

JDDC 2.1: A Multimodal Chinese Dialogue Dataset with Joint Tasks of Query Rewriting, Response Generation, Discourse Parsing, and Summarization

Nan Zhao*, Haoran Li*, Youzheng Wu, Xiaodong He

JD AI Research

{zhaonan8, lihaoran24}@jd.com

Abstract

The popularity of multimodal dialogue has stimulated the need for a new generation of dialogue agents with multimodal interactivity. When users communicate with customer service, they may express their requirements by means of text, images, or even videos. Visual information usually acts as discriminators for product models, or indicators of product failures, which play an important role in the E-commerce scenario. On the other hand, detailed information provided by the images is limited, and typically, customer service systems cannot understand the intent of users without the input text. Thus, bridging the gap between the image and text is crucial for communicating with customers. In this paper, we construct JDDC 2.1, a large-scale multimodal multi-turn dialogue dataset collected from a mainstream Chinese E-commerce platform¹, containing about 246K dialogue sessions, 3M utterances, and 507K images, along with product knowledge bases and image category annotations. Over our dataset, we jointly define four tasks: the multimodal dialogue response generation task, the multimodal query rewriting task, the multimodal dialogue discourse parsing task, and the multimodal dialogue summarization task. JDDC 2.1 is the first corpus with annotations for all the above tasks over the same dialogue sessions, which facilitates the comprehensive research around the dialogue. In addition, we present several text-only and multimodal baselines and show the importance of visual information for these tasks. Our dataset and implements will be publicly available.

1 Introduction

With the development of the Internet, multimodal dialogue has become much more natural and prevalent in many scenarios, such as e-commerce, restaurant, travel, and so on, which stimulates research

on dialogue agents with multimodal perceptions, understanding the interaction between vision and language. Take the scene of e-commerce as an example, when users resort to customer service for solving the difficulties they encounter in the process of online shopping, there could be various forms of information, including text, images, or even videos, which establishes a great challenge for customer service systems to understand users' requirements. Generally, users tend to express their needs with text. While sometimes, the text fails to convey enough information, and in that case, users may upload some images. For example, in Figure 1, images are used for distinguishing different product models for the same brand or used for identifying the location and cause of product failures. Therefore, there is an urgent need for customer service systems to understand multimodal information sent by users to provide proper responses.

Multimodal information processing has been widely explored recently. Researches on image captioning (Xu et al., 2015; Anderson et al., 2018; Pan et al., 2020), visual question answering (Antol et al., 2015; Yang et al., 2016; Lu et al., 2016), multimodal machine translation (Calixto et al., 2017; Caglayan et al., 2017; Helcl et al., 2018), multimodal summarization (Li et al., 2017, 2018a; Zhu et al., 2018; Li et al., 2018b, 2020a,b; Zhu et al., 2020a), and visual or multimodal dialogue (Das et al., 2017a,b; Murahari et al., 2020; Kottur et al., 2021) have made remarkable progress. However, most of the existing researches on multimodal dialogue are based on independent single-turn question answering or simulated dialogue flows, and the application of multimodal multi-turn dialogue in real E-commerce scenarios remains to be studied. In this paper, we build a large-scale multimodal dialogue dataset in E-commerce that aims to boost the research on multimodal dialogue.

Early work (Ritter et al., 2011; Wang et al., 2013; Sordani et al., 2015; Lowe et al., 2015; Li et al.,

*Equal contribution.

¹<https://JD.com>

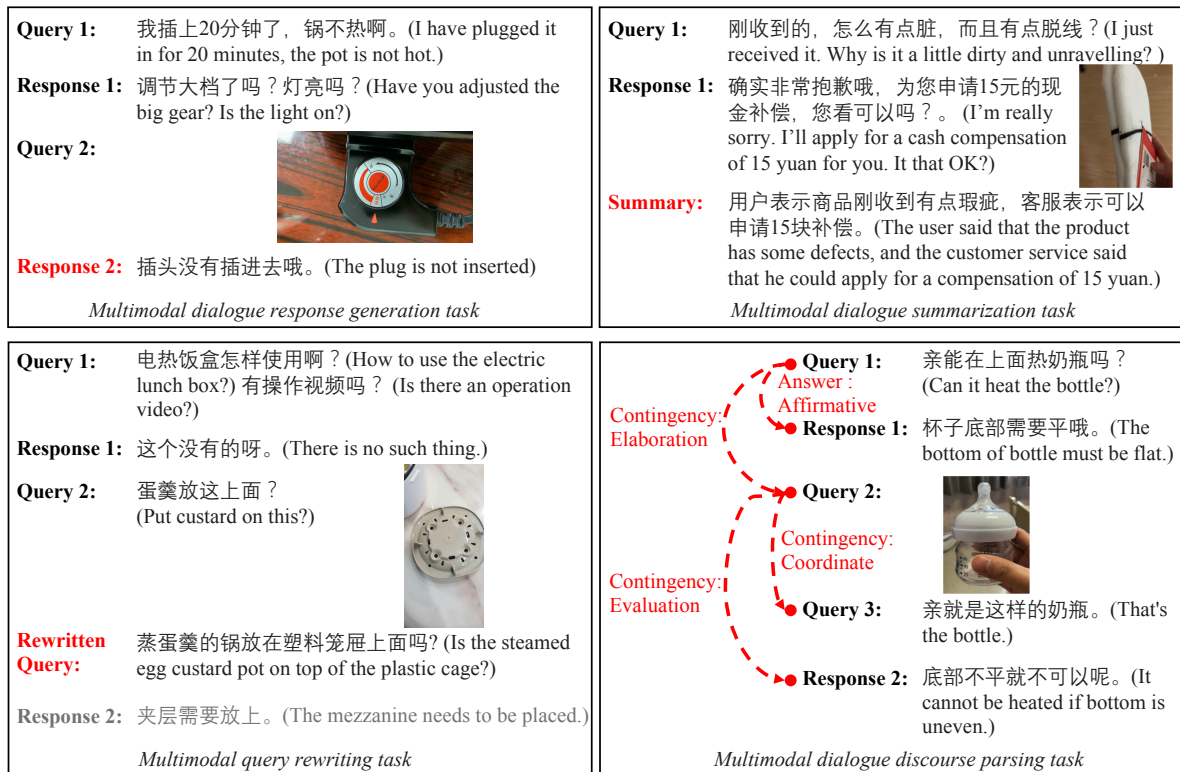


Figure 1: Four segments of dialogue sampled from our JDDC 2.1 corpus, corresponding to four tasks we define over the dataset. For each task, the output is marked in red. We show the second-level and third-level discourse relations for the dialogue discourse parsing task. Note that these samples are truncated for demonstration purposes.

