Improving Situated Conversational Agents with Step-by-Step Multi-modal Logic Reasoning

Yuxing Long[1] [*], **Huibin Zhang**[1] [*], **Binyuan Hui**[1] [*], **Zhenglu Yang**[2], **Caixia Yuan**[2], **Xiaojie Wang**[2], **Fei Huang**[1], **Yongbin Li**[1] [†]
[1] DAMO Academy, Alibaba Group, [2] Independent Researcher
{longyuxing.lyx,zhanghuibin.zhb,binyuan.hby,f.huang,shuide.lyb}@alibaba-inc.com

Abstract

To fulfill complex user requirements in a situated conversational scenario, the agent needs to conduct step-by-step multi-modal logic reasoning, which includes locating objects, querying information and searching objects. However, existing methods omit this multi-step procedure and therefore constitute the risk of shortcuts when making predictions. For example, they may directly copy the information from the dialogue history or simply use the textual description without visual reasoning. To address this issue and further boost the system performance, we apply the dual process theory to plug a reasoner into the original transformer-based model for step-by-step reasoning. When system 2 completes multi-step reasoning, its output is regarded as the final prediction. Our proposed method achieved the 1st rank on the summing scores across all four DSTC-11 SIMMC 2.1 sub-tasks.

1 Introduction

A situated conversational agent (Moon et al., 2020) engages in a conversation in an embedded multi-modal context, which may involve a virtual environment, a real physical world, etc. This task features language understanding, visual perception, decision feedback, and abundant interactions and reasoning among various modalities. Recently, (Kottur et al., 2021) proposed virtual shopping scenarios based on fashion and furniture, i.e. SIMMC, and it becomes a popular benchmark for the situated conversational task.

Current efforts (Lee et al., 2022; Lee and Han, 2021; Nguyen et al., 2021) on SIMMC target this task by designing unique model structures, auxiliary losses, and pre-training strategies. Although their approaches could improve the model's performance in predicting metadata related answer, such improvements partially come from shortcuts that have been revealed in recent studies (Ye and Kovashka, 2021; Dancette et al., 2021; Si et al., 2022; Ye et al., 2023). These shortcuts include directly copying keywords from dialogue history or making predictions simply based on textual mappings without performing visual reasoning. Apparently, shortcuts could harm the model performance especially for complex user requests. Instead, explicitly modeling the multi-step reasoning chain can provide a clear clue to help the model yield more reliable predictions. For instance, as shown

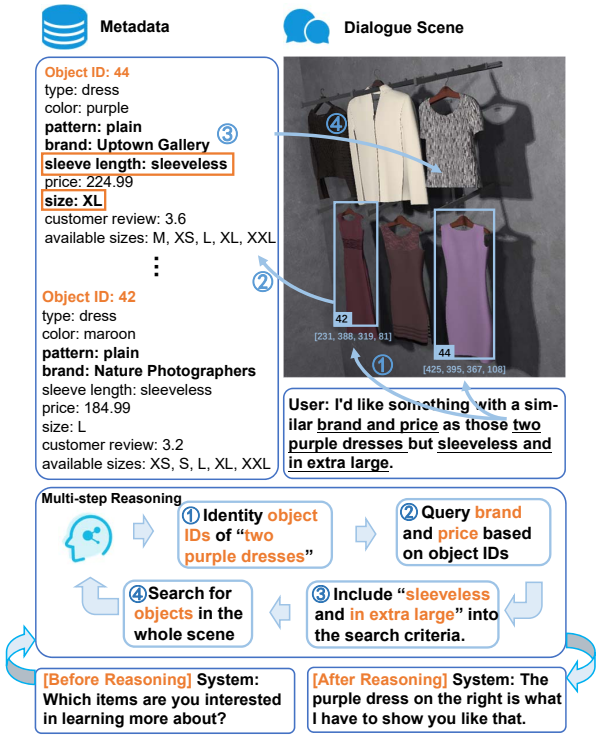

Figure 1: The multi-modal multi-step reasoning process for complex user requirements in the real world. Numbers ❶❷❸❹ represent the sequence of reasoning steps. To make the response to user, the situated conversational agent needs to identify objects on the scene image, query information from metadata, add extra user requirements and search for the target object in the whole scene. Without such a reasoning process, the agent fails to give an effective answer.

[*] Equal Contribution
[†] Corresponding Author

Proceedings of The Eleventh Dialog System Technology Challenge, pages 15–24

in Fig. 1, when the agent responds to a user query, *I'd like something with a similar brand and price as those two purple dresses but sleeveless and in extra large*, it needs to go through the subsequent step-by-step deliberation process. Step ❶, clarify the user's reference, i.e., identify the object id of the `two purple dresses`, based on the dialogue history and the current visual scene. Step ❷, check the corresponding brands and prices for the `two purple dresses` based on their object ids. Step ❸, include the newly stated requirements of the user in the search criteria: `sleeveless` and `extra large`. Step ❹, the eligible objects in the current visual scene are identified by incorporating the information from our query and the new requirements stated by the speaker.

