Reflecting on Experiences for Response Generation

Chenchen Ye
National University of Singapore
chenchenye.ccye@gmail.com

Lizi Liao*
Singapore Management University
lzliao@smu.edu.sg

Suyu Liu
Singapore Management University
suyuliu.2022@phdcs.smu.edu.sg

Tat-Seng Chua
Sea-NExT Joint Lab, National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

Multimodal dialogue systems attract much attention recently, but they are far from skills like: 1) automatically generate context-specific responses instead of safe but general responses; 2) naturally coordinate between the different information modalities (*e.g.* text and image) in responses; 3) intuitively explain the reasons for generated responses and improve a specific response without re-training the whole model. To approach these goals, we propose a different angle for the task — Reflecting Experiences for Response Generation (RERG). This is supported by the fact that generating a response from scratch can be hard, but much easier if we can access other similar dialogue contexts and the corresponding responses. In particular, RERG first uses a multimodal contrastive learning enhanced retrieval model for soliciting similar dialogue instances. It then employs a cross copy based reuse model to explore the current dialogue context (*vertical*) and similar dialogue instances' responses (*horizontal*) for response generation simultaneously. Experimental results demonstrate that our model outperforms other state-of-the-art models on both automatic metrics and human evaluation. Moreover, RERG naturally provides supporting dialogue instances for better explainability. It also has a strong capability in adapting to unseen settings by simply adding related samples to the retrieval datastore without re-training the whole model.

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**.

KEYWORDS

case-based reasoning, response generation, contrastive learning

ACM Reference Format:
Chenchen Ye, Lizi Liao, Suyu Liu, and Tat-Seng Chua. 2022. Reflecting on Experiences for Response Generation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548305

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548305

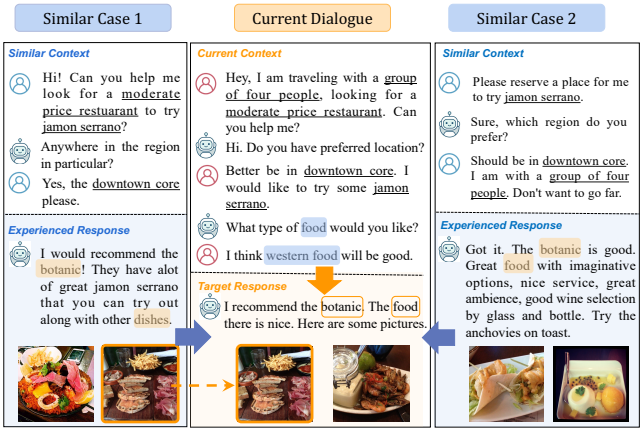

Figure 1: Example dialogue for venue recommendation, where two similar cases provide useful guidance.

1 INTRODUCTION

Multimodal dialogue system receives great attention in both academia and industry due to its growing application in reality. Generating fluent and informative natural responses is the ultimate goal of such systems. The task typically involves two sub-problems: 1) learning fixed-sized representations for the multimodal dialogue context, and 2) mapping the representations to the responses in various forms. Many methods have been developed to address these two sub-problems separately or in an end-to-end fashion. For example, Zhang et al. [45] presented a relational graph-based context-aware question understanding scheme to achieve global user intention comprehension. Nie et al. [28] designed adaptive decoders to generate the desired responses, and Liao et al. [24] tried to integrate domain knowledge for more intelligent response generation.

Although these existing multimodal dialogue systems have shown promising performance, they still suffer from the following issues: 1) Most existing models emphasize the context-response mapping that maximizes $P(R|C)$ over the whole training corpus, where C is the given dialogue context and R is the ground-truth response, or $P(R|C,K)$ when some external knowledge K is incorporated. It tends to assign high probabilities to safe but universal responses [21]. Especially for task-oriented multimodal dialogues, such general responses would possibly fail to fulfill the user's specific requirements, leaving the user unsatisfied; 2) The responses in different modalities are often treated separately (*e.g.* CNN for image response ranking and RNN for textual response generation). This makes the coordination between these response components a hard task, which may lead to unnatural responses; 3) The current

popular end-to-end modeling scheme hinders the explainability of generated responses. If we hope to improve the responses for a certain fraction of dialogue situations, re-training the whole model is usually required, and catastrophic forgetting will be an issue.

To address the aforementioned challenges, we tackle the response generation task from a different angle — explicitly making use of similar experiences. This roots from the observation that humans often solve a new problem by recollecting and adapting the solution to multiple related problems that they encountered in the past [33, 35]. This model of reasoning has been applied by case-based reasoning (CBR) [34] in classical artificial intelligence [20, 32]. A sketch of a CBR system typically consists of (a) a retrieval module which retrieves similar cases to the given problem and (b) a reuse module where the solutions of the retrieved cases are reused to synthesize a new solution. When the new solution cannot be used directly, there might be a revise module to do some revision. Intuitively, such a system has the potential to mitigate the challenges we are facing in multimodal dialogues, because the similar cases would allow us to refine the existing training corpus level analysis while zoom into similar dialogue sessions for more targeted and detailed analysis. For example, as shown in Figure 1, the similar cases provides rather specific response samples with natural modality coherence. These cases also explain where the final response came from and the response can be further improved by adding better cases.

However, the components of CBR are typically implemented with symbolic systems in the early days [1], which has largely limited its applications. For instance, finding similar dialogue contexts and synthesizing new responses will be a challenging task for a CBR system with symbolic components. Fortunately, the recent advancements in representation learning and various neural models have shed light on the possibility of applying CBR with neural components, which can largely boost the generality and applicability of such reasoning scheme in practical tasks. For instance, very recently, CBR has been successfully applied in KB reasoning [7, 8], natural language modeling [18] and machine translation [17] *etc.* Nonetheless, these approaches do not handle complex dialogue queries and only operate on structured triple queries or pure textual sequences.

