# MaTCR: Modality-Aligned Thought Chain Reasoning for Multimodal Task-Oriented Dialogue Generation


Yiting Liu
Institute of Computing Technology, Chinese Academy of Sciences & University of Chinese Academy of Sciences
liuyiting21s@ict.ac.cn

Liang Li*
Institute of Computing Technology, Chinese Academy of Sciences
liang.li@ict.ac.cn

Beichen Zhang
University of Chinese Academy of Sciences
zhangbeichen14@mails.ucas.ac.cn

Shan Huang
Tencent Research, Beijing, China
lattehuang@tencent.com

Zheng-Jun Zha
University of Science and Technology of China
Zhazj@ustc.edu.cn

Qingming Huang
University of Chinese Academy of Sciences
qmhuang@ucas.ac.cn



## ABSTRACT

In recent years, multimodal task-oriented dialogue systems have attracted increasing attention from communities, owing to their ability to naturally and efficiently provide user service. Despite the commercial value of multimodal dialogue systems, they are still confronted with two challenges: (1) capture users' intention from lengthy context and side knowledge for question comprehension; (2) jointly consider the multimodal information for response generation. In view of the challenges, previous methods designed for specific scenario lack auxiliary reasoning structures with effective modality interaction, which hinders the comprehension of user's needs and impedes the generation of desired responses. To address these issues, we propose a Modality-aligned Thought Chain Reasoning (MaTCR) framework to insert explicit reasoning process for multimodal task-oriented dialogue generation. We construct a multimodal thought chain by summarizing intermediate user queries from aligned visual and textual context, which helps to guide the comprehension of user intentions for generating reasonable responses. To effectively extract and integrate multimodal information for thought chain reasoning, we design a multimodal reasoner consisting of visual representation learning and modality-aligned fusion. We comparatively justify MaTCR with several strong baselines, including highly regarded LLM. Extensive experiments over a benchmark dataset demonstrate that MaTCR outperforms existing methods and provides stronger interpretability.



## CCS CONCEPTS

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

## KEYWORDS

multimodal task-oriented dialogue system, thought chain reasoning, visual representation learning, multimodal alignment.

**ACM Reference Format:**
Yiting Liu, Liang Li, Beichen Zhang, Shan Huang, Zheng-Jun Zha, and Qingming Huang. 2023. MaTCR: Modality-Aligned Thought Chain Reasoning for Multimodal Task-Oriented Dialogue Generation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612268



*Corresponding author.




*MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.*



ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612268


## 1 INTRODUCTION

With the advance in interactive intelligent assistants, multi-modal task-oriented dialogue systems have gained increasing interest due to their commercial value in domains such as retail and fashion. Different from the traditional dialogue agents which only focus on textual information, multimodal dialogue systems allow users to express their intentions with various modalities, thus providing a more intuitive and interactive manner to satisfy user needs.

Owing to the notable development in multimodality research [19, 23, 36, 37, 39–41, 45], many multimodal task-oriented dialogue models have been proposed in recent years [4, 10, 26, 34, 47]. Most of them are designed for the pre-sales guidance scenario of the manually constructed MMD dataset [34]. These methods mainly focus on incorporating knowledge base [20, 28], or studying users' attention to the different visual attributes of products [10, 26]. However, these methods are constrained by the single scene and data format of the MMD dataset, and lack strong reasoning structure to understand more complex user intentions from multi-source information, making it difficult to adapt to the service scenarios in the real world.

In practical customer service conversations, the requests provided by users are often ambiguous. In order to satisfy users with appropriate responses, the system needs to combine multimodal information and product knowledge to grasp the complete user expectations. As shown in Figure 1, a user faces a product problem: "machine cannot be used properly", and mentions an ambiguous concept "here" which is the root cause of the problem. To understand that the word "here" refers to the lid of the food processor, and to meet the user's expectations for solutions to product issues, it is essential for the system to complement multi-source information

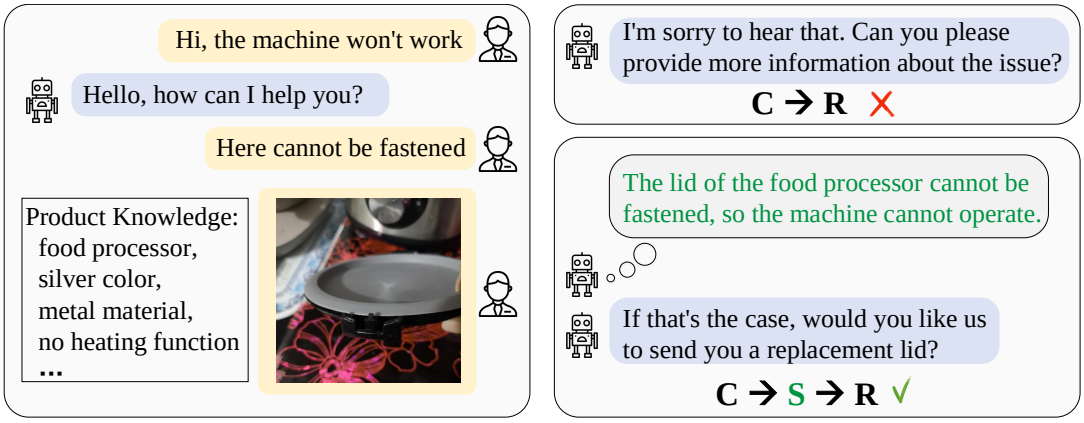


**Figure 1: An example of a multimodal task-oriented dialogue between a user and a customer service system. The green part indicates the reasoning process for the system before answering with a proper response.**

with each other. Based on this, the system provides a reasonable suggestion of "replacing the cup lid."