Additionally, according to the *Dual process theory* (Evans, 2003), human cognition is produced by the interaction of ***System 1***, which is in charge of quick, unconscious, implicit judgments, and ***System 2***, which is in charge of complex, explicit, step-by-step reasoning. The cognitive ability of humans is mediated by the interaction of *System 1* and *System 2*. However, current situated conversational agents solely rely on intuition to derive answers from straightforward processing of input data (analogous to *System 1*). These responses are frequently inaccurate when the questions are complex because of the lack of interlocking logical reasoning (analogous to *System 2*). Therefore, we argue that the acquisition of the reasoning capabilities performed by humans in *System 2*, especially step-by-step multi-modal reasoning, is a crucial component of situated conversational agents.

To this end, inspired by cognitive theory, we propose a step-by-step multi-modal reasoning framework. Concretely, in our framework, **System 1**'s implicit understanding is handled by the conventional encoder-decoder, while the explicit step-by-step reasoning of **System 2** is handled by the newborn reasoner. Besides, we propose a simple, unsupervised step-by-step reasoning process generation approach based on rewriting the original user utterance as a learning signal for the reasoner.

We evaluate our model on the 11th Dialog System Technology Challenge (DSTC-11) SIMMC 2.1 (Kottur et al., 2021) track. SIMMC 2.1 includes four sub-tasks of situated conversational task, i.e., ambiguous candidate identification (Task 1 Ambigu.), multi-modal coreference resolution (Task 2 MM-Coref), multi-modal dialog state tracking (Task 3 MM-DST) and multi-modal dialog response generation (Task 4 Res. Gen.). Notably, after the official evaluation of the DSTC-11 committee, **our proposed model achieved the 1st rank on the summing scores across all four sub-tasks**.

2 Related Work

2.1 Situated Conversational Agents

Situated conversational agents aim at holding a meaningful conversation with humans based on natural images or real visual scene (Chen et al., 2020c). Recently, the flourishing of visual and language representation learning poses great demand for multi-modal dialog systems. Several datasets have been proposed to evaluate multi-modal dialog system, e.g., VisDial (Das et al., 2017), MMD (Saha et al., 2018), SIMMC (Moon et al., 2020), JDDC (Chen et al., 2020b) and JDDC 2.0 (Zhao et al., 2021). These datasets, however, are image-oriented dialogue systems without a comprehension of real-world visual scenes. To address the aforementioned issue, SIMMC 2.0 and 2.1 (Kottur et al., 2021) have been proposed, which are datasets for situated interactive multi-modal dialogue system and serve as official datasets for DSTC-10 and DSTC-11 challenge respectively. To solve this challenge, (Kottur et al., 2021) proposes a multi-modal GPT-2 (Radford et al., 2019) model as task baseline to generate current belief states and system response auto-regressively. (Huang et al., 2021) suggested using the rich multi-modal input and the pre-trained UNITER (Chen et al., 2020d) model to predict user mentioned objects. (Lee et al., 2022) encodes multi-modal context by special tokens and designs auxiliary tasks to predict object metadata so that the single-stream model can better utilize this information to respond to user requirements. All existing methods make efforts to improve representations of scene objects and fluency of agent response, but they ignore modeling logic in the multi-modal multi-step reasoning. Therefore, these methods often make mistakes when dealing with complex user requirements. Unlike them and (Long et al., 2023; He et al., 2022c,a,b), we first notice the step-by-step multi-modal reasoning based on cognitive theory in situated conversational agents.

2.2 Multi-modal Reasoning

Widely used multi-modal datasets, such as VQA (Goyal et al., 2017), VisDial (Das et al., 2017), FVQA (Wang et al., 2018) and

KVQA (Shah et al., 2019), all contains complex multi-modal questions or user utterances which require multi-modal agent to conduct multi-step reasoning on the image, text or structural data to predict answers. For VQA task and VisDial task, existing models, like ReDAN (Gan et al., 2019), DMRM (Chen et al., 2020a) MUREL (Cadène et al., 2019), GoG (Chen et al., 2021), design special modules or graphs to increase multi-modal agent's attention on the target image objects step-by-step under user guidance. In fact, these methods just complete visual grounding of objects by understanding the multi-hop keywords in textual instruction. when dealing with multiple actions in context, they are not equipped with abilities of multi-step logic reasoning. For FVQA task, (Zhu et al., 2020) builds convolutional graph network for retrieved facts and performs graph multi-step reasoning through message passing scheme utilizing graph convolution. For KVQA task, (Heo et al., 2022) constructs hypergraphs with multi-modal multi-hop information and creates Hypergraph Transformer to predict answers. Although their methods conduct multi-step logic reasoning, their reasoning actions are limited to querying information. Besides, the reasoning chain of all these methods is invisible to users, which makes their prediction result lack of interpretation. Compared with their method, our approach can generate a detailed multi-step logic reasoning process dealing with multiple actions according to user requirements, which is more interpretable and can be easily plugged into existing model architecture.