In this work, we propose a neural components based CBR framework to Reflect on Experiences for Response Generation (**RERG**). The key lies in (1) an effective *retrieval* module that learns discriminative representations for multimodal dialogue contexts and selects similar experiences or cases for a given context accurately; and (2) an adaptable *reuse* module that abstracts the common characteristics over the retrieved dialogue sessions, and incorporates them into the new response generation based on current dialogue context. Specifically, we adapt unsupervised contrastive learning on both text and image part of the dialogue context to learn better intra-modality representations. Then, for each dialogue context with its positive and negative similar cases, we use triplet ranking loss to enforce the retrieval model learn better context representations and inter-modality relations. With well-selected similar cases, the reuse model predicts responses by automatically copying segments from its context vertically and copying segments from similar case responses horizontally, while selects image response accordingly. This allows the contentful patterns in previous contexts and other appropriate responses to be easily leveraged explicitly, rather than being memorized implicitly in latent representations or neural network parameters.

Moreover, the retrieve and reuse nature of the RERG framework enhances the explainability of generated responses via these retrieved cases. It also improves the generalizability of RERG in adapting to unseen dialogue situations simply by adding related samples to the retrieval datastore, while most of the current neural models cannot handle such cases without a time-consuming re-training or finetuning process. When dealing with new situations after the similar cases addition, RERG is able to adapt their responses to compose a targeted solution. Without the need of re-training neural parameters, it not only offers the fast adaptability but also overcomes the obstacle of catastrophic forgetting [19] which is commonly faced by other re-trained models.

To sum up, the contributions of this work are as follows:

- We propose to tackle the response generation task by reflecting on experiences. Instead of generating from scratch, the proposed RERG method recollects and adapts the responses from similar dialogue situations.
- We propose a multimodal contrastive learning enhanced neural retrieval model for selecting similar dialogue cases, and design a cross copy based reuse model to leverage contentful response patterns in both vertical and horizontal directions.
- Extensive experiments show that the proposed RERG method significantly outperforms several state-of-the-art models both on automatic evaluation metrics and human evaluation. We also demonstrate the explainability and good generalizability of RERG to unseen situations by experiments.

2 RELATED WORK

2.1 Task-oriented Response Generation

Task-oriented dialogue system has been largely explored in pure-text settings. Traditionally, it is constructed in a pipe-lined structure with four separated modules [11, 27, 47]. More recently, the end-to-end neural network has been largely employed to alleviate the information loss problem in pipe-lined systems [41, 48**?**]. One line of work benefits from the fast-learning capability and generalizability of large pre-trained language models, such as BERT [9] or GPT-2 [31]. Subsuming different modules into a single language model, these methods concatenate dialogue context, intermediate results, and response into a long sequence [15, 30, 43]. However, these methods oversimplify the task into a unidirectional language modeling task, where the likelihood of the generated word sequence is maximized over the whole training corpus, leading to safe but general responses. Another line of work adopts a variational autoencoder structure that decouples the direct context-response mapping by introducing latent variables in the middle [40, 49]. Although enhanced performance in fulfilling user requests has been achieved by these methods, the comprehensibility of the generated responses is often corrupted due to its single optimization goal towards task completion.

Under such background, multimodal dialogue systems further incorporate other information modalities such as image to make communications more vivid and attract much attention. The response generation task of it generally follows a similar end-to-end

modeling tradition — learning better multimodal context representations and better mapping the representations to responses. In [24, 39], they incorporated external domain knowledge and sent these combined high-level representations to generate more desirable responses in fashion domain. Firdaus et al. [10] utilized graph convolutional networks to extract information from dependency parsing tree of text and combined this with image encoder to generate responses. There are also works that introducing hierarchical attention mechanism or adaptive decoding scheme to improve the representation learning and context-to-response mapping [6, 45]. However, these methods suffer from the problems of generating safe answers, weak coordination between response modalities and lack of explainability as well as generalizability. In our work, instead of directly optimizing the mapping from context representation to responses, we first look for some similar cases and then leverage these experiences to help learning better mappings.

2.2 Case-based Reasoning

Reflecting experiences for response generation is naturally linked to the case-based reasoning (CBR), in which a reasoner remembers previous situations similar to the current one and uses that to solve the new problem. It is a recent approach to problem solving and learning that has got a lot of attention over the last few years [1]. In its early days, CBR typically consists of components with symbolic systems, which largely confined its apllications.

Recently, the development of deep learning models enables wider applicability of CBR. For example, it has achieved promising results in Knowledge Base (KB) related tasks. The model in [18] retrieved similar entities and utilized the reasoning path from them for KB completion. Das et al. [8] proposed to form a new structured logical KB query for the given natural language question by reusing similar logical queries from similar questions. However, the former one only performs reasoning over structured triplet queries and generates simple path patterns. Although the latter handles natural language questions, our model deals with more complex dialogue situations and constructs human-like informative responses instead of purely structured logical forms.

Most CBR applications in other NLP tasks are based on the k-nearest neighbors (kNN) model to retrieve relevant training examples at the test stage. For instance, Khandelwal et al. [18] utilized k nearest explicit training examples to extend the pre-trained language model and obtained improved performance. In the field of machine translation, Khandelwal et al. [17] applied kNN to retrieve similar cases at the token level for better next word prediction. Zheng et al. [51] instead proposed an adaptive kNN that dynamically determines the number of referenced neighbors. Different from these works that directly reuse the chosen retrieved results, the final responses of our model incorporate both knowledge from the current case context and the retrieved similar cases' solutions. Also, compared to the simple kNN retrieval model in these works, the multimodal contrastive learning and triplet ranking loss in our model help to gather more accurate experiences. Our model also shares some similarities with the prototype editing paradigm in [42].However, they only retrieve one single prototype, and thus the performance strongly depends on the relevance and quality of it. Our model instead leverages multiple relevant instances.