Another line of dialogue method based on the large language model, such as ChatGPT, shows promising performance in solving user problems. Although powerful at understanding user's questions, these methods inherit the shortcomings of large language models, making them prone to producing unrealistic outputs, and often exhibit suboptimal performance in specific domains that require expertise, since they are designed to be general. Especially when facing visual modalities like images, these methods trained with single language modality are limited to effectively integrate multimodal information, even with the conjoined vision processing block [43].

To address the above issues, we present MaTCR, a thought chain reasoning framework that inserts explicit reasoning process into multimodal dialogue generation. Our approach is inspired by the success of the chain-of-thought method (CoT) in question answering task [12, 14, 42, 51], which has been shown effective to elicit the multi-step reasoning abilities of the language models. Still, how to effectively utilize modalities from different semantic spaces remains a key challenge for the CoT method [48]. In this paper, we construct a thought chain for multimodal task-oriented dialogue generation, and design a multimodal reasoner with modality alignment to effectively leverage textual and visual dialogue context.

Specifically, we decompose the task into a thought chain with two stages: user query summarization and system response generation. At each dialogue turn, our method produces a summarized user query based on the multimodal dialogue context before generating a response. The summarized query acts as an intermediate rationale of the thought chain, which provides an auxiliary basis for the target task. The proposed multimodal reasoner aims to unify these two stages. To learn effective visual representation that can be easily accepted by the language modal, we devise a lightweight visual representation learning module. It takes a small set of visual queries to capture image features most relevant to the downstream tasks. In order to seamlessly combine information from dialogue history, knowledge, and image, we devise a modality-aligned fusion module. It aligns multimodal information using image-text-matching, then adaptively integrate them with a gate-controlled fusion layer. By doing so, we bridge the semantic gap between vision and language models, thus effectively leveraging multimodal information, resulting in more accurate user query and system response generation.

To validate the effectiveness of our approach, we conduct experiments on a benchmark multimodal task-oriented dialogue dataset that provides real customer service conversations with diverse scenarios.

The main contributions of our present work are as follows:

- We propose a MaTCR framework, which strengthens the multimodal dialogue generation by generating intermediate user queries to form the thought chain. To the best of our knowledge, this is the first work that introduces thought chain reasoning into multimodal dialogue system.
- We design a multimodal reasoner to perform thought chain reasoning. It extracts effective visual information with a visual representation learning module. Then it leverages a modality-aligned fusion module to align relevant multimodal information and adaptively integrate them.
- We conduct extensive experiments on JDDC 2.1 multimodal dialogue dataset with complex application scenarios. The experimental results verify the effectiveness of MaTCR with favorable performance and stronger interpretability[1].

## 2 RELATED WORK

### 2.1 Multimodal Task-oriented Dialogue System

The research of multimodal dialogue system can be roughly divided into two categories: open-domain conversations with casual chi-chat that involves specific images [7, 24, 35, 46, 50], and task-oriented dialog systems [13, 34, 49] which are designed to assist users in achieving specific goals.

To promote the study on multimodal task-oriented dialogue, Saha et al. [34] first constructed a multimodal task-oriented dialogue dataset MMD for the fashion domain, which consists of over

[1]We will release all the source code and model parameters soon.