3 Task Descriptions

Ambiguous Candidate Identification (Ambigu.) During the conversation, the user's expressions tend to be more colloquial and accompanied by ambiguous object references. The ambiguous candidate identification task focuses on identifying all possible candidates that match the user's description. (*e.g.* U:"Can you tell me how much the brown piece costs? I'd like the rating too." ⇒ All of the brown objects in the current scene need to be predicted by the model.).

Multi-modal Coreference Resolution (MM-Coref) Similar to the ambiguous candidate identification task, the user will refer to specific items in the current scene when they are expressing their meaning accurately. Based on the current dialogue context and visual scene, the task requires the model to anticipate which object the user is referring to.

Multi-modal Dialog State Tracking (MM-DST) Dialogue state tracking task requires the model to track the speaker's belief state cumulative across multiple turns. In addition to comprehending the dialogue history, the MM-DST task requires the model to reason about the objects and their attributes in the current scene. (*e.g.* U: "Do you have something in a similar brand as that grey coat" ⇒ It is necessary to predict the object id and brand attribute of that grey coat.)

Multi-modal Dialog Response Generation (Res. Gen.) Similar to text-driven dialog response generation, this task requires the model to generate assistant responses based on the dialog history and the current scene.

4 Method

Our model contains a dual system, similar to the human cognitive system, consisting of System 1 and System 2. System 1 consists of a conventional Encoder-Decoder, which is dedicated to a specific task. System 2 consists of a generative Reasoner capable of producing step-by-step multi-modal reasoning path generation.

4.1 System 1: Encoder-Decoder

The design of System 1 is followed by (Lee et al., 2022) and it can perform SIMMC 2.1 sub-tasks end-to-end by unique task loss. Here we describe the architecture of System 1 in detail.

Encoder As Figure 2 shows, the encoder takes multi-modal information as input. The textual input includes dialogue history, current user utterance and special tokens, e.g., `"[act]"`, `"[obj]"` and `"[type]"`. The visual input of the encoder is all scene object region features encoded by pretrained models, e.g., ResNet-50 (He et al., 2016), which are projected to visual embedding with the same size as textual embedding. The projected bounding box coordinates are added to object visual embedding to represent spatial relations. Besides, similar to (Lee et al., 2022), we add auxiliary tasks for encoder to enhance the understanding of visual and non-visual attributes.

For predicting user act, user mentioned object IDs and slot values, the additional heads on special tokens as classification layer, which are optimized by cross-entropy losses.