2016b; Mazaré et al., 2018) construct dialogue corpus with discussion records on the social media, such as Twitter, Sina Weibo, and Reddit. Although the discussions on social media consist of multiple turns, they are quite different from real dialogues due to the lack of an explicit goal in the conversation (Budzianowski et al., 2018). To mitigate this problem, researches (Zhou et al., 2018; Budzianowski et al., 2018; Zhang et al., 2018; Dinan et al., 2019; Quan et al., 2020) build the dialogue datasets through crowd-sourcing by asking annotators to talk to each other according to given dialogue objectives. While in the real world, topic switches across multi-domain and emotional interactions are more abundant, and thus, dialogue datasets in real scenarios are valuable for research. The JDDC corpus (Chen et al., 2020) is a dialogue dataset consisting of conversations about after-sales topics in E-commerce scenarios, which is goal-driven and with long-term dependency on the context. In addition, JDDC contains task-oriented, chitchat, and question-answering dialogues.

With the widespread use of smartphones, taking screenshots and photos has been very convenient, and users often describe their issues by a

combination of text and image when they communicate with customer service, which motivates us to explore multimodal dialogue tasks. Before this work, we construct the JDDC 2.0 corpus (Zhao et al., 2021) that is composed of multimodal dialogues where each dialogue session contains multiple pieces of text and at least one image, containing about 246 thousand dialogue sessions, 3 million utterances, and 507 thousand images, along with product knowledge bases and image category annotations. In this paper, on the foundation of JDDC 2.0 (Zhao et al., 2021), we build the JDDC 2.1 corpus. Over this dataset, we define four tasks, shown in Figure 1, including (1) a multimodal dialogue response generation task that predicts the current response given historical conversation records, (2) a multimodal query rewriting task that transforms multimodal queries into texts, aiming to reduce the difficulty of understanding multimodal utterances, (3) a multimodal dialogue discourse parsing task that converts a dialogue session into a discourse tree, aiming to analyze the discourse structure for dialogue sessions, and (4) a multimodal dialogue summarization task that generates a summary for a dialogue session, aiming to mine key informa-

tion among utterances. Note that the annotation of query rewriting, discourse parsing, and dialogue summarization is conducted with the same dialogue sessions, making it possible to explore the relationship between these tasks. So far as we know, JDDC 2.1 is the first dataset with comprehensive annotations for all these above dialogue-related tasks. For all the tasks, we conduct experiments with text-only and multimodal baselines and verify the necessity of visual information for these tasks.

2 Related Work

Visual Dialogue dataset (VisDial) (Das et al., 2017a) first introduces visual contents into dialogues, in which the utterances are a group of questions and answers towards the corresponding image. This dataset focuses on understanding the given images. Mostafazadeh et al. (2017) observes that, in social media, information for conversations around images is beyond what is visible in the image. They propose a new task called image-grounded conversation (IGC) that aims to constitute conversations with the images as the grounding, where the objects in images may not be mentioned in the conversation. In other words, images in IGC act as conversation topics. Similar to IGC, Image-Chat dataset (Shuster et al., 2020) is also an image-grounded dialogue dataset, where the dialogue is performed based on a given emotional mood or style, which are key factors in engagingness (Guo et al., 2019).

Multimodal dialogue is the focus of this paper, which is different from the visual dialogue. We summarize the characteristics of multimodal dialogue as follows. (1) There may be more than one image for a multimodal dialogue session. (2) The images can be updated with the advance of dialogue. (3) The questions and answers can be either multimodal or monomodal. (4) Knowledge bases may be useful for multimodal dialogue. (5) Multimodal dialogue models sometimes need to clarify users' requirements with some dialogue strategies like rhetorical questions.

As dialogue systems are widely deployed in the E-commerce domain, and nowadays, many online customer service robots have been put into use to provide 24-hour services to help customers solve various problems in the process of online shopping. There have been existing multimodal dialogue datasets in the E-commerce domain. Saha et al. (2018) build the Multimodal Dialogs dataset

(MMD) that consists of over 150K dialogue sessions between shoppers and sales agents, with 84 dialog states and various conversation flows suggested by fashion retail experts. The dialogue scene in MMD is limited for the pre-sales guidance, while other scenes, such as payment, logistics, and after-sales maintenance, are not covered. The dataset of SIMMC 2.0 (Kottur et al., 2021) contains 11K dialogue sessions between customers and virtual assistants for situated and photo-realistic VR applications. Similar to MMD, the target scene for SIMMC 2.0 is the pre-sales guidance. In fact, changes in scenes are quite frequent. For example, a dialogue system usually needs to solve customers' problems ranging from product selection, and payment, to logistics and distribution in a dialogue session. Thus, in this paper, we collect the JDDC 2.1 dataset that covers almost the complete process in E-commerce. In addition, to our knowledge, JDDC 2.1 is the first dataset with joint annotations of multimodal dialogue response generation, multimodal query rewriting, multimodal dialogue discourse parsing, and multimodal dialogue summarization tasks. Table 1 shows the detailed comparison of JDDC 2.1 with other existing visual and multimodal dialogue datasets.