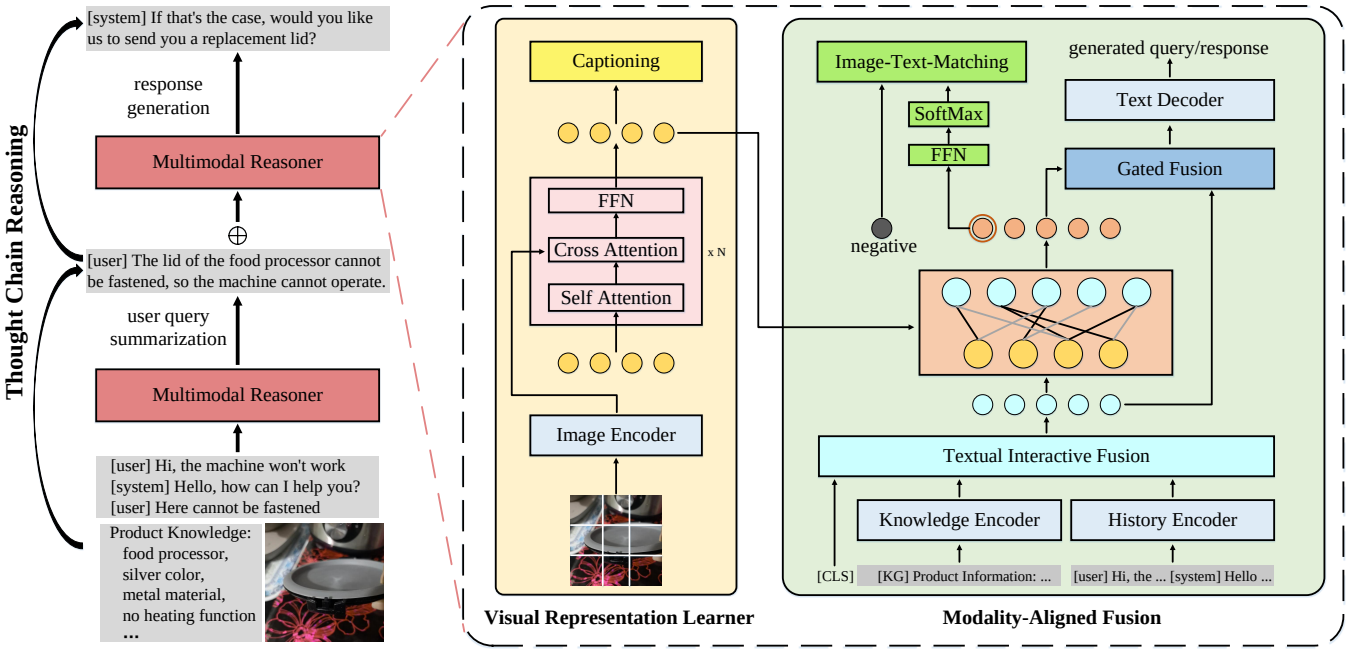


**Figure 2: The proposed MaTCR framework.**

150K conversation sessions and includes domain knowledge curation. Based on this dataset, Cui et al. [4] designed a user attention-guided multimodal dialog system by additionally considering the hierarchical product taxonomy. Later, Zhang et al. [47] devised a relational graph-based context-aware model to better understand user questions. Recently, Ma et al. [26] designed a unified representation framework that uses cross-modal alignment and key-value reasoning to generate accurate responses.

The shortcoming of MMD dataset is that the dialogue scene is limited in the pre-sales guidance, while other scenes such as payment, logistics, and after-sales maintenance are not covered. This reduces the practicality of the relevant models in real-world scenarios. The same problem exists in SIMMC dataset [13]. Therefore, Zhao et al. [49] collected the JDDC dataset that contains about 246K dialogue sessions and covers almost the complete process in E-commerce. The increased complexity of scenarios poses greater challenges and higher demands on dialogue systems to effectively utilize multimodal information for better comprehension of user intentions.

## 2.2 Chain-of-thought Method

Recently, chain-of-thought (CoT) has been widely studied to elicit the multi-step reasoning abilities of language models [12, 14, 42, 51]. This idea is inspired by the recognition that in neuro-symbolic computing, generating intermediate results significantly improves performance when a desired input-output mapping involves multiple computational steps [5, 27]. Thus, it is reasonable to recover complex input-output mappings more accurately by training on <input, intermediate results, output> triples, rather than simple <input, output> pairs [2, 3].

Following original CoT, Lampinen et al. [14] proved that explanations of context can improve the performance of large models. Kojima et al. [12] invoked zero-shot CoT reasoning by adding a prompt like "Let's think step by step" after the test question. In specific application domains. Lu et al. [25] first introduced CoT into multimodal research, by providing an image-based science QA dataset with manually annotated explanations for correct answers. Based on this work, Zhang et al. [48] proved that fine-tuning small models can also benefit from CoT, on condition that the necessary vision context is accessible to perform effective reasoning. However, the current chain of thought methods still lack mechanisms for effectively utilizing multimodal information, which limits their application.

# 3 THE PROPOSED METHOD

In this section, we first introduce the task formulation for multimodal task-oriented dialogue generation, and then illustrate the main components of the proposed MaTCR framework.

## 3.1 Task Formulation

Let $D_n = \{(Q_1, R_1), (Q_2, R_2), \ldots, (Q_n - 1, R_n - 1), Q_n\}$ denotes the textual dialogue history of the $n$-th turn, where $Q_i$ and $R_i$ are the user utterance and system response in the $i$-th turn respectively. $V_n$ is the current referenced image submitted by the user. $K$ denotes product knowledge bases. $R_n$ denotes the corresponding system response in the $n$-th turn.

Given multimodal dialogue context $(D_n, V_n)$ and product knowledge $K$, our goal is to learn a generation model $P(R_n|D_n, V_n, K)$ that integrates multimodal information and generates a coherent and reasonable response $R_n$.

## 3.2 Modality-aligned Thought Chain Reasoning

The core idea of thought chain reasoning is to decompose the complex original task <input, output> into intermediate steps <input, intermediate results, output>. The generation of the intermediate results (called rationale) provides a reasoning process which facilitates inferring the final output.

For multimodal dialogue generation task, we construct modality-aligned thought chain reasoning to enhance the generation of reasonable responses. Specifically, we assume that for each dialogue turn, there exists a user query $S_n$ that summarizes the current user intention and question, and can be inferred from $(D_n, V_n, K)$. Thus, we take $S_n$ as the rationale of our thought chain, decompose the original task into user query summarization and system response generation, and propose a multimodal reasoner to unify these two stages. In the first stage, we send $(D_n, V_n, K)$ triple into a multimodal reasoner to align and integrate multimodal information, then generate the summarized user query $S_n$. In the second stage, we concatenate $D_n$ with $S_n$, and send the updated triple into the same reasoner to generate the final system response. Thus, the whole process is as follows:

$$\begin{aligned} S_n &= \mathrm{MR}(D_n, V_n, K), \\ R_n &= \mathrm{MR}(D_n \oplus S_n, V_n, K), \end{aligned} \tag{1}$$

where $\oplus$ is the concatenation operation. $\mathrm{MR}(\cdot)$ denotes the multimodal reasoner, which captures user intention from complementary multimodal information. The multimodal reasoner mainly consists of two parts: the visual representation learning module and the modality-aligned fusion module. The former learns to extract effective visual representation from images submitted by user, while the latter aligns relevant information in different modalities and adaptively fuses them.

## 3.3 Visual Representation Learning

As an old saying goes, "a picture is worth a thousand words". Since the images provided by users are important supplements or explanations to text content, introduction of visual modality has been known to provide complete details for building effective end-to-end dialogue systems [15, 34].