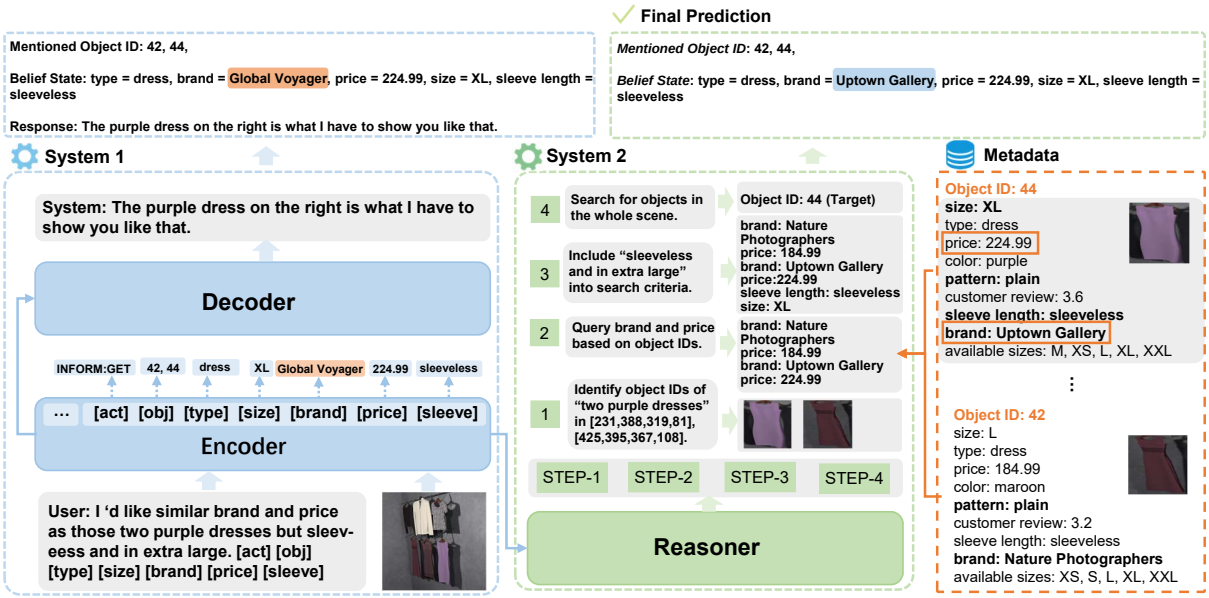

Figure 2: Overview of our proposed framework. We reuse the example in Figure 1. According to the user utterance, the encoder predicts the belief state and intent action of user, and the reasoner generates the detailed reasoning steps of situated conversational agents. The Decoder is in charge of generating system response based on the current multi-modal context. In the inference stage, the output of the reasoner is regarded as the final prediction if it completes multi-step reasoning.

Decoder The decoder takes hidden representation H from the encoder as input and generates output auto-regressively. For response generation, the response decoder is optimized by the standard left-to-right language modeling loss $\mathcal{L}_{\text{dec}}$ as the following.

$$\mathcal{L}_{\text{dec}} = \sum_{i=1}^{L} -\log P_{\theta_{dec}}\left(\omega_{dec}^{i} \mid \omega_{dec}^{1}, \ldots, \omega_{dec}^{i-1}, H\right) \tag{1}$$

where H denotes the last hidden states of encoder, ω_{dec}^{i} represents i-th target response token, L_{dec} is the total length of target response and θ_{dec} means decoder learnable parameters.

4.2 System 2: Reasoner

The aim of system 2 is to generate a concrete inferring process step-by-step and implement corresponding actions on the scene image and metadata to get information about object IDs, bounding boxes and attributes. The core module of system 2 is a reasoner whose architecture is also a Transformer-based auto-regressive decoder. As Figure 2 shows, the reasoner takes encoded multi-modal representation from the encoder as module input and auto-regressively generates reasoning steps.

According to reasoning intention, the generated reasoning steps can be categorized into: **Locate Object**, **Query Information**, **Add Requirement**, **Search Object**, **Compare** and **Disambiguate**. **1)** In the step of locating objects, the reasoner extracts object detailed description from the input text and then predicts corresponding bounding box coordinates on the scene image, which can be used to crop object region visual features. The cropped region feature is compared with object screenshots in the metadata to identify IDs of objects mentioned in the current user utterance. **2)** In the step of querying information, the reasoner extracts user required attribute types, such as *"brand"* and *"price"*, about mentioned objects from inputted text and queries metadata values based on identified object IDs and attribute types. These queried values are regarded as search requirements. **3)** In the step of adding requirement, the reasoner extracts user extra requirements attribute values, like *"sleeveless"* and *"in extra large"*, from inputted text and adds them into search requirements. **4)** In the step of searching objects, the reasoner takes all search requirements as criteria and searches for objects satisfying all requirements among metadata. **5)** In the step of comparing, the reasoner retrieves all metadata attribute of object 1 and object 2 and compare their similarity and difference. **6)** In the step of disambiguation, the reasoner extracts object detailed description from the input text and transverses all metadata of scene objects to find out

REASONING TYPE	REASONING TEMPLATE
`Locate Object`	Identify object IDs of "**[Referring Expression]**" in **[x1, y1, x2, y2]**.
`Query Information`	Query **[Metadata Type]** of object **[Object ID]**.
`Add Requirement`	Include "**[Attribute Value]**" into search criteria.
`Search Object`	Search for objects in the whole scene.
`Compare`	Compare metadata of object **[Object 1 ID]** and **[Object 2 ID]**.
`Disambiguate`	Find all possible "**[Referring Expressions]**" in the whole scene.

Table 1: Six types of reasoning step templates designed for SIMMC 2.1. "**[*]**" is the template slots to be filled. By extracting information from user utterances and filling the templates, multi-step reasoning data can be generated as reasoner training data.

matched objects.

