- `MHRED` [25] : This is the first work on a multimodal task-oriented dialog system in the retail domain. It incorporates visual features into the hierarchical recurrent encoder-decoder model [32] to form the multimodal hierarchical encoder-decoder model, achieving impressive performance.
- `UMD` [6] : It is a user attention-guided multimodal dialog system built on top of MHRED, which jointly considers hierarchical product taxonomy and the user's attention to products. In particular, it designs an attention mechanism that leverages textual features and multiple features extracted from the taxonomy-attribute tree to extract visual features.
- `MAGIC` [22] : This is currently the strongest baseline on the MMD dataset in the image response selection task. It incorporates a variety of domain knowledge and presents adaptive decoders, to dynamically generate different responses.

Note that the results of these baselines are obtained utilizing the codes provided in their original papers. For fair comparison, all baselines and our model adopt the same experimental setup (*e.g.*, the ratio of positive and negative products). The comparison results are summarized in Table 1. By analyzing the results, we found that our proposed model TREASURE outperforms the compared baselines regarding all metrics with different depths. For instance, compared with the state-of-the-art baseline MAGIC, our approach obtains relative Recall@5, Precision@5, and NDCG@5 with 9.92%, 8.34%, and 7.32% gains, respectively. Moreover, it separately achieves improvement with nearly 31.05% and 48.38% NDCG@5 gains as compared to UMD and MHRED. The improvement indicates that 1) the feasibility and importance of highlighting attribute-aware keywords in the textual sentences; and 2) the remarkable ability of our sparse GAT module. Specifically, the former could enhance the local representation of each multimodal utterance, while the latter could capture relational utterances to promote the global comprehension of user intentions. Therefore, our model could deliver more precise system responses to users.

[7] https://www.amazon.com/.
[8] https://www.jabong.com/.
[9] https://www.abof.com/.
[10] We appropriately expanded the number of negative products for each group of positive ones to 1k according to the matching relationship between product attributes and dialog retrieval requirements, increasing the difficulty of image retrieval.

Table 3: The performance of UMD-based variants in the text response generation task. The best performance is highlighted in bold.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	NIST
UMD	0.3470	0.2620	0.2108	0.1735	2.7566
UMD+TE	0.3944	0.3061	0.2504	0.2079	3.2942
UMD+TE+CE	**0.4457**	**0.3581**	**0.3034**	**0.2612**	**4.1110**

4.4 Ablation Study

4.4.1 Image Response Selection Task. We studied variants of our model to further investigate the effectiveness of the multi-attribute attention and the sparse GAT in the image response selection task:

- **w/o Attribute**: We eliminated the multi-attribute attention module from the textual encoder. That is, we directly utilized the LSTM to encode the word sequence, same as MAGIC.
- **w/o Graph**: We removed sparse GAT module from the relational context modeling module. In other words, following the same settings as other baseline methods, we directly applied LSTM to encode utterance sequence and treated the final hidden state as the question representation.
- **w/o Sparsity**: Instead of using sparse adjacency matrix update strategy, we adopted the original GAT with fully connected node relationship.

As reported in Table 2, compared with our model, the performance of **w/o Attribute** degrades dramatically. Particularly, it drops absolutely by 1.88%, 2.21%, and 3.07% on Precision@5, Recall@5, and NDCG@5, respectively. This demonstrates the vital importance of the multi-attribute attention as it can capture crucial attribute information related to the products from the given utterances. Besides, our model achieves better results than **w/o Graph**, indicating that adaptively considering relational context information of each utterance is beneficial to strengthen the complete comprehension of the user intentions. Moreover, the performance of **w/o Sparsity** drops, reflecting that it is crucial to filter useful relational context information from dense connections to enhance the global representation of each utterance. In general, our proposed model largely exceeds all variants, verifying the effectiveness of multi-attribute attention as well as the sparse GAT.

4.4.2 Text Response Generation Task. We also verified the effectiveness of our model in the text response generation task. Particularly, we selected the UMD [6] that focuses on question understanding as our baseline and set the following variants:

- **UMD+TE**: We replaced the textual encoder in UMD with our multi-attribute attention guided textual encoder.
- **UMD+TE+CE**: We replaced the whole question understanding module of UMD with our network.