Previous methods acquire visual information by adopting CNN-based [1, 9] or transformer-based [6] pretrained visual encoders. They freeze these encoders and extract vision features from images, then send these features into the text generator to perform modality interaction with textual representation. However, the semantic gap between different modalities makes it difficult for downstream modules to fuse multimodal information, while maintaining their general generation ability [17]. Besides, the unprocessed images submitted by users may contain partial redundant information, which could adversely impact the system's ability to correctly interpret user intent.

To extract effective visual representation that can be easily accepted by text generators, we design a visual representation learning module based on a lightweight transformer. As shown in Figure 2, this module performs as a bridge between the text generator and the pretrained visual encoder, whose parameters are frozen to maintain its visual processing ability. We use the pre-trained visual encoder to extract original visual features from the user-uploaded image. Then we employ a set of trainable query vectors to learn pertinent information from the original image features. Finally, we take the processed query vectors to replace the original visual features as the learned visual representation and send them into the downstream model. Inspired by [17], we significantly reduce the number and dimension of query vectors compared to the original image features. This forces the query vectors to capture visual representation that is most relevant and useful for the target task.

Specifically, we adopt a vision transformer from CLIP [31] pretrained on vision-language task as our backbone visual encoder $\mathrm{VE}(\cdot)$. Given image $V$, we slice the image into a sequence of 2D patches and send it into the visual encoder to extract the original image features $H_{vo}$. Then, we initiate a sequence of trainable query vectors $q$, and send it into a $L$ layer interaction block with $H_{vo}$. At each layer, these queries interact with $H_{vo}$ through cross-attention layers, and interact with each other through self-attention layers. In the end, we project the learned visual queries into the unified semantic space to get the visual representation $H_{vq}$, which bridges the modality gap. The process is as follows:

$$\begin{aligned} & H_{vo} = \mathrm{VE}(V), q^0 = q \\ & \begin{cases} q^a = \mathrm{MultiHead}(q^{l-1}, W_q H_{vo}, W_q H_{vo}), \\ q^b = \mathrm{MultiHead}(q^a, q^a, q^a), \\ q^l = \mathrm{FFN}(q^b), \end{cases} \\ & H_{vq} = \mathrm{Linear}(q^L), \end{aligned} \tag{2}$$

where $W_q$ is a trainable matrix, $\mathrm{MultiHead}(\cdot)$ denotes the multi-head attention block [38], and $\mathrm{FFN}(\cdot)$ denotes the feed-foward network.

We pretrain the interaction block and the query vectors on a commonly used image-caption dataset. In particular, we adopt a pretrained T5 [33] model as the text generator, and prepend a short sentence "this is a picture of " as the textual prompt. Then we send the projected visual representation $H_{vq}$ of a given image into the T5 decoder for caption generation, and compute the cross-entropy loss with the ground truth caption for modal optimization. During this process, we freeze the parameters of the T5 model, forcing the visual representation learning module to acquire the capability of extracting useful visual information that can be directly interpreted by the text model.

Note that in the subsequent training, the visual encoder will remain frozen, while the $L$ layer interaction block and the query vector will be fine-tuned with the rest of our MaTCR model.

## 3.4 Modality-Aligned Fusion

This component aims to fully exploit the complementary information in multimodal context and product knowledge, capture user intention, and generate user query and system response in the two stages of the thought chain accordingly.

The summarized user query acts as a rationale of the thought chain, representing the model's understanding of the current dialogue state. Based on this understanding, the model is able to generate a response that is more aligned with the user's expectations. As demonstrated in [48], during the thought chain reasoning, the hallucinated rationales may mislead the generation of the final results, and the accuracy of intermediate rationales heavily relies on the interaction of different modalities and side knowledge. Thus, we

design this modality-aligned fusion module for the effective usage of multimodal context. It first extracts unified textual representation from dialogue history and product knowledge, then achieves the integration of multi-source information using multimodal alignment and gate-controlled fusion.

*3.4.1* ***Textual Representation Extraction.*** In this task, accessible textual information exists in both dialogue history and product knowledge. The product knowledge contains detailed attribute information that can assist in understanding the dialogue context and answering user's question. Therefore, we equip the dialogue history with corresponding knowledge from the knowledge base, searched by the product ID. We concatenate the searched key-value knowledge pairs together with a special token [KG] at the head as knowledge sequence $K$.

We take the encoder from a pretrained T5 model as product knowledge and dialogue history encoder, which extracts knowledge representation $H_k$ from $K$ and history representation $H_d$ from $D$ respectively. Then we concatenate a special token [CLS] with $H_k$ and $H_d$, send it into a textual interactive layer to obtain the unified textual representation $H_t$:

$$\begin{aligned} H_k &= \text{TE}(\text{Emb}(K)), \\ H_d &= \text{TE}(\text{Emb}(D)), \\ H_t &= \text{TB}(\text{Emb}([CLS]) \oplus H_k \oplus H_d), \end{aligned} \tag{3}$$

where $\text{Emb}(\cdot)$ is the embedding operation that sums up word embedding and position embedding of each token in the sequence, since the self-attention computation is order-less. $\text{TE}(\cdot)$ denotes the pretrained textual encoder, and the textual fusion layer $\text{TB}(\cdot)$ is a single layer transformer block.