Reasoning Process Generation By analyzing complex user utterances in the multimodal conversation datasets (Kottur et al., 2021; Moon et al., 2020; Saha et al., 2018), we find that there are signal words, like *"similar"*, *"but"* and *"as"*, exist in the user utterance, which indicates the start of essential information. In this case, original user utterance can be divided into different reasoning semantics by these signal words. For example, user utterance *"I'd like similar brand and price as those two purple dresses but sleeveless and in extra large"* can be splitted into referring expression *"as those two purple dresses"*, information inquiry *"similar brand and price"* and new requirements *"but sleeveless and in extra large"*. Based on common reasoning semantics in the SIMMC 2.1, we design six types of reasoning step templates as Table 1 shows. After determining the template by signal words, regular expressions are utilized to extract attribute types and values from splitted reasoning semantics to fill template slots. In this way, a multi-step reasoning process can be generated from the original user utterance. If there are no signal words in the utterance, the generation target is set to *"No multi-step reasoning"*.

Reasoning Task The reasoner is optimized by the standard left-to-right language modeling loss $\mathcal{L}_{\text{rea}}$ as the following.

$$\mathcal{L}_{\text{rea}} = \sum_{i=1}^{L} -\log P_{\theta_{rea}}\left(\omega_{rea}^{i} \mid \omega_{rea}^{1}, \ldots, \omega_{rea}^{i-1}, H\right) \tag{2}$$

where ω_{rea}^{i} represents i-th target reasoning process token, L_{rea} is the total length of target reasoning process and θ_{rea} is reasoner learnable parameters.

After the reasoner executes all steps, it can obtain object IDs mentioned by current user utterance, user utterance slot values and object information to be responded by agent, which is collected as the output of system 2. Compared with system 1, system 2 is more reliable because system 2 completes the whole reasoning process explicitly instead of utilizing shortcuts to respond to complex user requirements. Therefore, in the inference stage, system 2 output is chosen as the final prediction result if its output exists.

5 Experiments

5.1 Dataset

The SIMMC 2.1 dataset follows the setting of SIMMC 2.0 (Kottur et al., 2021), which is geared towards building virtual assistant that generates conversations with users in the form of co-observation and immersive virtual reality (VR) environments. The conversation's content may take several turns, and the context is dynamically modified based on the user's activities at each turn. There are a total of 11244 dialogues and 1566 scenes gathered in the SIMMC 2.1 dataset, which uses a two-stage pipeline to collect dialogues (multi-modal dialogue simulation and manual interpretation). The dialogue of SIMMC 2.1 involves two domains: fashion (7.2k dialogs) and furniture (4k dialogs), which are randomly divided into four splits: train (64%), dev (5%), dev-test (15%), and test-std (15%). The test-std split is used as the evaluation dataset for the DSTC-11 Challenge[2] and its data annotation is not publicly available.

5.2 Evaluation Metric

In order to evaluate the performance of the multimodal dialogue system, SIMMC 2.1 employs different evaluation metrics for each subtask. For Ambiguous Candidate Identification task and Multimodal Coreference Resolution task, the performance is mainly evaluated by the F1 score of object

[2]https://dstc11.dstc.community

	Model	Task 1 Ambigu.	Task 2 MM-Coref	Task 3 MM-DST		Task 4 Res. Gen.
		Object F1 (↑)	Object F1 (↑)	Slot F1 (↑)	Act F1 (↑)	BLEU-4 (↑)
Baseline	GPT-2	18.00%	26.50%	73.50%	93.00%	19.20
	MTN	43.20%	–	75.00%	94.30%	21.00
	BERT	43.90%	–	–	–	–
DSTC-10	I2R Singapore	–	40.50%	86.40%	96.30%	34.62
	Sogang University	–	46.97%	88.25%	96.13%	28.38
	Kakao Enterprise	–	60.80%	–	–	28.50
	KAIST	–	73.50%	88.30%	96.30%	33.10
DSTC-11	**#2**	62.45%	73.78%	92.48%	97.30%	30.00
	#3	69.28%	–	–	–	–
	#5	68.46%	78.64%	92.05%	97.82%	39.09
	(Ours) #1	**70.30%**	**94.40%**	**93.32%**	**99.19%**	**42.55**
	- Reasoner	65.10%	79.65%	89.42%	96.53%	41.24

Table 2: Results on the SIMMC 2.1 devtest set. The first block shows the baselines, which are separately trained on each subtask. The second block and third block provide the complete results and ablation study on DSTC-10 & DSTC-11 model: "- Reasoner" means taking system 1 prediction as the final prediction under all cases without system 2 participation. Task 1 is the new task of DSTC-11, hence it isn't included in the result of DSTC-10.

prediction (Object F1). For the Multi-modal Dialog State Tracking task, the performance is mainly evaluated by the F1 score of dialog action and slot-values prediction (Slot F1 & Act F1). For the Multi-modal Dialog Response Generation task, the performance is mainly evaluated by the BLEU-4 for generated system response (BELU-4).