3 The JDDC 2.1 Corpus

3.1 Data Collection

We collect our dataset from JD.com, a large E-commerce platform in China. We select the conversations between users and customer service for two categories of products with large sales volume, including *small home appliances* and *fashion*, as the source of our dataset. In real E-commerce scenarios, the type of dialogue is diversified, where the customer service needs to answer the questions posed by users passively and recommend products to users actively. To ensure high quality in diversity, in the process of data selection, we only select conversations of customer service staff with gold medals, who tend to answer the questions more accurately and recommend products more suitably than the general staff. In addition, according to our observations, dialogue behaviors for customer service staff with gold medals are richer and more natural than general ones, and thus we select the dialogue sessions of these staff. We collect conversation logs for one month and finally maintain the conversations containing at least one image as our dataset.

Dataset	Language	# Dialogues	# Average Utterances	# Images	Additional Tasks		
					Query Rewrite	Discourse	Summary
VisDial	English	120,000	20.0	120,000	✗	✗	✗
IGC	English	4,222	6.0	4,222	✗	✗	✗
Image-Chat	English	201,779	2.0	201,779	✗	✗	✗
MMD	English	150,629	40.0	385,969	✗	✗	✗
SIMMC 2.0	English	11,244	10.4	1,566	✗	✗	✗
JDDC 2.1	Chinese	246,153	14.1	507,678	✓	✓	✓

Table 1: Comparison of JDDC 2.1 with other datasets. The first three datasets are for visual dialogue, and the latter three are for multimodal dialogue.

	All	Training Set		Validation Set		Test Set	
		Home appliances	Fashion	Home appliances	Fashion	Home appliances	Fashion
# Dialogues	246,153	103,555	93,371	12,941	11,674	12,935	11,674
# Utterances	3,459,888	1,481,151	1,284,819	185,879	162,630	185,197	160,217
# Avg. utterances per dialogue	14.06	14.30	13.76	14.36	13.93	14.32	13.72
# Images	507,678	210,386	195,549	26,218	24,696	26,031	24,888
# Avg. images per dialogue	2.06	2.03	2.09	2.03	2.11	2.01	2.13
Avg. conversation length	27.24	23.77	31.25	23.81	31.09	23.68	31.29

Table 2: Statistics of the JDDC 2.1 dataset.

3.2 Dataset Annotation

For the multimodal query rewriting, dialogue discourse parsing, and dialogue summarization tasks, we select a group of the same dialogue sessions for human annotation.

For the multimodal query rewriting task, we employ human annotators to transform the multimodal query consisting of texts and images into a textual query. Before rewriting, annotators first need to judge whether the multimodal query can be reconstructed from a text². Then, for the queries that are suitable for rewriting, annotators are instructed to generate the written query with a simple text and guarantee the written query can cover information of the original multimodal query.

For the multimodal dialogue discourse parsing task, following the STAC corpus (Asher et al., 2016), we adopt the discourse structures of SDRT (Cadilhac et al., 2013). For human annotation, first, for simplicity, we regard each utterance as the elementary discourse unit (EDUs). Second, for a given EDU, we recognize a previous EDU that the current EDU should be linked to. Last, we classify the relation between the linked EDU pairs. We define three-level discourse relation, and the first level includes *Question&Answer (QA)*, *Imperative*, *Promise*, *Communication*, and *Contingency*.

²According to the annotation results, about 14.28% multimodal queries cannot be rewritten.

More details can be found in Appendix D.

For the multimodal dialogue summarization task, human annotators are first asked to filter out the trivial utterances, e.g., greeting, waiting, and self-introduction. Then, they write a short summary that covers the key information in the multimodal dialogue session.

3.3 Annotation Quality Control

To ensure a satisfactory annotation quality of our dataset, we apply group-by-group acceptance testing. We take each 100 sessions as a testing group, and for each group, we randomly sample 10% instances for acceptance testing. If the acceptability rate is lower than 90%, the corresponding group needs to be re-annotated.

3.4 Dataset Statistics

Table 2 shows the statistics of JDDC 2.1 that contains 246,153 dialogue sessions, 3,459,888 utterances, and 507,678 images. The dataset is divided into the training set, the validation set, and the test set according to the ratio of 80%, 10%, and 10%. The number of dialogues in the two categories is roughly equal. More statistics are in Appendix A.

For the multimodal query rewriting, dialogue discourse parsing, and dialogue summarization task, we annotate the same 2,000 sessions from JDDC 2.1, which results in around 10,000 “original multimodal query, rewritten query” pairs, and 2,000

“dialogue session, dialogue summary” pairs. The average number of characters in the summary is 38.3. The statistics for the discourse parsing task are shown in Table 3.

	Total	Train	Valid	Test
# Dialogues	2,000	1,800	100	100
# Turns	27,832	24,956	1,437	1,439
# EDUs	45,173	40,664	2,269	2,240
# Discourse Relations	43,173	38,864	2,169	2,140

Table 3: Statistics of the discourse parsing task.

Categories	Sub-categories
Screenshot	Screenshot of product
	Screenshot of product order
	Screenshot of logistics order
	Screenshot of after-sales service order
	Screenshot of text message
	Screenshot of user comment
	Screenshot of system or software
Photo	Screenshot in other scenes
	Photo for purchasing consultation
	Photo of product with damaged appearance
	Photo of products with malfunction
	Photo of product with missing items
	Photo for product recommendation and comparison
	Photo for product installation
Photo of user screen shot	

Table 4: Definition of image categories and sub-categories.

3.4.1 Image Category

The images in JDDC 2.1 can be divided into two categories: screenshots and real photos. Further, we classify these two categories into 15 sub-categories according to the characteristics of the business scenarios involved in the entire shopping process, which are shown in Table 4. More statistical results are shown in Appendix B.

3.4.2 Knowledge Base

In the E-commerce domain, whatever the scenes of pre-sales purchasing consultation or after-sales return of a product, the conversation always involves at least one product. Thus, for the product mentioned in the conversation, we provide the corresponding knowledge base (Zhu et al., 2020b; Xu et al., 2021) that describes attribute information in detail, which can be useful for improving the faithfulness of the generated text (Li et al., 2018c; Yuan et al., 2020). The knowledge base contains 30,205 products, involving 231 product sub-categories,

and 759 types of product attributes. The total number of “product, attribute, value” triples is 219,121. The overall statistics of the knowledge base are shown in Appendix C.