*3.4.2* ***Multimodal Alignment.*** Aligning unimodal representation before multimodal fusion is highly beneficial for learning interactions between modalities [18]. Here we use Image-Text Matching (ITM) to conduct representation alignment. Specifically, we leverage textual representation $H_t$ as text query and compute multi-head cross-attention with visual representation, obtaining the image-enhanced text representation $H_{vt}$:

$$H_{vt} = \text{MultiHead}(H_t, H_{vq}, H_{vq}). \tag{4}$$

For ITM, we take the representation corresponding to [CLS] token at the head of the sequence as the joint matching representation. Then we append a fully-connected layer followed by sigmoid function to predict the matching similarity score:

$$\text{score}(V, T) = \text{sigmoid}(\text{FFN}(H_{vt,0})), \tag{5}$$

where $(V, T)$ is a positive or negative image-text pair sampled from the dataset $M$, and the negative pair is created by randomly replacing the image or text in the same batch. Note that $T$ refers to the combination of dialogue history $D$ and the corresponding product knowledge $K$. the value of the score is between 0 and 1. Finally, a binary cross-entropy loss is applied for optimization:

$$\begin{aligned} \mathcal{L}_{ITM} = -\mathbb{R}_{(V,T)\ M}(&y\log(\text{score}(V, T)+ \\ &(1-y)\log(1-\text{score}(V, T)))), \end{aligned} \tag{6}$$

where y is a 2-dimensional one-hot vector representing the ground truth label.

Considering that there is relevant but not completely overlapping content between textual context and the user image, image-text alignment facilitates the model to learn interrelated complementary information across different modalities.

*3.4.3* ***Gate-controlled Fusion.*** In order to obtain the complete contextual information for understanding the user's intention, we fuse the multimodal representation after alignment. Since different conversation turns may require various information, it is crucial to balance the contribution from the visual and textual modalities. Thus, we apply the gated fusion mechanism [16, 44] to fuse $H_{vt}$ and $H_t$. We take a weighted average score $\beta$ to adaptively combine the multimodal representation, and the fused representation is obtained by

$$\begin{aligned} \beta &= \text{sigmoid}(W_t H_t + W_v H_{vt}), \\ H_f &= \beta H_t + (1-\beta) H_{vt}, \end{aligned} \tag{7}$$

where $W_t$ and $W_v$ are trainable parameters.

*3.4.4* ***Query & Response Generation.*** After the multimodal fusion, we feed the fused representation into a transformer-based text decoder to generate the summarized user query $S$ or system response $R$, depending on the reasoning stages. The generation loss is computed with cross entropy, and the complete training objective of the model is as follows:

$$\begin{aligned} \mathcal{L}_{GEN} &= -\log(P(S/R|D, V, K)), \\ \mathcal{L} &= \gamma \mathcal{L}_{ITM} + \mathcal{L}_{GEN}, \end{aligned} \tag{8}$$

where $\gamma$ is a hyper-parameter to control the trade-off between the two losses. We initiate text encoder and decoder from the same pretrained T5 model, as corresponds to the pretraining of the visual representation learning module.

Note that in the response generation stage, the dialogue history is updated by concatenating with $S$ from the previous user query summarization stage. Thus, we seamlessly unify the two stages of our thought chain reasoning with this multimodal reasoner.

# 4 EXPERIMENTAL SETTINGS

## 4.1 Dataset

In this paper, we conduct experiments on JDDC 2.1 dataset, a Chinese multimodal task-oriented dialogue dataset collected by Zhao et al. [49]. The JDDC 2.1 dataset comprises approximately 246,000 dialogue sessions collected from a mainstream Chinese E-commerce platform[2]. Each dialogue session contains multiple pieces of text and images, along with product knowledge base, and describes a complete online customer service process, covering various scenarios like payment, logistics, pre-sales, and after-sales-maintenance. Meanwhile, Zhao et al. [49] defined four tasks over this dataset, of which our proposed method focuses on the multimodal dialogue response generation, which is the most natural application of customer service robots.

Additionally, we also observe the multimodal query rewriting task, which requires a model to produce a textual user query that covers the multimodal information of the original dialogue history, aiming to reduce the difficulty of understanding multimodal utterances. We leverage the annotated rewritten query from this task

[2]https://www.jd.com

as the ground truth summarized user query to train our MaTCR model for thought chain reasoning.

## 4.2 Comparison Methods

To demonstrate the effectiveness of our proposed model, we compare it with the following representative methods:

**Unimodal GPT** [49]: it provides a text-only generative model that takes textual context as input, and uses GPT-2 [32] decoder to generate the response.

**Multimodal GPT** [49]: it provides a multimodal generative model, which feeds the visual feature extracted from the last pooling layer of ResNet-18 into GPT-2 after a dimension transformation based on a feed-forward layer.

**ChatGPT**: it provides a large language model built upon InstructGPT [29], trained with instruction learning and reinforcement learning from human feedback, and specifically designed to interact with users in a genuinely conversational manner.

**Visual ChatGPT** [43]: it designs a prompt manager to combine ChatGPT with a series of pretrained visual foundation models, thus it enables ChatGPT to handle complex visual tasks.

**MATE** [10]: it introduces the transformer network [38] to capture the context semantic relation between the textual context and the visual context, and devises the Transformer-based decoder to generate the text response.

Among these methods, Unimodal and multimodal GPT [49] are the vanilla baselines provided with the JDDC 2.1 dataset. Unlike most previous multimodal dialogue models, MATE [10] is less restricted by specific data formats, making it easier to transfer to the general scenarios we study here, thus we choose it as a strong baseline in this work. In addition, considering its excellent generalization ability and problem-solving capability, we also introduced ChatGPT and Visual ChatGPT [43] as comparison methods[3].

## 4.3 Evaluation Metrics

To examine the quality of the generated responses, we adopt both automatic and human evaluation methods to compare the performance of different models.