5.3 Baselines

The official SIMMC 2.1 proposer provided three baseline models, the finetuned **GPT-2** baseline generates the results of tasks 1 to 4 in a way similar to the language model. The **MTN** and **BERT** baselines combine visual scene information and dialogue history to generate user belief state and system response (these two models are not involved in all tasks). The visual features for the corresponding bounding boxes in each scene have been extracted by using pretrained ResNet-50 (He et al., 2016). Additionally, we have included the result from other DSTC-10 and DSTC-11 participating teams. In table 2, We display the teams with publicly available results due to certain teams still not publishing their performance on devtest set. In table 3 we display the official leaderboard of DSTC-11 & DSTC-10 on the teststd set of SIMMC 2.1.

5.4 Implementation Detail

For the encoder module in System 1, we employ the pre-trained longformer-base (Beltagy et al., 2020) model provided by HuggingFace's transformers library (Wolf et al., 2020). The decoder in System 1 and the reasoner in System 2 are designed as 6 hidden layers with a size of 512 and 8 attention heads. In our training stage, the batch size is set to 64 and the number of training epochs is set to 100. We introduce a cosine warm-up mechanism and set the warm-up rate and learning rate to 0.1 and 5e-5. The AdamW optimizer is used to optimize all models and all experiments are conducted with PyTorch. Our source code will be released online.

5.5 Experimental Results

5.5.1 Comparison Analysis

From the evaluation results in Table 2, it can be observed that our model achieves state-of-the-art performance on each task. In particular, the performance on multi-modal coreference resolution (Task 2) and response generation (Task 4) exceed the runner-up for a large margin (15.76 (78.64 $\rightarrow$ 94.40) in object f1-score and 6.30 (39.09 $\rightarrow$ 45.39) in Bleu-4 score), which demonstrates the effectiveness of introducing the step-by-step multi-modal reasoning strategy. By introducing the reasoning generation process of system 2, our model can better comprehend the coreference relationship and implicit reasoning logic in the dialogue. For the ambiguous candidate identification (Task 1) and multi-modal dialog state tracking task (Task 3), we apply a similar framework and achieve favorable performance. Additionally, in contrast to other teams which employ an autoregressive language approach to deal with the MM-DST, We first utilize the encoder module for processing. We also conducted an ablation analysis on the primary modules in the framework.

	Model Entry	Task 1 Ambigu.	Task 2 MM-Coref	Task 3 MM-DST		Task 4 Res. Gen.
		Object F1 (↑)	Object F1 (↑)	Slot F1 (↑)	Act F1 (↑)	BLEU-4 (↑)
DSTC-10	I2R Singapore	–	42.20%	87.80%	96.20%	25.60
	Sogang University	–	52.10%	88.30%	96.30%	28.50
	QS goal Diggers	–	56.40%	89.30%	96.40%	32.20
	Kakao Enterprise	–	68.20%	4.000%	41.40%	29.70
	Hariot-Watt University	–	68.20%	87.70%	95.80%	32.70
	New York University	–	73.30%	–	–	–
	KAIST	–	75.80%	90.30%	95.90%	29.50
	UCLA	–	78.30%	–	–	–
DSTC-11	**#2**	65.17%	–	–	–	–
	#3	63.84%	75.85%	90.48%	96.77%	30.29
	#4	–	–	–	–	25.19
	#5	**70.50%**	80.28%	92.66%	**97.75%**	36.50
	(Ours) #1	67.26%	**94.29%**	**94.24%**	95.98%	**40.93**

Table 3: The official leaderboard of DSTC-11 & DSTC-10 on the teststd set of SIMMC 2.1. The first block shows the results of DSTC-10 and the second block shows the results of DSTC-11. The task winners are bold-faced and runner-ups are marked with underline.'-' means that the entry did not participate in that task. Task 1 is the new task of DSTC-11, hence it isn't included in the result of DSTC-10. Our entry is officially announced as the Winner of Subtask 2,3,4 (MM-Coref, MM-DST, Res. Gen.) and Runner-up of Subtask 1 (Ambigu.)[1].

As indicated by the last row of Table 2, the removal of the reasoner weakens our model's performance on all sub-tasks. Especially, reasoner has a huge influence on the Task 1 and Task 2. The reason may lay on the fact that these two tasks require the model to ground all mentioned objects or possible referred objects in the scene image, which leads to the failure of shortcuts. The ablation study reflects the significance of the step-by-step multi-modal reasoning for the multi-modal dialog system.

For the official evaluation, we generate the evaluation file in accordance with the competition settings[3]. As shown in table 3 our model achieved the 1st rank when summing the metrics across all four sub-tasks. We withheld the names of the DSTC-11 participating teams in order to comply with the anonymity requirement.