2.3 Contrastive Learning

Learning distinctive representations for multimodal dialogue context is one of the key problems for response generation. Contrastive learning aims to learn effective representations by pulling semantically close neighbors together and pushing apart non-neighbors [13], which naturally fits our goal. In recent years, many prominent approaches have come into play and drawn much attention, such as MoCo [14], SimCLR [4], SimCSE [12] *etc.* In the dialogue generation area, there also has emerged a trend of applying contrastive learning. For instances, in [3], the authors made contrastive pairs between contextual sentences and responses for learning high-quality sentence embeddings from dialogue turns. To capture and summarize various topic information from dialogue turns, Liu et al. [26] designed two contrastive objectives which consist of coherence detection and summary examination.

Different from the above efforts, instead of just focusing on the uni-modal information, we utilize contrastive learning to learn better representations for multimodal dialogue contexts. Although there are works that targeted at cross-modal contrastive learning [22, 52] or multimodal contrastive training [44], these efforts tend to emphasize on the intra-modality and inter-modality similarity at the same time. In our work, since the text and image contents are less coupled than these image captioning datasets, we use unsupervised contrastive learning to ensure intra-modality similarity while adopt triplet ranking loss to ensure dialogue context level similarity between cases.

3 THE RERG METHOD

In this section, we formally introduce our proposed RERG approach as illustrated in Figure 2. Following the typical CBR sketch, it consists of a neural retrieval model for selecting similar cases and a neural reuse model for generating the final multimodal responses. In our multimodal dialogue response generation setting, a case is a multimodal dialogue context paired with its corresponding response. For example, Figure 1 shows a specific dialogue setting with two similar cases.

In what follows, we will first give a formal definition of the multimodal response generation task and the setting of reflecting experiences. Then, the retrieval and the reuse models will be introduced subsequently. Generally speaking, the retrieval model first leverages unsupervised contrastive learning to harvest effective representations for both image and text parts of dialogue contexts. It then adopts a triplet ranking loss to enforce similarity relations among the dialogue context level representations. The reuse model on the other hand, generates high-quality responses by dynamically integrating information from both the retrieved similar cases' responses and the dialogue context.

3.1 Formulation

We first denote the multimodal dialogue training dataset as $D = \{(C_1, R_1), (C_2, R_2), \cdots, (C_N, R_N)\}$, which comprises N multimodal dialogue context-response pairs. Specifically, C_i represents dialogue context while R_i denotes its corresponding response. Since the context and response may contain both textual part and image part, we denote the corresponding text components as C_i^S, R_i^S, and the image components as C_i^I, R_i^I in context and response respectively. In

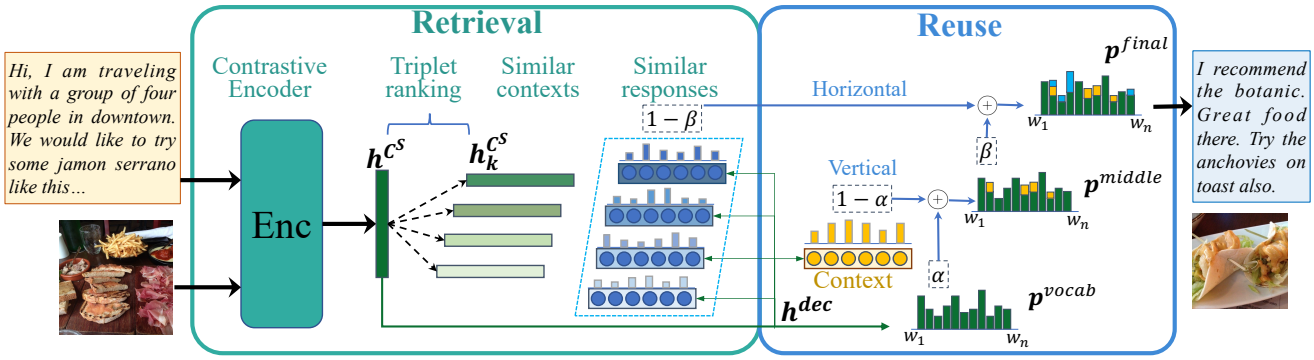

Figure 2: Architecture of the proposed RERG model. The retrieval model learns distinctive representations for dialogue context via multimodal contrastive learning and triplet ranking loss. With retrieved similar cases, the reuse model integrates contentful patterns from previous context vertically and other responses horizontally.

detail, the textual context part consists of belief state and three turns of utterances. We concatenate these into a long textual sequence to form C_i^S. For image part, since not every dialogue turn involves image, we use the most recent image within the three turns as C_i^I. Preliminary experiments show that this helps to yield coherent contexts.

Hence, the response generation task can be formulated as: given a set of training instances D, we aim to learn a model that can predict response R' for given dialogue context C where R' should be close to the target R. In the reflecting experiences setting, instead of predicting R' from scratch, the proposed RERG model first retrieves K similar cases D_K from D (Subsection 4.5.2). It then predicts the response R' by learning to reuse segments from responses of the retrieved cases (Subsection 3.3). We evaluate the model by calculating the difference between the generated response R' and the target response R while also check the task completion metrics.

3.2 Retrieval

3.2.1 Intra-modality Contrastive Learning. The green box (left) in Figure 2 illustrates the retrieval model. It first encodes the textual and visual part of the dialogue context via self-supervised representation learning. We give more details as follows.