*4.3.1* ***Automatic Evaluation.*** Task-oriented dialogue generation requires the system responses to be reasonable and coherent to the context. Following the existing baseline [49], we adopt the commonly-used **BLEU** [30] and **Rouge-L** [21] as the automatic metrics, which analyze the co-occurrences of n-grams in the the generated responses and the ground truth.

*4.3.2* ***Human Evaluation.*** Considering that the automatic metrics are not always accurate to evaluate the responses [22], we further conduct manual evaluation following previous works [10, 26]. Specifically, we randomly sample 200 testing pairs from the JDDC 2.1 test set. Given the dialogue context, five annotators are asked to conduct pair-wise comparison between the responses generated by MaTCR and two strong baselines from two perspectives: 1) Context Coherence: Whether the response is in accordance with the issue being discussed. 2) Informativeness: Whether the generated response contains informative content for satisfying user needs.

[3]These two methods generate responses in a few-shot way, by referring to the paralleled conversation in the dialogue history

The annotators need to judge which response is better independently. If the two responses are both proper or inappropriate, the comparison of this pair is treated as "draw". Ultimately, we average the results of three annotators and calculate their Fleiss' kappa scores [8].

## 4.4 Implementation Details

We adapt ViT-L/14 model[4] from CLIP [31] as our frozen visual encoder. For visual representation learning, we set the size of query vector $q$ to be 16x768, and set the number of layers for the interaction block to be $L = 2$. We pretrain the query vector and the interaction block on Flickr30k-CNA[5] image-caption dataset. To initiate the text encoder and decoder used in modality-aligned fusion, we adapt a Chinese version of mT5-large model[6] with 784M parameters pretrained on large-scale Chinese corpora. All the multi-head attention layers used in this work have a structure of 768 hidden size and 6 heads.

To train the model on JDDC 2.1 dataset, we set the maximum length of dialogue history to 160, and use the Adam optimizer [11] with a learning rate of 3e-4 for model optimization. At the inference stage, the maximum decoding length of the queries and responses is set to 40, and we adopt beam search decoding with a beam size of 3. All our experiments are implemented with PyTorch, and the entire model is trained on RTX3090 GPUs.

# 5 RESULTS AND ANALYSIS

## 5.1 Automatic Evaluations

Table 1 shows that MaTCR outperforms the compared baselines regarding all automatic evaluation metrics. The significant improvement on Bleu and Rouge-L metrics shows that the responses of MaTCR are relevant and coherent with the dialogue context. The results indicate that our model can effectively utilize multimodal information in the context to generate appropriate responses.

We also analyze the advantages of MaTCR over other methods. The text-only methods lack reference to visual information, leading to inferior performance. In multimodal setting, both Multimdal-GPT and Visual-ChatGPT lack the necessary interaction process to allocate limited attention on task-related visual information. Although MATE employs modality interaction by using cross-model attention computation, the absence of multimodal alignment makes it vulnerable to the semantic gap problem, which in turn hinders the effective integration of complementary multimodal information.

## 5.2 Human Evaluations

The human evaluation results in able 2 show that MaTCR consistently outperforms all the strong baselines and achieves significant improvements in both context conherence and informativeness. The strengths in these two aspects respectively indicate that MaTCR's responses are more relevant to dialogue context, and contain more reasonable information. This validates the advantages of the multimodal thought chain reasoning, which endows the model with superior multimodal comprehension ability, and thus it can generate helpful responses preferred by the annotators. Besides, we

[4]https://github.com/OpenAI/CLIP
[5]https://zero.so.com/download.html
[6]https://fengshenbang-doc.readthedocs.io/zh/latest/

**Table 1: Automatic evaluation results. The best results are highlighted in bold.**

| Methods | | Bleu-1 | Bleu-2 | Bleu-3 | Bleu-4 | Rouge-L |
|---|---|---|---|---|---|---|
| Text-only | Uni-GPT [49] | 16.02 | 12.83 | 8.17 | 6.63 | 17.57 |
| | ChatGPT | 17.65 | 10.98 | 6.91 | 4.56 | 14.39 |
| Multimodal | Multi-GPT [49] | 17.90 | 14.81 | 8.72 | 7.23 | 20.36 |
| | Visual-ChatGPT [43] | 17.92 | 11.05 | 6.51 | 4.41 | 14.01 |
| | MATE [10] | 22.44 | 17.56 | 15.11 | 13.66 | 24.94 |
| | MaTCR | **29.03** | **24.69** | **22.43** | **21.04** | **27.79** |

**Table 2: Human evaluation results in two aspects: Context Coherence and Informativeness.**

| | Context Coherence | | | | Informativeness | | | |
|---|---|---|---|---|---|---|---|---|
| Opponent | win | loss | draw | kappa | win | loss | draw | kappa |
| vs. MATE | 37.9% | 16.2% | 45.9% | 0.51 | 56.2% | 12.4% | 31.4% | 0.58 |
| vs. Visual-ChatGPT | 38.1% | 26.3% | 35.6% | 0.56 | 43.7% | 19.5% | 36.8% | 0.44 |

**Table 3: Evaluation results of ablation study.**

| Model | Bleu-1 | Bleu-2 | Bleu-3 | Rouge-L |
|---|---|---|---|---|
| MaTCR | **29.03** | **24.69** | **22.43** | **27.79** |
| w/o.TCR | 25.75 | 21.99 | 19.83 | 25.27 |
| w/o.MAF | 26.61 | 22.54 | 20.03 | 25.11 |
| w/o.VRL | 28.11 | 22.98 | 20.50 | 27.24 |

**Table 4: Comparison of different user query settings**

| Model | Bleu-1 | Bleu-2 | Bleu-3 | Rouge-L |
|---|---|---|---|---|
| MaTCR | 28.25 | 22.84 | 21.61 | 27.36 |
| Vanilla | 26.41 | 21.19 | 20.39 | 26.20 |
| Gold | 33.68 | 26.83 | 23.26 | 34.14 |

find that both MATE and Visual-Chatgpt often can not effectively leverage the visual information provided by images. The former frequently fails to generate useful suggestions, while the latter tends to produce a large number of negative responses starting with apologies. This results in their inferior performance on the informativeness metric compared to MaTCR.