3.5 Dataset Quality Testing

To examine how related the images are to the responses, we sample 200 sessions for manual annotation. As a result, we find that all images are related to the responses to some degree. 74% of the responses cannot be produced without the images because images provide indispensable information. For the remaining 26% of the sessions, customs would explain the images with texts, and thus, the images provide complementary information in that case.

4 Task Definition

We specify the definition of four tasks, namely the multimodal dialogue response generation task, the multimodal query rewriting task, the multimodal dialogue discourse parsing task, and the multimodal dialogue summarization task.

The task of multimodal dialogue response generation is defined as:

$$(H_{<n}, Q_n, K, V, \rightarrow R_n)$$

where $H_{<n}$ denotes the dialogue context before the n -th turn, Q_n denotes user query in the n -th turn, K denotes product knowledge bases, V denotes images, R_n denotes response in the n -th turn. That is, the task is to predict the current response on the condition of the dialogue context, a user query, images, and the related knowledge base information.

The multimodal query rewriting task aims to produce a text that covers the textual and visual information of the original multimodal query, which consists of two sub-tasks. The first one is to predict whether the multimodal query can be rewritten by a text without sacrificing critical information, and the latter one is rewritten query generation task that can be defined as:

$$(H_{<n}, Q_n, K, V, \rightarrow Q_n^R)$$

where Q_n^R denotes rewritten query in the n -th turn.

Multimodal dialogue discourse parsing task is designed to convert the discourse session into a discourse tree, which can be formalized as:

$$(H_{<n}, Q_n, K, V, \rightarrow T)$$

where T denotes a discourse parsing tree. Note that images in the discourse session are also regarded as EDUs, and correspondingly, the discourse relations between images and texts are involved.

For multimodal dialogue summarization task aims at generating a condensed summary for a given dialogue session, which can be defined as:

$$(H_{<n}, Q_n, K, V, \rightarrow S)$$

where S denotes a dialogue summary.

5 Experiments

In this section, we present experimental results for each task, human evaluations, error analyses, and case studies.

5.1 Multimodal Dialogue Response Generation Task

We conduct experiments with retrieval-based and generative-based models. For retrieval-based models based on textual information, we first extract token embeddings for the dialogue context and the current query with BERT (Devlin et al., 2019), and then we use the averaged token embedding to represent the instance. Next, we use the k -NN algorithm³ to retrieve the top 40 candidates for further re-ranking. For retrieval-based models based on multimodal information, beyond these 40 candidates, we additionally retrieve the top 10 candidates based on images. Specifically, we extract the activations from the last pooling layer of ResNet-18 (He et al., 2016) as the features for images, which are used to retrieve candidate responses. For both models, we use the TransResNet model (Shuster et al., 2020) (text-only or multimodal retrieval models correspondingly) to re-rank the candidates and maintain the one with the highest score as the final response.

We use GPT-2 (Radford et al., 2019) as the generative-based baseline. For text-only settings, the dialogue context and the current query are the input for GPT, and then a response is generated by the decoder. For multimodal settings, beyond textual information, we feed 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. The experiment settings, *i.e.*, hyper-parameters, for all the models used in

³We implement it with Faiss library at <https://github.com/facebookresearch/faiss>.

the paper are set as the same as the original implementations, which can be available in our open-source project.

Automatic evaluation results for the multimodal dialogue response generation task are shown in Table 5. We use BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) to evaluate the overlap between the model output and the ground-truth result. In addition, we present distinct n-grams (Li et al., 2016a) to evaluate the diversity of the generated responses. From the results, we can conclude that generative-based models are significantly better than retrieval-based models. Multimodal models outperform retrieval-based models, which corroborates the necessity of visual information for this task.

Models	BLEU	RG-L	Dt-1	Dt-2
Text-only Retrieval	5.73	14.45	0.45	6.65
Multimodal Retrieval	5.99	14.71	0.46	6.89
Text-only GPT	9.62	21.40	0.23	2.86
Multimodal GPT	10.21	22.14	0.25	3.27

Table 5: Results (%) for the multimodal dialogue response generation task. RG and Dt are short for ROUGE and Distinct, respectively.

5.2 Multimodal Query Rewriting Task

The multimodal query rewriting task consists of two sub-tasks. For the first one, *i.e.*, identifying whether the multimodal query can be rewritten by a text without loss of important information, we adopt models of the text-only BERT and the multimodal TransResNet (Shuster et al., 2020) as the baseline methods. The results are shown in Table 6, where TransResNet exhibits the obvious advantage over BERT.

Models	F1 score
Text-only BERT	88.42
Multimodal TransResNet	89.64

Table 6: Classification results (%) for whether the multimodal query can be rewritten.

For the latter one, rewritten query generation task, similar to the multimodal dialogue response generation task, we apply text-only and multimodal GPT models. Once again, the results shown in Table 7 verify the effectiveness of the multimodal model.

Models	BLEU	RG-L	Dt-1	Dt-2
Text-only GPT	2.38	19.57	1.54	6.15
Multimodal GPT	5.45	26.45	3.19	22.86

Table 7: Results (%) for the multimodal query writing task.

5.3 Multimodal Dialogue Discourse Parsing Task

For this task, we apply the model of Deep Sequential (Shi and Huang, 2019) as the text-only baseline⁴. For the multimodal model, similar to the dialogue response generation task, we extract the visual features extracted using ResNet-18, which are then fed into the Deep Sequential model through a feed-forward layer. Following Shi and Huang (2019), we use micro-averaged F1 score to evaluate the model performances for the link prediction and relation classification. The results shown in Table 8 prove that images contribute to the discourse parsing task.

Models	Link	Link & Relation
Text-only Deep Sequential	77.37	57.01
Multimodal Deep Sequential	78.12	57.77

Table 8: Results (%) for the multimodal dialogue discourse parsing task. “Link” denotes the link prediction and “Relation” denotes the relation classification.