Textual Contrastive Learning. For the textual dialogue context part, we apply dropout masks in a way similar to SimCSE [12] to conduct self-supervised representation learning. Suppose there is a collection of textual contexts $\{C_i^S\}_{i=1}^N$, we design a text encoder f_s to encode each specific context C_i^S as:

$$s_i^{z_i} = f_s(C_i^S, z_i; \theta_s)$$

where z_i is a random mask for dropout. The key for contrastive learning here is to feed the same input context C_i^S to the BERT encoder f_s twice and get two embeddings as a positive pair with different dropout masks z_i and z_i'. Hence, the training objective inside a minibatch becomes:

$$L_{textual} = -log\frac{exp(s_i^{z_i} \cdot s_i^{z_i'}/\tau)}{\sum_{j=0}^{N'} exp(s_i^{z_i} \cdot s_j^{z_j'}/\tau)}, \quad (1)$$

where τ is a temperature parameter and N' is the minibatch size.

Visual Contrastive Learning. Following MoCo-v2 [5] for visual representation learning, we denote the image encoder as $f_q(\cdot;\theta_q)$ and momentum image encoder as $f_k(\cdot;\theta_k)$, where θ_q and θ_k are the network parameters. The weights θ_k are updated with momentum coefficient m: $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$. Suppose $C_i^{I'}$ and $C_i^{I''}$ are augmented examples for the same input image I_i in a minibatch, the image encoder and momentum encoder will embed them to *query* and *key* feature vectors:

$$q_i = f_q(C_i^{I'};\theta_q)$$
$$k_i = f_k(C_i^{I''};\theta_k).$$

We maintain a dynamic set of *key* feature vectors with length M by iterative dequeue and enqueue operations. For a *query* feature vector q_i in the current batch, if k_i in the queue is originated from the same image, we denoted it as k_i^+ to form a positive pair with q_i. Hence, the visual self-supervised contrastive loss is defined as :

$$L_{visual} = -log\frac{exp(q_i \cdot k_i^+/\tau)}{\sum_{j=0}^{M} exp(q_i \cdot k_i^j/\tau)}, \quad (2)$$

where all *key* feature vectors in the dynamic queue are considered.

3.2.2 Triplet Ranking. The intra-modal training scheme captures the intrinsic patterns of augmented text and image examples. However, self-supervised methods lack of the ability to learn semantic information from higher-level similarity among multimodal dialogue contexts. We address such limitation by using triplet ranking loss to enforce similarity relations between dialogue contexts. We first embed context c_i as $c_i = f_{MLP}([s_i;q_i])$ where $[\cdot;\cdot]$ is the concatenation operation and f_{MLP} is a MLP network. Suppose c_i^+ is the positive similar case to c_i and c_i^- is the batch-hardest case by online triplet mining [36], we train the whole representation learning network via triplet ranking loss with margin ϵ as:

$$L_{triplet} = max(0, \epsilon - sim(c_i, c_i^+) + sim(c_i, c_i^-)). \quad (3)$$

Finally, the retrieval model is trained via the total loss as below:

$$L_{retrival} = L_{textual} + \lambda_1 \cdot L_{visual} + \lambda_2 \cdot L_{triplet}, \quad (4)$$

where λ_1, λ_2 are coefficients. We use the trained model to extract features for contexts and calculate dot products between them for similarity ranking.

3.3 Reuse

The essence of reuse model is to learn a mapping $(D_K, C) \rightarrow R$ for each training instance. It contains reusing these resources for both text response generation and image response generation.

3.3.1 Reuse for Text Response. Inspired from [16], the text response generation model generates response from context C^S via encoder-decoder network. At each decoding step, it learns to attend to the dialogue context C^S vertically and the similar cases' responses $R_1^S, R_2^S, \cdots, R_K^S$ in D_K horizontally to select out useful information. It is illustrated in the blue box (right) in Figure 2.

We first introduce the encoder part. We use bi-directional gated recurrent units (GRU) to encode the dialogue contexts. For example, we feed the current dialogue context C^S into the encoder and get the encoded representation $H^{C^S} = [v_1, \cdots, v_{|C^S|}^{C^S}] \in \mathcal{R}^{|C^S| \times d_v}$, where d_v is the hidden size and $|C^S|$ denotes the token numbers in context. Moreover, we calculate a summarized vector representation h^{C^S} for context by token level attention:

$$h^{C^S} = \sum_{i=0}^{|C^S|} a_i \cdot v_i,$$

$$a_i = \frac{exp(tanh(W_1 \cdot v_i) \cdot v_i)}{\sum_{j=0}^{|C^S|} exp(tanh(W_1 \cdot v_j) \cdot v_j)},$$

where W_1 is a learnable weight matrix. Similarly, for each dialogue context of the retrieved cases in D_K, we obtain the encoded context representations $h^{C_1^S}, \cdots, h^{C_K^S}$ in sequence level. When feed the similar cases' responses into the encoder, we also get the encoded response representations $H^{R_1^S}, \cdots, H^{R_K^S}$ in token level.

We also use GRU as our decoder. At the decoding step t, the decoder GRU takes a token embedding w_{t-1} as its input and returns a hidden state h_t^{dec}. It first maps the hidden state h_t^{dec} into the vocabulary space using the trainable embedding $E \in \mathcal{R}^{|V| \times d_v}$:

$$P_t^{vocab} = softmax(E \cdot (h_t^{dec})^T) \in \mathcal{R}^{|V|},$$

where $|V|$ is the vocabulary size. This is the vallina decoding probability in a typical encoder-decoder network

Vertical Copy. The decoder will also attend to the current dialogue context C^S to pick up some useful information. The attention weight over the context tokens is calculated as in [2]:

$$P_t^{vertical} = softmax(v_C^T \cdot tanh(W_{enc1} \cdot H^{C^S} + W_{dec1} \cdot h_t^{dec})),$$

where v_C, W_{enc1} and W_{dec1} are learnable parameters. Hence, the output distribution of tokens is weighted sum of the two distributions:

$$P_t^{middle} = \alpha \times P_t^{vocab} + (1 - \alpha) \times P_t^{vertical},$$

$$\alpha = sigmoid(W_2 \cdot [h_t^{dec}; w_{t-1}; P_t^{vertical} \cdot H^{C^S}]),$$

where W_2 is a trainable weight matrix, and α is trainable to combine the two distributions together. It determines whether to i) predict a token from context mapping or ii) copy a token from the current dialogue context.