## 5.3 Ablation Study

We conduct experiments on different variants of MaTCR to investigate the effectiveness of each component. The variants include: (1) w/o. TCR: it disrupts the thought chain by removing the user query summarization process, resulting in a regression of the task to the original <input, output> pattern. (2) w/o. VRL: it removes the visual representation learning module, and takes the original image features extracted by the frozen visual encoder as visual representation. (3) w/o. MAF: it removes the modality-aligned fusion module, concatenates the visual and textual representation together, and sends them into text decoder for query/response generation.

As reported in Table 3, the performance of w/o. TCR degrades dramatically. This demonstrates the vital importance of the thought chain reasoning as it can decompose the complex task and enhance the model's ability to understand the user intention. Besides, our model achieves better results than w/o. VRL, indicating that learning to extract effective visual representation benefits the subsequent generation task. Moreover, the performance of w/o. MAF also drops, reflecting that it is crucial to align different modalities for bridging the semantic gap before fusing multimodal information. In general, our proposed model largely exceeds all variants, verifying the effectiveness of modality-aligned thought chain reasoning framework.

## 5.4 Thought Chain Analysis

Although the ablation study confirms the negative impact on model performance in the absence of thought chain training, it does not intuitively demonstrate the potential benefits of good thought chains for multimodal task-oriented dialogue generation in this work.

To verify this, we test the trained MaTCR on the query rewritting dataset mentioned in section 4.1, which has manually annotated summarized user queries, and compare the results with two variants of the model: (1) Vanilla: concatenates no query for the response generation process. (2) Gold: concatenates ground truth query provided by the dataset for the response generation. Table 4 shows the comparison results. We can see that the Gold variant outperforms the original MaTCR and the Vanilla variant performs the worst. The results indicate the great utility of inserting high-quality intermediate reasoning into multimodal task-oriented dialogue generation.

## 5.5 Case Study

To further investigate the quality of responses generated by MaTCR intuitively, we show two dialogue cases in Figure 3, which cover two common scenarios in the dataset: post-sales inquiries about product usage and pre-sales inquiries about product information.

As we can see, although all three models can generate fluent responses relevant to the dialogue context and product knowledge,