5.4 Multimodal Dialogue Summarization Task

For this task, we adopt Lead (first N characters according to the average length of the target) and LexRank (Erkan and Radev, 2004) as extractive baselines, Pointer-Generator (See et al., 2017) and BERTSUMABS (Liu and Lapata, 2019) as abstractive baselines. Multimodal summarization model based on BERTSUMABS and visual feature extracted from ResNet-18 is conducted, similar to the multimodal dialogue response generation task. We report ROUGE-1/2/L F1 scores as in Table 9. From the results, we can conclude that the multimodal model slightly outperforms text-only models. Looking through the data we find that most of the important information is contained in the text, and the image plays a supplementary role.

⁴For text-only settings, the image EDUs are ignored.

Models	RG-1	RG-2	RG-L
Lead	17.39	1.64	13.98
LexRank	25.68	10.14	21.94
Pointer-Generator	38.91	23.81	37.86
Text-only BERTSUMABS	49.71	29.97	44.64
Multimodal BERTSUMABS	50.31	29.76	44.71

Table 9: Results (%) for the multimodal dialogue summarization task.

5.5 Further Analysis for Dialogue Response Generation

In this section, we report the human evaluation, error analysis, and case study on the dialogue response generation task.

5.5.1 Human Evaluation

First, we report the human evaluations on the dialogue response generation task. We sample 500 instances from the test set and invite customer service experts to evaluate the responses generated by baseline models. Specifically, we evaluate the quality of response in terms of *Fluency* (how easy of understand) and *Relevance* (whether the response solves the problem) with a 3-point scale (3 for the best). The human evaluation results are shown in Table 10. Overall, all the models obtain comparative scores for *Fluency*, and GPT-based models achieve higher *Relevance* scores than retrieval-based methods. Multimodal GPT outperforms textual GPT, which demonstrates that visual information is necessary for this task. Note that the kappa value (Fleiss, 1971) is 0.711, indicating a high consistency among different evaluators.

Models	Fluency	Relevance
Text-only Retrieval	2.97	1.25
Multimodal Retrieval	2.93	1.28
Text-only GPT	2.90	1.75
Multimodal GPT	2.94	1.91

Table 10: Human evaluation for dialogue response generation task.

5.5.2 Error Analysis

To further understand how our model fails in some cases, we analyze 100 unsatisfactory generated responses. We observe four main types of error.

- **Response with wrong objects (25%).** Some generated responses talk about a wrong object. For example, for a query “*The motor doesn’t*

work”, the model may generate “You can clear out the drains”.

- **Response with wrong aspects (26%).** Sometimes models will make mistakes in identifying a correct aspect. For example, when customers ask about “logistics”, the query is about “logistics time”, while the response is about “logistics expense”.
- **Over-generalization (23%).** Dialogue models may resort to general solutions without specific analysis of problems. For example, for various of damage, the model tends to “return the product”.
- **Meaningless response (26%).** The remaining error appear to generate meaningless text, such as “Wait a minute.” and “How can I help you?”.

Despite these errors, multimodal models have shown advantages over text-only models, and we believe that the multimodal dialogue tasks are worth further studying.

5.5.3 Case Study

From the relevance evaluation from Table 10, we can conclude that there are still many challenges to be solved in the multimodal dialogue response generation task. We show some cases in Figure 2.

For case 1, to generate a proper response, a dialogue model needs to recognize the product, *i.e.*, a *blender*, in the image and understand the cause of product failures. For case 2, there is a color difference in the product, but it does not affect usage. Generally, human customer service tends to compensate for it with money, while the models may be more inclined to return the product. This challenge requires the model to have more sophisticated dialogue strategies. For case 3, for the common question of how to adjust the power of the *induction cooker*, almost all the models can give correct answers, while occasionally a few customers ask how to set the temperature on the *induction cooker*, most of the models fail to give proper responses. For case 4, it needs to effectively model the context to accurately express the detailed information such as the style favored by customers in the example.

6 Conclusions and Future Work

In this paper, we construct a real-scenario multimodal multi-turn Chinese dialogue dataset named

JDDC 2.1⁵. Along with this dataset, we present four tasks, including the multimodal dialogue response generation task, the multimodal query rewriting task, the multimodal dialogue discourse parsing task, and the multimodal dialogue summarization task. For these tasks, we conduct experiments with text-only and multimodal baselines and show the necessity of visual information. However, there are still challenges in multimodal dialogue, such as understanding image details, effective modeling context information, long-tail questions, and refined dialogue strategy.

In the future, we plan to further explore the multimodal joint modeling methods and approaches to using product knowledge bases and image category information, and we will provide more detailed annotation information on the dataset in terms of user emotions, user semantic analysis, and so on. Furthermore, considering that all the annotations are shared with the same dialogue sessions, a promising research direction is to explore the interaction between these tasks, *e.g.*, to investigate whether discourse structure across modality is helpful for understanding the multimodal dialogue, whether discourse information can indicate the important utterances, and so on.

Limitations

A limitation of our work is that it is designed for only the Chinese language, the applicability to other languages remains to be verified. We hope our work can trigger future work on the multilingual dialogue dataset.

Second, our dataset only covers the product categories of small home appliances and fashion. Although they are the categories with relatively higher purchase frequencies, they may not reflect the real data distribution for the whole product categories. We will collect more data for other product categories in the future.

Third, to ensure a high quality of the dataset, we only select dialogues of customer service staff with gold medals, who tend to answer the questions more accurately. There may be some biases in the evaluation of the customer service staff, which will influence the quality of the dataset.