From the experimental results summarized in Table 3, we can see that **UMD+TE** surpasses the baseline UMD in all evaluation metrics. Particularly, it has achieved 19.83% and 19.50% relative gains over UMD on BLEU-4 and NIST, respectively. This fully indicates that our multi-attribute attention based textual encoder could enhance the comprehension of user's questions, therefore improving the generation accuracy of text responses. In addition, **UMD+TE+CE** achieves superior performance with competitive results to **UMD+TE**. This reflects that adaptively capturing pivotal relational context information through sparse GAT for each node

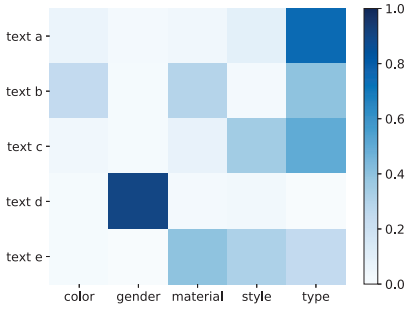

Figure 5: Visualization of the aggregation attention in Eqn. (3). The attribute attention is presented with different colors, and the darker color states the higher value.

is conducive to understanding global user intentions and generating the correct text responses. The results reported in Table 2 and Table 3 adequately demonstrate the effectiveness of our question understanding model in both the image response selection and the text response generation tasks.

4.5 Attention Visualization

4.5.1 Visualization of Attribute Attention. Apart from achieving the superior performance, one of the key advantages of TREASURE over other methods is that its multi-attribute attention module is able to distinguish the most relevant attributes to the products. Towards this end, we illustrated five textual sentences describing different product intentions of users, *i.e.*, from Text a to Text e, and then visualized their attention values over five attributes[11].

- Text a : "I would love to see **jeans** that would suit me."
- Text b : "I intend to see a few of your **dark** color **culotte** in **cotton** that she would like, just like this one."
- Text c : "I am here to see some **espadrill** with a **casual** fit that would suit me."
- Text d : "**Male**."
- Text e : "Can you show me some **business** type soft material **driving-shoes** containing sole made out of **leather** material that my buddy would like?"

From the attention results shown in Figure 5, we found that our model could adaptively aggregate crucial attribute information to enhance the utterance representation. For example, in the Text c, the bold words "espadrill" and "casual" respectively indicate the user's "type" and "style" intentions on the target product. Intuitively, to enhance the representation of this utterance, we should aggregate the information related to "type" and "style" attributes. In the third row of Figure 5, the words "style" and "type" are marked in the darkest blue, reflecting that these attributes attract the most attention. These findings are consistent with our expectation, demonstrating that our proposed module is capable of adaptively identifying the useful attributes, hence, further verifying the effectiveness of our multi-attribute attention module.

4.5.2 Visualization of Context Attention. To gain deeper insights into our sparse GAT module, in this section, we visualized the adjacent context of the N-th node (*i.e.*, the final question utterance in the multimodal context), as demonstrated in Figure 6.

[11] Due to the large number of attribute values, we only show the attention coefficient of each attribute.

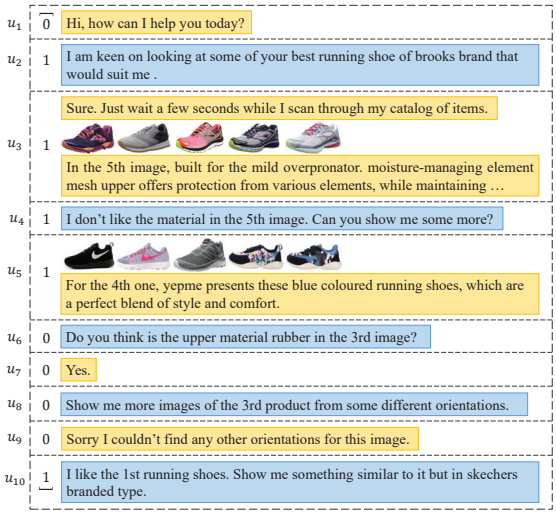

Figure 6: Relational context visualization of the last utterance node. The first column represents the order of each utterance in the dialog. The second shows the sparse adjacency relationship between the last utterance and other ones, where 1 refers to the relevant relationship. The third displays the corresponding content of each utterance, where the blue and yellow boxes denote the responses of the user and the agent, respectively.

From Figure 6, we found that the relational score of u_1 is 0. It makes sense since the greeting sentence from the agent is useless in answering the user's question (*i.e.*, the last utterance). From u_2 to u_5, their relational values are all 1. This is because the user gradually puts forward her/his intentions for the target product in these utterances. In other words, these utterances involve valuable information for question understanding. Although some textual sentences in these utterance may be useless, such as "Sure. Just wait a few seconds while I scan through my catalog of items.", their visual information is vital in response generation. For the utterance u_6 to u_9, their relational scores are 0, reflecting they are meaningless to the final response generation. This is consistent with our expectation, since the user and the agent have a detailed discussion about the third product in these utterances, rather than the first product referred in the question utterance. By analyzing the results shown in Figure 6, we can see that merely part of the utterance context is effective for understanding the user's question. Therefore, simply encoding the whole utterance context holistically may be not appropriate. This further justifies the necessity and effectiveness of our proposed sparse GAT.