Horizontal Copy. Since responses often contain some rare phrases such as locations, venue names, or phone numbers *etc.* which are hard for models to memorize in network parameters, we thus allow the decoder to attend to similar responses for copying.

First of all, we use context similarity to calculate the weight for each reference response R_k^S in D_K:

$$\gamma_k = \frac{exp(h^{C^S} \cdot h_k^{C^S})}{\sum_{j=0}^{K} exp(h^{C^S} \cdot h_j^{C^S})}.$$

Then, for each reference response R_k^S, the attention weight over each response token is calculated as

$$P_{t_k}^{horizontal} = softmax(v_R^T \cdot tanh(W_{enc2} \cdot H^{R_k^S} + W_{dec2} \cdot h_t^{dec})),$$

where v_R, W_{enc2} and W_{dec2} are also learnable parameters. Then, each output distribution of tokens after copying from a reference response is a weighted sum of it with former integrated distribution, and they are weightedly summed up to final output distribution of tokens as:

$$P_t^{final} = \sum_{k=0}^{K} \gamma_k \cdot (\beta_k \times P_t^{middle} + (1 - \beta_k) \times P_{t_k}^{horizontal}),$$

$$\beta_k = sigmoid(W_3 \cdot [h_t^{dec}; w_{t-1}; P_t^{horizontal} \cdot H^{R_k^S}]),$$

where W_3 is a trainable weight matrix and β integrates distributions.

Training Objective For each training instance, the objective for the textual response generation part is formed using negative log-likelihood as follows:

$$\begin{aligned} L_{text\_R} &= -\log P(R^S \mid C, D_K) \\ &= -\sum_{t=1}^{|R^S|} \log P^{final}(w_t \mid w_{1:t-1}\ C, D_K), \end{aligned} \tag{5}$$

where there are $|R^S|$ tokens in the ground truth response R^S.

3.3.2 Reuse for Image Response. Image response is also an important part to consider in multimodal dialogues. Different from the existing methods that rank images by simply considering the visual features, or jointly incorporating both the textual context and the visual features, we instead rely on the generated text response $R^{S'}$ and the images $R_1^I, \cdots, R_K^I$ used by other similar responses. The intuition behind is that the images in reference responses would carry most of the semantic information already, and the generated text response as part of the inputs would help to enhance the coordination between the textual and image response components.

In particular, the image ranking model aims to learn the mapping $(R^{S'}, [R_1^I, \cdots, R_K^I]) \rightarrow R^{I'}$ from training instances. It first encodes the generated text response $R^{S'}$ in the way similar to the above mentioned context encoder and gets the summarized representation vector $h^{R^{S'}}$. It also integrates image representations from reference images $[R_1^I, \cdots, R_K^I]$ via the former calculated attention weights γ:

$$q_{sum} = \sum_{k=0}^{K} \gamma_k \cdot q_k.$$

We use q_{sum} as the final query vector to search for image responses, and the ranking process uses cosine similarity as score. Images in appeared in training set with the highest score are provided to user along with our generated textual response.

4 EXPERIMENTS

4.1 Dataset

We conduct experiments on the public multimodal conversational search benchmark dataset MMConv [23]. It contains 5,106 dialogues that spans over five distinct domains with 39.8K utterances. The dialogues are grounded on a venue database with 1,771 venues and 113,953 associated images which are annotated with 315 unique classes. During experiments, we randomly split the data, and each turn is paired with 5 similar responses and 495 random responses.

4.2 Evaluation Metric

For textual response, we measure the fluency of the generated response using *BLEU* [29] score. *NIST* further considers how informative a particular n-gram is, while *ROUGE-L* further takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically [25]. For task completion, we use *Entity-F1* to report how are the important information pieces such as venue name, address, phone number *etc.* predicted in each turn. We also report *Match Rate* to check how accurate are the target venues predicted in the dialogue level. For image response, we judged it by the *Recall@k* metric, where k varies from 1 to 5. Since images for the same venue about the same concept usually look similar, we treat these images as the same when evaluating image responses following [23]. Furthermore, we carry out human evaluation to measure the quality of generated responses. More details are provided in Section 4.5.4.

4.3 Training Details

The training for the retrieve module mainly includes two stages: an intra-modality level pretraining for texts and images separately, and a case-level ranking that incorporates both textual and visual information in each dialogue case. Following SimCSE [12] for text part, we set dropout rate as 0.1 and the minibatch size N' is 64. τ is set to be 0.05. The textual contrastive learning stage includes 50 epochs with an initial learning rate of 3e-05. For visual part, we follow MoCo-v2 [5] structure. The augmentation follows [5]. During training, the *key* set length is set to be 16384, and the momentum m of updating the key encoder is 0.999. The temperature τ is set to be 0.2. The image encoder is trained using an SGD optimizer with an initial learning rate of 0.03, and we feed in 128 images each batch. The retrieve module for RERG is trained on case-level. Both text and image representations are passed into MLP layers and the features are concatenated to compute the triplet ranking loss with the margin ϵ equal to 1.5. Adam optimizer is used with an initial learning rate of 3e-06 and weight decay 1e-08.

During the reuse module training, we set batch size as 32. It is optimized by Adam optimizer with an initial learning rate of 0.001 and weight decay 1e-05. Each token in contexts and responses is embedded to be a 100-dimension vector. The GRU cell size for the context encoder, response encoder and decoder is set to be 300.