Finally, we collect our dataset from the conversation logs for one month. The dialogue contents may vary slightly across different times or seasons,

⁵Our dataset is publicly available at <https://github.com/hrlinlp/jddc2.1>





<p>Query: 电源插起来没反应, 是不是坏了。(It does not work when it is plugged in. Is it broken?)</p>  <p>Response: 杯子内加水。(Add water to the cup.)</p> <p>Multimodal GPT: 请问您是加水到刻度线之间的嘛?(Do you add water to the scale line?)</p> <p style="text-align: center;">Case 1</p>	<p>Query: 我买的一双42.5的, 发现鞋后跟颜色不一样。(I bought a pair of shoes in size 42.5 and found that the heel color was different.)</p>  <p>Response : 您发的照片这边已看到了, 非常抱歉给您造成困难, 我们也会积极为您处理的, 为了不耽误您穿着, 如果您愿意留下的话, 这边可以给您申请20元补偿可以么。(I have seen the photo you sent. I'm very sorry for that trouble. We will take care of it actively. In order not to delay your dressing, if you want to stay, you can apply for 20 Yuan compensation here.)</p> <p>Multimodal GPT: 这边先看下您发的图片哈。(We will take a look at the photo you sent.)</p> <p style="text-align: center;">Case 2</p>
<p>Query: 能调瓦数, 不能调温度哦。(The wattage can be adjusted, but the temperature cannot be adjusted.)</p>  <p>Response: 您好, 该款商品在火锅档只能调瓦数, 不能调温度的。(Hello, this product can only adjust the wattage but not the temperature in the hot pot.)</p> <p>Multimodal GPT: 尊敬的客户您好, 请问有什么可以帮到您的呢?(Dear customer, how can I help you?)</p> <p style="text-align: center;">Case 3</p>	<p>Query: 我买男款, 还是女款的? 女款的贵10块钱, 是不是要好一点。(Should I buy men's or women's versions? Women version are 10 yuan more expensive, isn't it better)</p>  <p>Response: 您好, 建议您选购正确款式哦。男女版型是不同的。(Hello, I suggest you choose the correct style. Male and female versions are different)</p> <p>Query: 有什么不同? 我看这都一样啊 我男的 (What's the difference? I think it's all the same.)</p> <p>Response: 建议您选购男款哦! (It is recommended that you choose men's version.)</p> <p>Query: 我咋觉得女款好看点呢? (But I think the women version is more attractive.)</p> <p>Response: 😊 您也是可以收到后试穿 按实际穿着的为准哦。(You can also try it on after receiving it, according to what you actually wear.)</p> <p>Multimodal GPT: 男款女款都是可以的哦。(Both male and female versions are OK.)</p> <p style="text-align: center;">Case 4</p>

Figure 2: Four typical examples in JDDC 2.1.

especially for the product category of fashion. We are planning to extend to a longer time span.

Ethical Statement

Privacy Protection

The construction of this dataset is licensed by the service platform and complies with the platform's agreement with merchants and customers that the data can be applied to non-profit service quality improvement research. To protect the private information of merchants and customers in the conversation, we have desensitized the data. Product-related information, including store names, product brands, and model numbers, and customer-related information, including names, addresses, phone numbers, and order numbers, is all masked. Specifically, we first use OCR to detect these information, and mask the recognized regions automatically. Among total 850,524 images, 161,459 are detected and masked. For the remaining images, we employ human annotators to check whether private information is in the image. As a result, 6,979 images are detected and masked manually. The annotation interface for data desensitization is shown in Appendix E.

Intended Use

We expect the JDDC 2.1 dataset can advance research in multimodal dialogue systems. We strictly restrict the use of our dataset to academic research.

Furthermore, we are looking forward to defining new tasks based on our dataset.

Labor Compensation

We employ crowdsourcing workers to complete the multimodal query rewriting, discourse parsing, and summarization tasks. Annotators are compensated at a rate of 2.55 Yuan per session. The annotation for each session takes 200 seconds on average, and thus crowdsourcing workers are paid 45.9 Yuan per hour, higher than the local living wage (25 Yuan per hour).

For data desensitization manual verification, 10,000 images can be checked per day per person, taking 87 man-days for all images. Masking the 6,979 images with privacy information takes 7 man-days. Ultimately, crowdsourcing workers are paid 60.3 Yuan per hour.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600. We thank the anonymous reviewers for their helpful comments and suggestions.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: visual question answering](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.
- Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. [Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727.
- Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. [Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.
- Anaís Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. [Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368.
- Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. [LIUM-CVC submissions for WMT17 multimodal translation task](#). In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. [Doubly-attentive decoder for multi-modal neural machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.
- Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [The JDDC corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service](#). In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 459–466. European Language Resources Association.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017a. [Visual dialog](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.
- Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. [Learning cooperative visual dialog agents with deep reinforcement learning](#). In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Günes Erkan and Dragomir R Radev. 2004. [Lexrank: Graph-based lexical centrality as salience in text summarization](#). *Journal of artificial intelligence research*, 22:457–479.
- Joseph L Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological bulletin*, 76(5):378.
- Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. [Mscap: Multi-style image captioning with unpaired stylized text](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4204–4213.
- 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*, pages 770–778.
- Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. [CUNI system for the WMT18 multimodal translation task](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. [SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations](#). *CoRR*, abs/2104.08667.

- Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. [Aspect-aware multimodal summarization for chinese e-commerce products](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8188–8195.
- Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018a. [Multi-modal sentence summarization with modality attention and image filtering](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.
- Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. [Multi-modal summarization for asynchronous collection of text, image, audio and video](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.
- Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018b. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):996–1009.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. [Multimodal sentence summarization via multimodal selective encoding](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018c. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. [A diversity-promoting objective function for neural conversation models](#). In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. [A persona-based neural conversation model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.
- Yang Liu and Mirella Lapata. 2019. [Text summarization with pretrained encoders](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems](#). In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. [Hierarchical question-image co-attention for visual question answering](#). *Advances in neural information processing systems*, 29:289–297.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. [Training millions of personalized dialogue agents](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2775–2779. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. [Image-grounded conversations: Multimodal context for natural question and response generation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 462–472. Asian Federation of Natural Language Processing.
- Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. [Large-scale pretraining for visual dialog: A simple state-of-the-art baseline](#). In *European Conference on Computer Vision*, pages 336–352. Springer.
- Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. [X-linear attention networks for image captioning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980.
- 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.
- Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. [Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*,