5.5.2 Human Evaluation

The human evaluation for response generation mainly focuses on 4 aspects: **fluency**, **relevance**, **correctness**, and **informativeness**, which are important for task-oriented dialogue systems. First, We randomly select 500 dialogues from SIMMC 2.1 dev-test dataset as candidates. According to SIMMC rule, only the last response of each dialogue in devtest is evaluated. Therefore, we choose the last responses of these selected 500 dialogues generated by KAIST, I2R Singapore, and our model. We release human evaluation task on Amazon Mechanical Turk (AMT) platform and hire 10 evaluators with different AMT ID. Every evaluator is assigned with these 1500 randomly shuffled evaluation data samples. Each data sample contains complete GT dialogue, scene images and last generated response without the model name. Evaluators need to score from four aspects on a scale of 1-5. The final score of each aspect of each model is the average score of 10 evaluators on 500 responses generated by corresponding model on this aspect. To guarantee the effectiveness of our human evaluation result, we respectively calculate Fleiss Kappa scores for fluency, relevance, correctness, and informativeness, which are 0.9077, 0.9189, 0.8813, 0.8698. As shown in 3, it can be observed that our model consistently outperforms

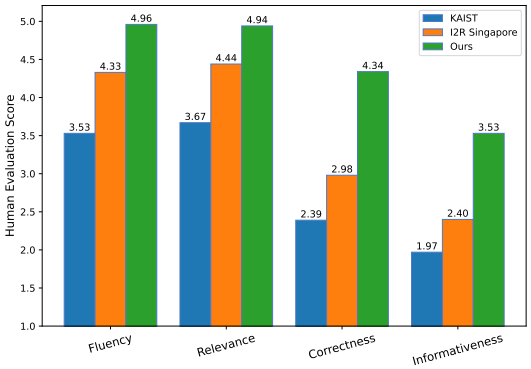

Figure 3: The human evaluation results on SIMMC 2.1 from four aspects. Our model displays significant improvement on **correctness** and **informativeness**.

[3]https://github.com/facebookresearch/simmc2

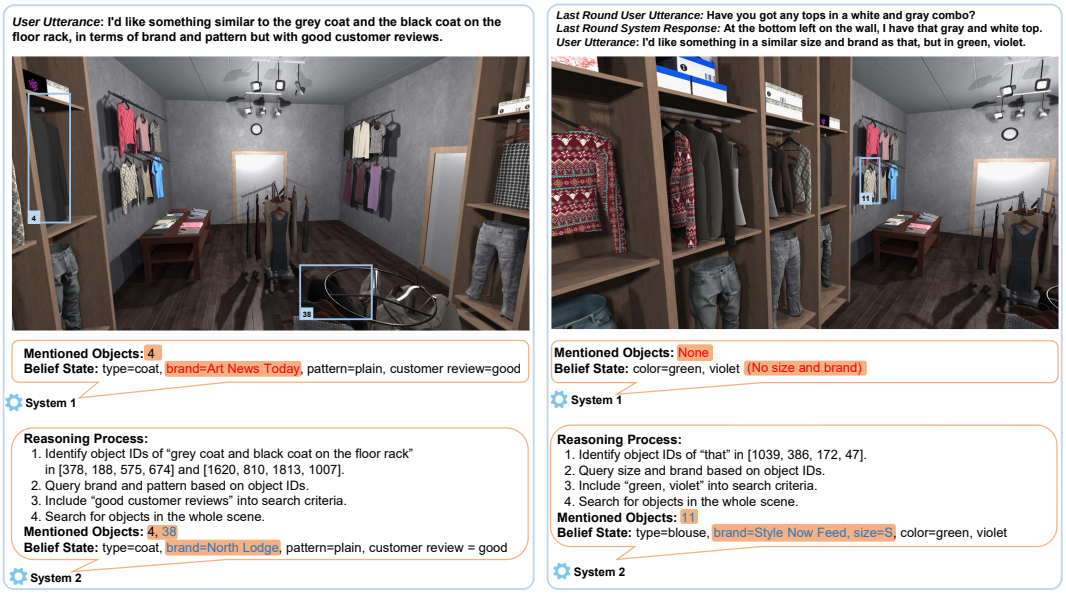

Figure 4: The left case: in which system 1 fails to predict user mentioned "*black coat*" and slot value of the brand while system 2 conducts a complete reasoning process and predicts the correct mentioned object ID "38" and brand slot value "*North Lodge*". The right case: in which system 1 fails to resolve user coreference "*that*" and misses slot values of size and brand while system 2 identifies "*grey and white top*" referred by "*that*" in the last round system response and predicts correct mentioned object ID "11", brand "*StyleNow Feed*" and size "*S*".

the other two models on all metrics, which is in line with automatic evaluation results.