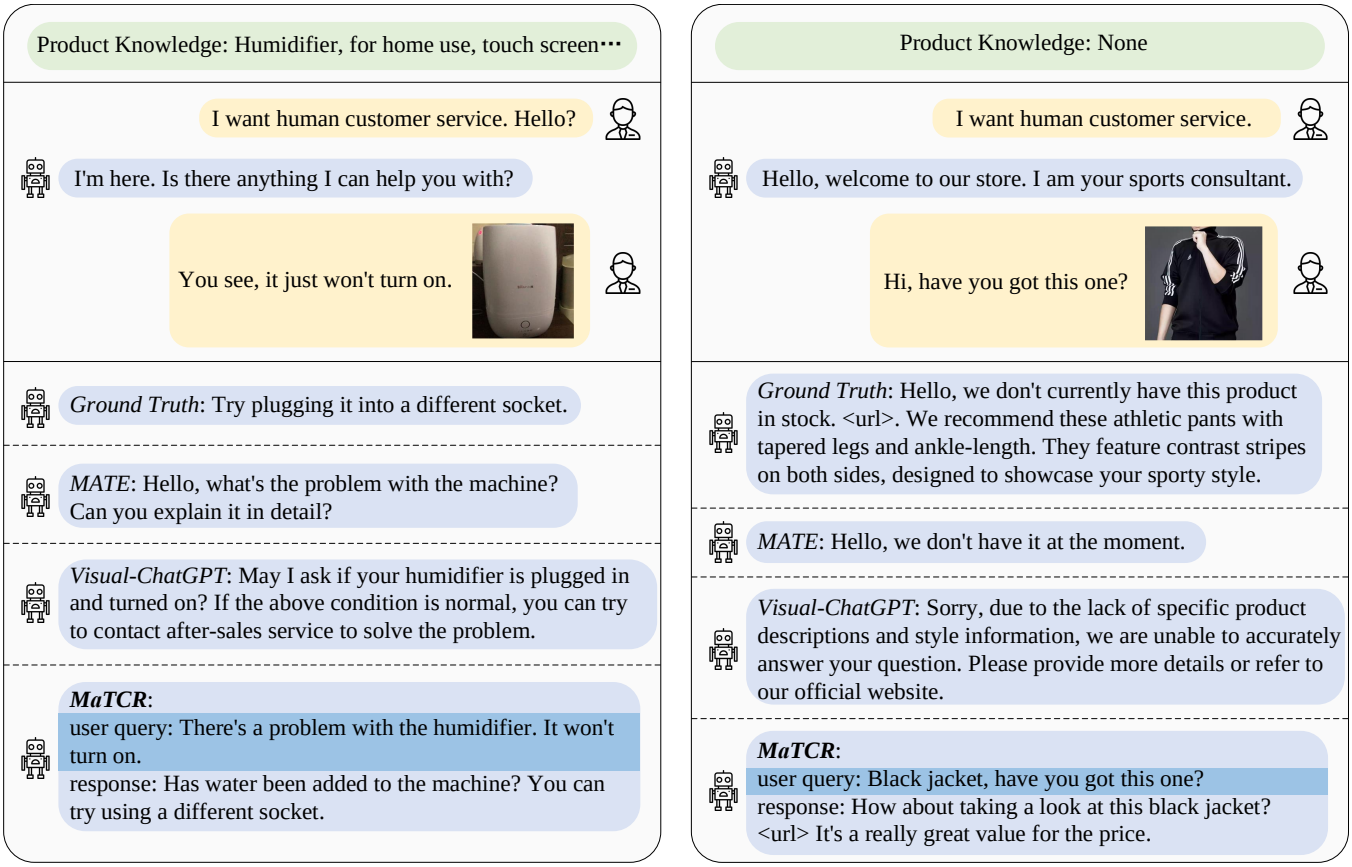


**Figure 3: Two cases from JDDC 2.1 dataset (translated from Chinese).**

the responses generated by MaTCR are more reasonable and comprehensive. Specifically, in Case 1, MaTCR can understand that the user's purpose is to require a solution for the issue that their humidifier cannot be turned on. Therefore, it proposes two suggestions: checking the water level and changing the socket, both of which are reasonable and helpful. Case 2 shows a more complex scenario, where there is no product knowledge and the user's expression is ambiguous. Although MaTCR does not provide the same answer as ground truth regarding the existence of the product (this would require access to a more comprehensive database), it obtaines the correct image information and performs a reasonable product recommendation operation.

We also show the summarized user queries generated by MaTCR in these two cases. These queries indicate that the model is capable of combining multimodal information to comprehend user intentions. Again, it demonstrates the effectiveness of our modality-aligned thought chain reasoning framework.

In addition, we analyze the unsatisfactory performance of Visual-ChatGPT in Case 2. We found that it generates such an image description in its pipeline process:"adidas adidas adidas adidas adidas ...". While this description is technically correct, it does not effectively assist in the subsequent task. This further demonstrates the validity of incorporating visual representation learning and multimodal information interaction modules into our model.

## 6 CONCLUSION

In this work, we propose a modality-aligned thought chain reasoning framework for multimodal task-oriented dialogue generation. To elicit the multi-step reasoning abilities of the model, We construct a thought chain by decomposing the complex original task into user query summarization and system response generation, then we design a modality-aligned fusion module to unify these two process. Extensive experiments show that our proposed method is superior to existing methods, demonstrating the effectiveness of our framework. What's more, it verifies the potential of exploiting multi-step reasoning for intelligent multimodal conversation agents, which benefits both the model performance and the interpretability.

In the future, we will extend this work by studying the transition of user intention and exploring the application of external knowledge.

## ACKNOWLEDGMENTS

This work was supported in part by the National Key RD Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: 62322211, 62236008, U21B2038, 62225207, U19B2038 and 61931008, and Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2020108.

# REFERENCES


[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 213–229.

[2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[4] 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 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 445–454.

[5] Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. 2019. Neural logic machines. *arXiv preprint arXiv:1904.11694* (2019).

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

[7] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. *arXiv preprint arXiv:2211.05719* (2022).

[8] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[9] 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.

[10] 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 28th ACM International Conference on Multimedia*. 2755–2764.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916* (2022).

[13] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: a task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667* (2021).

[14] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329* (2022).

[15] Hung Le, S Hoi, Doyen Sahoo, and N Chen. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.

[16] Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173* (2022).

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[19] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing* 31 (2022), 2726–2738.

[20] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*. 801–809.

[21] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[22] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2122–2132.

[23] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3003–3018.

[24] Yiting Liu, Liang Li, Beichen Zhang, and Qingming Huang. 2022. Think Beyond Words: Exploring Context-Relevant Visual Commonsense for Diverse Dialogue Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3106–3117.

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[26] Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. UniTranSeR: A Unified Transformer Semantic Representation Framework for Multimodal Task-Oriented Dialog System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 103–114.

[27] Jiayuan Mao, Freda Shi, Jiajun Wu, Roger Levy, and Josh Tenenbaum. 2021. Grammar-based grounded lexicon learning. *Advances in Neural Information Processing Systems* 34 (2021), 7865–7878.

[28] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal dialog system: Generating responses via adaptive decoders. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1098–1106.

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[32] 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.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[34] 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*, Vol. 32.

[35] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945* (2018).

[36] Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021. R^3Net:Relation-embedded Representation Reconstruction Network for Change Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9319–9329.

[37] Yunbin Tu, Chang Zhou, Junjun Guo, Huafeng Li, Shengxiang Gao, and Zhengtao Yu. 2023. Relation-aware attention for video captioning via graph learning. *Pattern Recognition* 136 (2023), 109204.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[39] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2022. Semantic and relation modulation for audio-visual event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[40] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. 2023. Context-Aware Proposal–Boundary Network With Structural Consistency for Audiovisual Event Localization. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[41] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7026–7035.

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[43] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

[44] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462* (2021).

[45] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2561–2569.

[46] Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453* (2021).

[47] 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*. 695–703.

[48] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).

[49] Nan Zhao, Haoran Li, Youzheng Wu, and Xiaodong He. 2022. JDDC 2.1: A Multimodal Chinese Dialogue Dataset with Joint Tasks of Query Rewriting, Response Generation, Discourse Parsing, and Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 12037–12051.

[50] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2021. MMChat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154* (2021).

[51] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625* (2022).