4.4 Baseline Models

The proposed method is compared with two groups of baselines: pure text-based methods and multimodal methods. All these models leverage oracle dialogue states. We also denote the variations of the proposed RERG as follows: RERG$_gt$ corresponds to the proposed model using ground truth similar cases; the subscript k indicates the number of used similar cases.

- **DialoGPT** [46]: DialoGPT is based on GPT-2 [31]. It is pretrained on large-scale Reddit conversation-like comments.
- **LaRL** [50]: It is the first to map contexts to a latent action space for response decoding and apply RL optimization.
- **HDNO** [40]: It adopts the option framework [38] to model a high-level mapping from contexts to latent action variable and then a low-level mapping to word sequence.
- **MMD** [37]: It adopts the hierarchical recurrent encoder decoder network to learn the mapping from multimodal context to response in an end-to-end fashion.
- **MMConv** [23]: It is also developed based on GPT-2 [31] that subsumes different sub-tasks into a single language model.

4.5 Main Results

4.5.1 Response Generation Results. The main response generation results are shown in Table 1. On textual responses, RERG$_5$ surpasses the best performed baseline MMConv by 3.54 on Entity F1, 10.1 on Match Rate and 0.0183 on ROUGH-L. Specifically, we observe that the textual responses generated by GPT based methods, such as DialoGPT and MMConv, score higher than the other baselines, and MMConv obtains the highest BLEU and NIST, which reflects higher occurrence of each n-gram. Their performance benefits from the fast-learning ability and robustness of large pretrained models. However, RERG still outperforms these methods with large leading margin on ROUGE-L, Entity F1 and Match Rate. Instead of decoding based on a training corpus level context-response mapping, each response generated by RERG incorporate contentful segments from its corresponding contexts and multiple similar responses, leading to better sentence-level structure, as well as richer and more accurate information in the final response. On image responses, RERG also shows big performance jump over other methods. The results of MMD is inflated due to their ranking setting where each true image is paired with 300 random negative images for ranking. Still, the performance is lower.

We also test the effect of k, the number of similar cases. We first train RERG with $k = 5$ as the dataset setting. As shown in Table 1, a leading performance is achienved. Then, when directly using all five ground truth similar cases (RERG$_gt_{k=5}$), the responses generally show better performance, which is as expected. However, these two results are relatively close, which indicates that our retrieval model can be further improved but still works reasonable well. Further comparison will be provided in Subsection 4.5.2. The results also suggest that k indeed affects RERG's performance. When $k = 2$, higher venue match and image recall are achieved, while when $k = 10$, a lower performance is obtained, possibly due to the increase of noise and response variety when taking more similar responses.

4.5.2 Retrieval Results. We show the results for various retrieval model that selects similar cases in Table 2. It shows that the proposed multimodal contrastive learning based model performs the best. For example, it exceeds the pure text-based SimCSE by 4.32% on Recall@1 and 43.05% on Recall@10. We omit the retrieve performance of the pure image-based Moco-v2 since it is only applicable

Table 1: Main multimodal response generation results on MMConv dataset.

Group	Method	Textual Response					Image Response		
		BLEU	NIST	ROUGE-L	Entity-F1	Match Rate	Recall@1	Recall@3	Recall@5
Text-based	DialoGPT [46]	18.32	3.160	0.4419	18.89	24.7	–	–	–
	LaRL [50]	13.33	2.496	0.3214	5.36	1.5	–	–	–
	HDNO [40]	14.79	2.745	0.3663	8.23	2.3	–	–	–
Multimodal	MMD [37]	16.60	3.062	0.3728	11.08	5.1	4.69	8.33	11.98
	MMConv [23]	32.33	5.758	0.5402	49.01	69.2	17.85	–	–
	RERG$_{_5}$	30.75	5.616	0.5585	52.55	79.3	22.83	24.88	26.33
	RERG$_{_gt_{k=5}}$	31.17	5.529	0.5776	54.36	80.6	23.43	25.60	36.57
	RERG$_{_2}$	29.66	5.374	0.5591	51.55	81.9	33.94	35.51	36.23
	RERG$_{_10}$	27.72	5.345	0.5322	46.69	69.8	14.37	16.67	17.75

Table 2: Performance results on similar case retrieval.

Method	HR@1	HR@3	HR@5	HR@10
BERT	9.30	17.87	21.84	27.48
SimCSE	14.54	32.89	42.87	53.76
RERG$_{_retrieval}$	18.86	55.48	88.74	96.81

to a small portion of dialogue turns that contain visual information. Contrarily, RERG can gather useful information from either textual or visual or both to support the retrieval process. Note that both SimCSE and MoCo-V2 apply self-supervised representation learning which has been demonstrated by various methods that it can help networks to learn more effective representations. For instance, SimCSE outperforms the vallina BERT as shown in Table 2. Therefore, the better performance of RERG$_{_retrieval}$ shows that our design manages to combine the strength of the two, hence resulting in better performance.

Table 3: Results for different reuse strategies (k=2).

Copy Strategy	BLEU	NIST	ROUGE-L	Entity-F1	Match
No Copy	16.02	2.976	0.3864	8.60	2.1
Vertical Only	16.74	3.142	0.3988	11.31	3.9
Horizontal Only	28.72	5.266	0.5268	48.66	78.4
Cross Copy	29.66	5.374	0.5591	51.55	81.9

4.5.3 Effect of Different Reuse Strategies. To further investigate the effectiveness of the reuse strategy, on top of a pure encoder-decoder model, we fix the retrieve module and train the reuse module with different copy strategies. The results in Table 3 show that the horizontal copy from similar cases largely benefits the quality of the generated responses in providing not only useful utterance patterns reflected by the increased BLEU, NIST, ROUGH-L scores, but also correct entities and requested venues which lead to a large improvement on Entity F1 and Match rate. This greatly aids the system's ability to respond to user-specific demands in dialogues. Moreover, combining both vertical and horizontal copy further improves the generated responses.