5.5.3 Case Study

To better illustrate the advantage of our method on multi-step reasoning, we visualize several system responses to complex user utterances generated by our model and existing SOTA models.

As shown in the left part of Figure 4, system 1 fails to predict the corresponding object ID of "*black coat*" and therefore directly uses the brand of "*grey coat*" and copies pattern "*plain*" and customer review "*good*" in the user utterance to answer. Compared with system 1, system 2 successfully identifies object IDs of "*grey coat*" and "*black coat*" and conducts multi-step reasoning to obtain the correct brand of target object. As shown in the right part of Figure 4, system 1 is confused about coreference word "*that*" and unable to link it to "*grey and white top*" in the dialogue history. In this case, system 1 simply judges there is no mentioned object in the current the user utterance and ignores brand and size slots. It only predicts color slot by copying user requirement about color "*green, violet*". Compared with system 1, system 2 successfully links "*that*" to "*grey and white top*" and predicts its IDs, brand and size by multi-step reasoning.

From the above cases, it can be observed that: (1) our model is able to locate user mentioned objects more accurately on the scene image when more than one object exists in the user utterance. (2) our model has the ability to resolve corefence like "*that*", "*those*" and "*these*" in the dialogue history and link them to corresponding object IDs. (3) our model predicts slot values about complex user utterances through implementing reasoning actions step-by-step instead of directly copying information from the current user utterance or simply using the textual description without visual reasoning.

6 Conclusion

In this paper, we propose to enhance the ability of situated conversational agents by emphasizing step-by-step multi-modal reasoning. Specifically, the original user utterance is rewritten into a step-by-step reasoning process by introducing a knowledge- and semantic-based extraction approach. In light of dual process theory, we propose a step-by-step multi-modal reasoning framework, and transmit complicated multi-step reasoning and implicit multi-modal comprehension to different modules. Extensive experiments on benchmarks SIMMC validate the superiority of our framework. After the official evaluation of the 11th Dialog System Technology Challenge (DSTC-11) SIMMC 2.1 track, our model is ranked at the 1st of the summing score of all sub-tasks.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Rémi Cadène, Hédi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. MUREL: multimodal relational reasoning for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1989–1998. Computer Vision Foundation / IEEE.

Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021. Gog: Relation-aware graph-over-graph network for visual dialog. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 230–243. Association for Computational Linguistics.

Feilong Chen, Fandong Meng, Jiaming Xu, Peng Li, Bo Xu, and Jie Zhou. 2020a. DMRM: A dual-channel multi-hop reasoning model for visual dialog. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7504–7511. AAAI Press.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. 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.

Xiaofan Chen, Songyang Lao, and Ting Duan. 2020c. Multimodal fusion of visual dialog: A survey. In *RICAI 2020: 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China, October, 2020*, pages 302–308. ACM.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020d. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1554–1563. IEEE.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. 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.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Zhe Gan, Yu Cheng, Ahmed Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474, Florence, Italy. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. Space-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Space-3: Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022c. Space: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph Transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 373–390, Dublin, Ireland. Association for Computational Linguistics.

Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. 2021. Uniter-based situated coreference resolution with rich multimodal input. *CoRR*, abs/2112.03521.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4903–4912. Association for Computational Linguistics.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, and Kee-Eung Kim. 2022. Learning to embed multimodal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 813–830. Association for Computational Linguistics.

Joosung Lee and Kijong Han. 2021. Multimodal interactions using pretrained unimodal models for SIMMC 2.0. In *DSTC10 challenge wokrshop at AAAI*.

Yuxing Long, Binyuan Hui, Fulong Ye, Yanyang Li, Zhuoxin Han, Caixia Yuan, Yongbin Li, and Xiaojie Wang. 2023. SPRING: situated conversation agent pretrained with multimodal questions from incremental layout graph. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1103–1121. International Committee on Computational Linguistics.

Thanh-Tung Nguyen, Wei Shi, Ridong Jiang, and Jung jae Kim. 2021. Multimodal and joint learning generation models for simmc 2.0. In *DSTC10 challenge wokrshop at AAAI*.

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):9.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal 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.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8876–8884. AAAI Press.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. *CoRR*, abs/2210.04692.

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3181–3189. AAAI Press.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *CoRR*, abs/2301.13808.

Nan Zhao, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. The JDDC 2.0 corpus: A large-scale multimodal multi-turn chinese dialogue dataset for e-commerce customer service. *CoRR*, abs/2109.12913.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1097–1103. International Joint Conferences on Artificial Intelligence Organization. Main track.