4.6 Qualitative Analysis

To qualitatively validate the effectiveness of our TREASURE model, we displayed two typical cases in Figure 7. In addition, we also displayed the results of the two best baselines. Based on these retrieval results, we could see that our model could comprehend the user intentions accurately. Specially, in Figure 7(a), compared to UMD and MAGIC, TREASURE can deliver the most correct images. More importantly, the top-10 results of our model are all related to the "shorts". This demonstrates the effectiveness of our multi-attribute attention module, which effectively enforces the

(a) Case one

(b) Case two

Figure 7: Top-10 image response selection results of our TREASURE and the baselines.

model to focus on the product attribute information. Moreover, in Figure 7(b), our TREASURE not only selects the correct images but also sorts them at the top positions. This reflects our model could capture the effective information from the utterance context to better understand user intentions for image response selection.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a relational graph-based context-aware question understanding scheme for multimodal dialog system, referred to TREASURE. To be specific, to obtain the global representation for each utterance, we regard each utterance as a node in the graph network to construct a relational graph, adaptively capturing the crucial relational context information for each node. At the same time, in order to simplify the complex multimodal context relationships, we design a sparse adjacency relationship learning method to make feature propagation between nodes more accurate and faster. Besides, to enhance the local representation of each node in the relational graph for better initialization, we propose a multi-attribute attention mechanism to highlight the product-related keywords in the textual information. Extensive experiments show that our proposed TREASURE model is superior to existing methods, demonstrating the effectiveness of our framework.

In the future, we will extend our work in two directions. First, we will continue to work on complex relational context modeling, aiming to seek a more efficient way to understand user intentions. Second, we will explore the application of external knowledge, especially the cross-modal knowledge.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:62006142, No.:61872270; the Shandong Provincial Key Research and Development Program, No.:2019JZZY010118; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Kuaishou; the Young creative team in universities of Shandong Province, No.:2020KJN012.

REFERENCES

[1] Diederik P. Kingma andJimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*. 1–15.

[2] Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Ordinal and attribute aware response generation in a multimodal dialogue system. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 5437–5447.

[3] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[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 Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. 1803–1813.

[5] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 225–234.

[6] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User attention-guided multimodal dialog systems. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 445–454.

[7] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the Annual Meeting of the Association for Computational*. 484–495.

[8] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the International Conference on Human Language Technology Research*. 138–145.

[9] Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7764–7771.

[10] Zan Gao, Yinming Li, Weili Guan, Weizhi Nie, Zhiyong Cheng, and Anan Liu. 2020. Pairwise view weighted graph network for view-based 3D model retrieval. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 129–138.

[11] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 583–592.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai, and Jing Yuan. 2020. Multimodal dialogue systems via capturing context-aware dependencies of semantic elements. In *Proceedings of the ACM International Conference on Multimedia*. 2755–2764.

[14] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information System* 38, 3 (2020), 1–33.

[15] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1192–1202.

[16] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the International Joint Conference on Natural Language Processing*. 733–743.

[17] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the ACM International Conference on Multimedia*. 1571–1579.

[18] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the ACM International Conference on Multimedia*. 801–809.

[19] Jianfeng Liu, Feiyang Pan, and Ling Luo. 2020. GoChat: Goal-oriented chatbots with hierarchical reinforcement learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1793–1796.

[20] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[21] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the ACM International Conference on Multimedia*. 843–851.

[22] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the ACM International Conference on Multimedia*. 1098–1106.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 311–318.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[25] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 696–704.

[26] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Natural Language Processing*. 1577–1586.

[27] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural compatibility modeling for clothing matching. In *Proceedings of the ACM International Conference on Multimedia*. 753–761.

[28] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 235–244.

[29] Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2020. Improving matching models with hierarchical contextualized representations for multi-turn response selection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1865–1868.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[31] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*. 1–12.

[32] Iulian Vlad 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 AAAI Conference on Artifcial Intelligence*. 3776–3784.

[33] Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the International Conference on Computational Linguistics*. 4100–4110.

[34] Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat more: Deepening and widening the chatting topic via a deep model. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 255–264.

[35] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29, 1 (2019), 1–14.

[36] 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 ACM International Conference on Multimedia*. 1437–1445.

[37] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. 438–449.

[38] Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *International Conference on Learning Representations*. 1–15.

[39] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 496–505.

[40] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artifcial Intelligence*. 3351–3357.

[41] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 55–64.

[42] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 685–694.

[43] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*. 245–254.

[44] Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9604–9611.