- November 16-20, 2020, pages 930–940. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*.
- Alan Ritter, Colin Cherry, and William B. Dolan. 2011. [Data-driven response generation in social media](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 583–593. ACL.
- Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. [Towards building large scale multi-modal domain-aware conversation systems](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 696–704. AAAI Press.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.
- Zhouxing Shi and Minlie Huang. 2019. [A deep sequential model for discourse parsing on multi-party dialogues](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7007–7014.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. [Image-chat: Engaging grounded conversations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2414–2429. Association for Computational Linguistics.
- Alessandro Sordani, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. [A neural network approach to context-sensitive generation of conversational responses](#). In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205. The Association for Computational Linguistics.
- Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. [A dataset for research on short-text conversations](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 935–945. ACL.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. [Show, attend and tell: Neural image caption generation with visual attention](#). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.
- Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. [K-plugin: Knowledge-injected pre-trained language model for natural language understanding and generation in e-commerce](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1–17.
- Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. [Stacked attention networks for image question answering](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Peng Yuan, Haoran Li, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [On the faithfulness for e-commerce product summarization](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5712–5717.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.
- Nan Zhao, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. [The jdcd 2.0 corpus: A large-scale multimodal multi-turn chinese dialogue dataset for e-commerce customer service](#). *arXiv preprint arXiv:2109.12913*.
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. [A dataset for document grounded conversations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jijun Zhang, and Chengqing Zong. 2018. [MSMO: Multimodal summarization with multimodal output](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9749–9756.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Multimodal joint attribute prediction and value extraction for e-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2139.

A Further Statistics for the JDDC 2.1 Dataset

Figure 3 demonstrates statistics of dialogue turns. The average number of dialogue turns in JDDC 2.1 is 14.06, while there is a large long-tail distribution for cases where some users interact with customer service with relatively more turns. Figure 4 shows statistics of the number of images. We can see that, in most cases, customers only use one or two images in conversations. Figure 5 illustrates statistics of dialogue utterance length. Users tend to use short sentences of about 20 characters to describe the problems they encounter, while customer assistants sometimes prefer longer sentences.

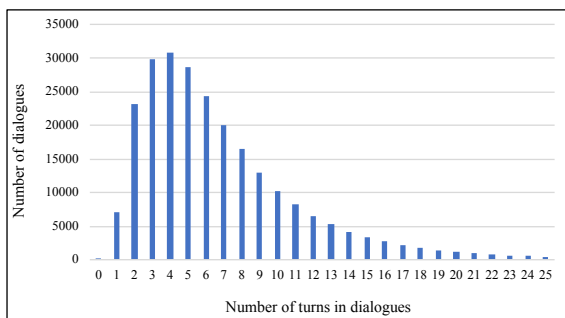


Figure 3: The statistics of dialogue turns.

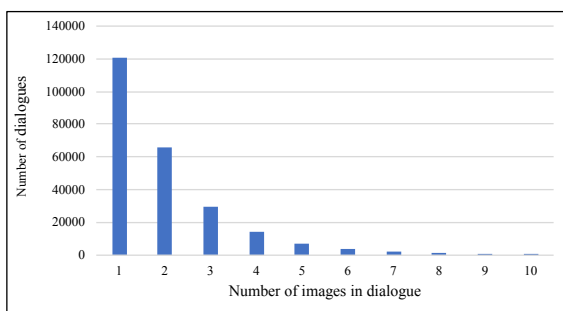


Figure 4: The statistics of numbers of image.

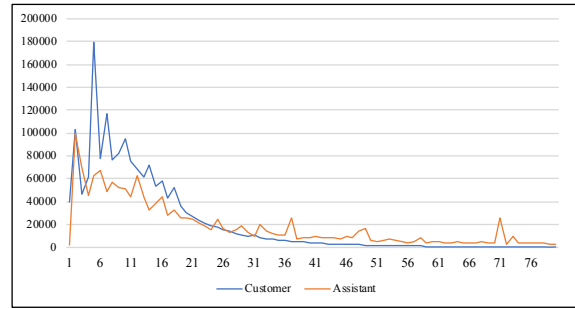


Figure 5: The statistics of dialogue utterance length.

B Statistics for Image Category

We randomly select 5,000 dialogue sessions from the product category of small home appliances, of which the user questions contain 8,218 images, and 5,000 from fashion, of which the user questions contain 8,731 images. We annotate these images according to the sub-categories, which are also be released as part of our dataset. The results are shown in Figure 6. We can find that the top-3 sub-categories for small home appliances are: 34% are photos for purchasing consultation, 18% are photos of products with the malfunction, and 13% are screenshots of products. The top-3 sub-categories for fashion are: 20% are photos of products with damaged appearance, 19% are screenshots of after-sales service order, and 14% are screenshots of products. The classification results show that, for small home appliances, the questions of users are mostly concentrated in pre-sales consultation and after-sales product usage. While for fashion, users are mainly concerned about products with damaged appearance.

C Statistics for Knowledge Base

The statistics of the knowledge base in our dataset are shown in Table 11.

D Distribution of Discourse Relations

The distribution of three-level dialogue discourse relations defined in our JDDC 2.1 dataset is shown in Table 12.

E Annotation Interface

To ensure the privacy information is completely masked, we employ human annotators to check whether private information is in the image, and the annotation interface is shown in Figure 7.