4.5.4 Human Evaluation. We conduct human evaluation to further investigate the quality of the generated responses. We recruit nine graduate students as participants. We test on 100 randomly sampled dialogue cases and compared the responses generated by RERG vs. MMConv. During the evaluation of each case, the participants

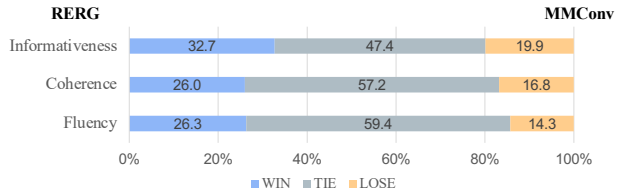

Figure 3: Human evaluation results.

are provided with all previous dialogue contexts and ground-truth responses as the reference. A fair blind evaluation is conducted with no information disclosure about the source model. The ranking is based on three criteria: (i) *fluency*: the response is grammatically correct, natural, and smooth. (ii) *coherence*: the response is coherent and follows the flow of the dialogue reasonably. (iii) *informativeness*: the response provided related information to solve the user's requests and complete the task.

After gathering the replies, we visualize the calculated statistics in Figure 3. The proportions of RERG outperforms, ties with, and loses to MMConv under each criterion are represented by the "Win", "Tie", and "Lose" portions in the stacked bars. We can observe that the generated responses by RERG outperform MMConv in all three aspects, indicating its strong capability in correctly inferring responses to fulfill user requests and generating human-like responses. Moreover, in the criteria of 'informativeness', the RERG model surpasses MMConv by a large margin, which further shows the necessity of the cross copy mechanism that can provide users with valuable information from other similar textual sentences.

4.6 Explainability

Most existing dialogue systems based on a corpus-level mapping from dialogue contexts to responses provide limited explanations for the generated responses. However, with the retrieved similar responses and the cross copy design in the reuse module, RERG naturally supports better explainability. For instance, we investigate a generated response in Figure 4. We first extract the probability of each predicted word at its corresponding decoding timestep. With values computed in the reuse module, we could inspect how the initial vocab distribution, vertical copy from the context, and horizontal copy from the retrieved experienced responses are contributed to the final probability. From the visualization, we could clearly observe that both vertical copy and horizontal copy contribute to the generated response. Also, it largely benefits from the retrieved responses to correctly recommend 'Pizza Express'.

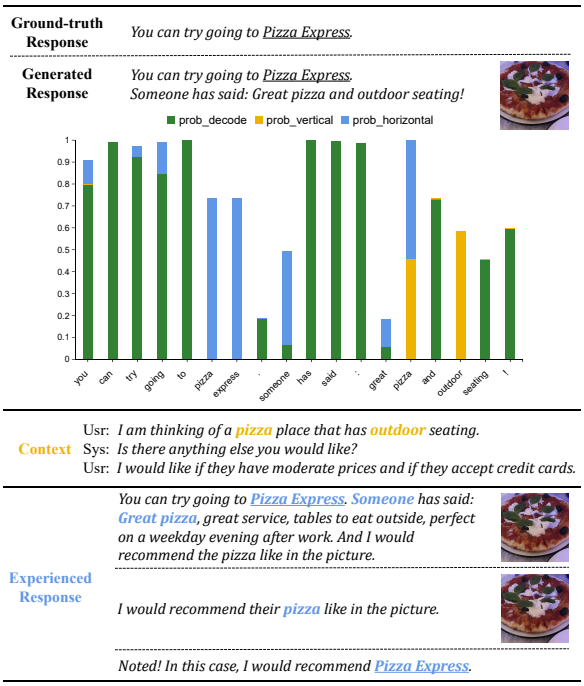

Figure 4: Example inspection of how initial vocab distribution, vertical copy, and horizontal copy contribute.

For example, the generated response does not end with a single sentence but provides extra review information. We notice that the probability of the word 'Someone' is mostly contributed by the horizontal copy. This suggests that there is an effective information support from the retrieved similar responses, especially from the first response which has a very similar replying structure and useful review contents and useful review contents. Meanwhile, the dialogue system also captures the user's specific request for 'outdoor seating' in the context, and supported by the vertical copy on the word 'outdoor'. Moreover, the system also gives a suitable pizza image with the text response, which is learned from similar cases.

4.7 Study on Unseen Situations

Most existing neural models for response generation require a time-consuming re-training or finetuning process to handle unseen situations. Such costly process may also lead to catastrophic forgetting [19]. RERG instead provides a computationally much cheaper way: add similar cases to the retrieve datastore, and let the reuse module constructs responses with new top-k responses. To verify the generalizability of the proposed RERG model, we create a controlled setup by removing all dialogues in the training set which happened under a specific goal setting. Specifically, the user plans to shop in the Jurong East area and seeks advice about the three different shopping malls there. Four dialogues in the training set are removed and four dialogues under the same goal in the testing set become the held-out set with the unseen situation.

Table 4: Entity F1 scores under unseen situations (k=2).

Method	Scenario	Remaining	Held-out
MMConv	Train on original cases	49.07	11.54
	+ Fine-tune on additional cases	44.06	69.23
	+ Fine-tune on all cases	47.39	57.69
RERG	Train on original cases	49.55	11.54
	+ Add back to retrieve datastore	49.55	65.38

As shown in Table 4, both transformer-based baseline MMConv and the proposed RERG model obtain low entity F1 scores on the held-out set without similar training dialogues. The correctly predicted turns are those that only needs very general entities, such as *inform building*. For RERG, we only added the four additional training dialogues back to the retrieve datastore without retraining the model. As shown in the last row, it achieves a significant improvement on the held-out set, indicating an effective retrieval and reuse of the added similar cases in response generation. Meanwhile, we observe that the added cases have a minor influence on the remaining test cases as they do not rank in the top-2 using the trained retrieval module. On the contrary, finetuning is necessary for MMConv to perform better in those unseen situations. We first try to finetune MMConv only on the additional training dialogues. As shown in the row 2 in Table 4, though an enhanced entity F1 score is achieved on the held-out set after finetuning, the performance on the remaining test cases is degraded. On the other hand, we find that in order to acheive reasonable performance, very specific settings need to be carefully designed for finetuning MMConv on both original and additional cases in line 3.