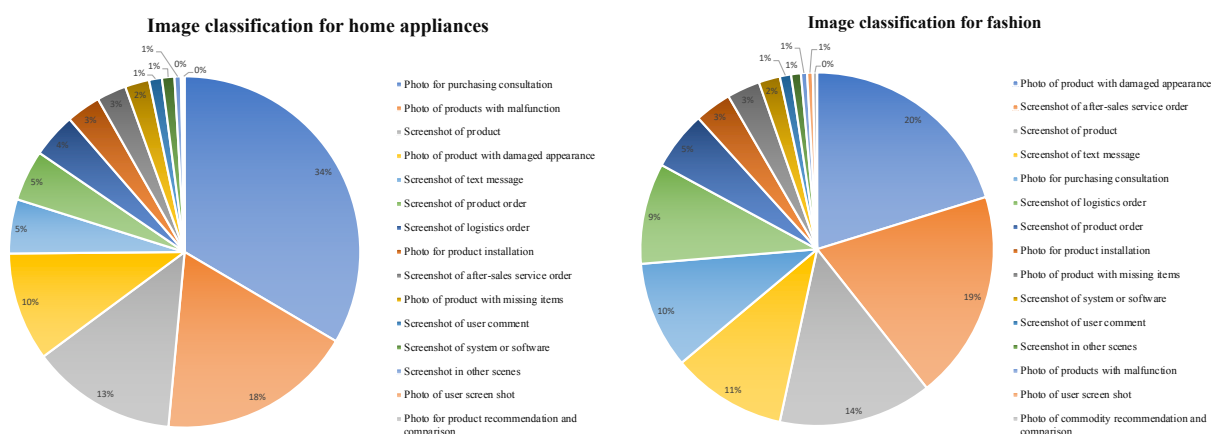


Figure 6: Image classification for 15 sub-categories.

Item	Number	Explanations
Entity	30,205	Products mentioned in the dialogue
Entity type	231	E.g., Rice cooker, air fryer, basketball shoes, jeans
Relation	759	E.g., Anti-dry function, removable basket, upper material, waist type
Triple	219,121	Detailed description of product attributes and product selling points

Table 11: The statistics for knowledge base.

1st Class	2nd Class	3rd Class	Explanations	Number of instances		
				Train	Valid	Test
Question&Answer	Question	Yes/No	Questions with WHETHER	4,148	268	237
		SWIH	Questions with WHO, WHAT, WHEN, WHERE, WHY, and HOW	3,202	205	170
		Choice	Questions for choosing a correct answer	150	5	7
	Answer	Affirmation	Positive answer to a question, it can be a positive word or with further explanations	3,868	236	221
		Deny	Negative answer to a question, it can be a negative word or with further explanations	856	54	47
Hold		Reply to a question, but no explicit answer is given	265	4	15	
Imperative	Suggestion	Recommend	Recommend some products	271	13	10
		Suggestion	Suggestion for some actions	828	33	33
		Request	Request for some actions	1,730	94	94
	Rely to suggestion	Acceptance	Accept some suggestions	1,087	54	60
		Rejection	Reject some suggestions	105	5	5
Promise	Promise	Promise	Make promise to do something	469	25	18
	Rely to promise	Acceptance	Accept some promise	94	2	4
		Rejection	Reject some promise	3	0	0
Communication	Social	Social	Say hello, greet, introduce yourself, apologize, thank you, bye	4,349	245	242
	Time	Suspend	Interrupt the current dialogue session	718	79	33
	Expansion	Correction	Correct the errors in another EDU	135	5	7
		Completion	Provide additional information towards another EDU	204	12	18
	Contact	Contact	Ask whether somebody is online	505	21	21
		Confirm	Confirm somebody is online	238	11	12
Feedback	Feedback	Interjection for response, e.g., "Mm-hmm", "well"	2,695	130	147	
Contingency	Contingency	Elaboration	Further explain the details in another EDU	6,751	315	378
		Summary	Summarize the key points in another EDU	53	1	0
		Restatement	Say something again about another EDU	1,166	87	86
		Background	Introduce the background for another EDU	1,516	64	73
		Evaluation	Evaluation and feeling towards another EDU	799	26	33
		Proof	Proof for another EDU	51	5	1
		Cause	Cause or result for another EDU	473	32	32
		Condition	Condition of another EDU	13	1	1
		Concession	Contrast, opposition, concession for another EDU	1,082	62	62
		Solution	Solutions to phenomena or faults in another EDU	319	8	17
		Purpose	Purpose or motivation of another EDU	52	3	3
		Coordinate	On a par with another EDU	669	64	53
		Total				38,864

Table 12: The distribution of three-level dialogue discourse relations in JDDC 2.1.

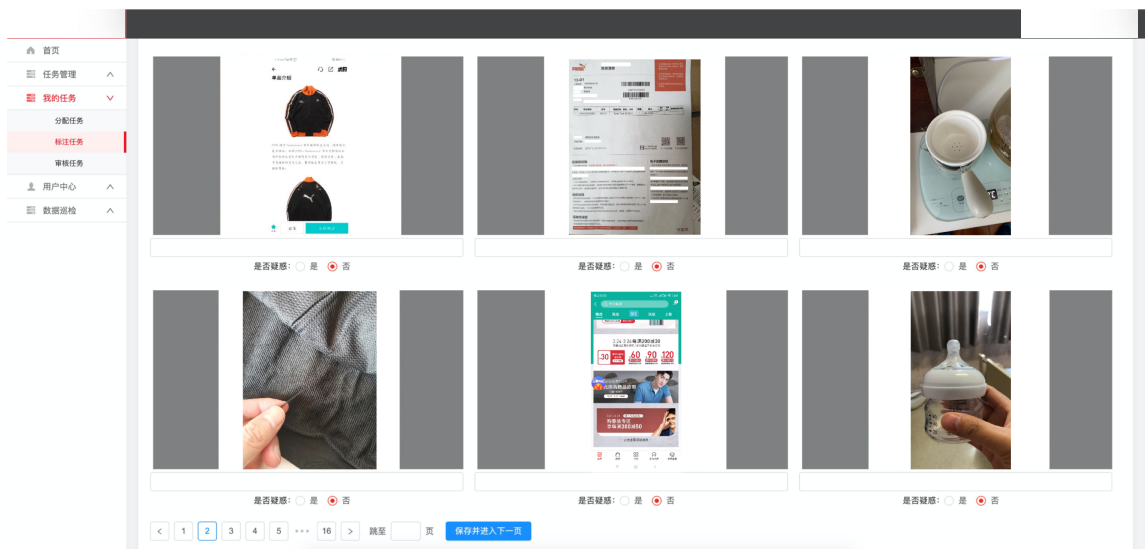


Figure 7: Annotation interface for data desensitization.