5 CONCLUSION

In conclusion, we proposed a neural case based reasoning framework to reflect on experiences for multimodal response generation (RERG). It roots from the fact that humans often solve a new problem by relying on the solutions of multiple related problems encountered in the past. Correspondingly, we first designed a multimodal contrastive learning enhanced retrieval model for soliciting similar dialogue instances. Given the selected similar instances, we then proposed a cross copy based reuse model. It learns to vertically copy useful segments from the current dialogue context and horizontally copy from similar dialogue instances' responses simultaneously for response generation. We carried out extensive experiments on a public large-scale dataset in comparison with a wide range of baselines. Both automatic and human evaluation are involved. The superior performance demonstrates that the proposed RERG method generates better responses. Moreover, experimental results demonstrate the explainability of the proposed RERG and the good generalizability of RERG to unseen situations.

Even though, RERG is modular by following the typical CBR sketch and has several advantages, the retrieve and reuse components of our model are trained separately. In the future, we plan to explore avenues for end-to-end learning for case based reasoning. We would also like to further improve the strategy planning part in handling dialogue situations that require consecutive turns of actions and analyze how our model performs in such situations.

ACKNOWLEDGMENTS

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and the Sea-NExT Joint Lab.

REFERENCES

[1] Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7, 1 (1994), 39–59.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

[3] Hengyi Cai, Hongshen Chen, Yonghao Song, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Xiaofang Zhao. 2020. Group-wise Contrastive Learning for Neural Dialogue Generation. In *EMNLP*. 793–802.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. 1597–1607.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[6] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User attention-guided multimodal dialog systems. In *SIGIR*. 445–454.

[7] Rajarshi Das, Ameya Godbole, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2020. A Simple Approach to Case-Based Reasoning in Knowledge Bases. In *Automated Knowledge Base Construction*.

[8] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based Reasoning for Natural Language Queries over Knowledge Bases. In *EMNLP*. 9594–9611.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[10] Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2020. MultiDM-GCN: Aspect-guided Response Generation in Multi-domain Multi-modal Dialogue System using Graph Convolutional Network. In *EMNLP*. 2318–2328.

[11] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. (2018), 2–7.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*. 6894–6910.

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, Vol. 2. 1735–1742.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[15] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796* (2020).

[16] Changzhen Ji, Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, Conghui Zhu, and Tiejun Zhao. 2020. Cross Copy Network for Dialogue Generation. In *EMNLP*. 1900–1910.

[17] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest Neighbor Machine Translation. In *ICLR*.

[18] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through Memorization: Nearest Neighbor Language Models. In *ICLR*.

[19] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[20] Janet L Kolodner. 1983. Maintaining organization in a dynamic long-term memory. *Cognitive science* 7, 4 (1983), 243–280.

[21] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*. 110–119.

[22] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *ACL*. 2592–2607.

[23] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *SIGIR*. 675–684.

[24] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *ACM Multimedia*. 801–809.

[25] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*. 605–612.

[26] Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization. *arXiv: Computation and Language* (2021).

[27] Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured Fusion Networks for Dialog. In *SIGdial*. 165–177.

[28] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *ACM Multimedia*. 1098–1106.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*. 311–318.

[30] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298* (2020).

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[32] Edwina L Rissland. 1983. Examples in Legal Reasoning: Legal Hypotheticals.. In *IJCAI*. 90–93.

[33] Brian H Ross. 1984. Remindings and their effects in learning a cognitive skill. *Cognitive psychology* 16, 3 (1984), 371–416.

[34] Roger C Schank. 1983. *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press.

[35] Henk Schmidt, Geoffrey Norman, and Henny Boshuizen. 1990. A cognitive perspective on medical expertise: theory and implications. *Academic medicine* 65, 10 (1990), 611–621.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.

[37] Agarwal Shubham, Dusek Ondrej, Konstas Ioannis, and Rieser Verena. 2018. A Knowledge-Grounded Multimodal Search-Based Conversational Agent. In *SCAI@EMNLP*.

[38] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.

[39] Deeksha Varshney and Asif Ekbal Anushkha Singh. 2021. Knowledge Grounded Multimodal Dialog Generation in Task-oriented Settings. In *PACLIC*. 425–435.

[40] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Modelling Hierarchical Structure between Dialogue Policy and Natural Language Generator with Option Framework for Task-oriented Dialogue System. In *ICLR*.

[41] Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022. State Graph Reasoning for Multimodal Conversational Recommendation. *IEEE Transactions on Multimedia* (2022).

[42] Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *AAAI*. 7281–7288.

[43] Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2. In *AAAI*. 14230–14238.

[44] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *CVPR*. 6995–7004.

[45] 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 *ACM Multimedia*. 695–703.

[46] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL*. 270–278.

[47] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural multimodal belief tracker with adaptive attention for dialogue systems. In *WWW*. 2401–2412.

[48] Zheng Zhang, Lizi Liao, Xiaoyan Zhu, Tat-Seng Chua, Zitao Liu, Yan Huang, and Minlie Huang. 2020. Learning Goal-oriented Dialogue Policy with opposite Agent Awareness. In *AACL*. 122–132.

[49] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *SIGdial*. 1–10.

[50] Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. In *ACL*. 1208–1218.

[51] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive Nearest Neighbor Machine Translation. In *ACL*. 368–374.

[52] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*. 13041–13049.