Multimodal Recommendation Dialog with Subjective Preference: A New Challenge and Benchmark

Yuxing Long[1], **Binyuan Hui**[2], **Caixia Yuan**[1], **Fei Huang**[2], **Yongbin Li**[2] [*], **Xiaojie Wang**[1] [*]

[1] Beijing University of Posts and Telecommunications, [2] Independent Researcher
{longyuxing,yuancx,xjwang}@bupt.edu.cn, lyb821@gmail.com

Abstract

Existing multimodal task-oriented dialog data fails to demonstrate the diverse expressions of user subjective preferences and recommendation acts in the real-life shopping scenario. This paper introduces a new dataset SURE (*Multimodal Recommendation Dialog with SUbjective PREference*), which contains $12K$ shopping dialogs in complex store scenes. The data is built in two phases with human annotations to ensure quality and diversity. SURE is well-annotated with subjective preferences and recommendation acts proposed by sales experts. A comprehensive analysis is given to reveal the distinguishing features of SURE. Three benchmark tasks are then proposed on the data to evaluate the capability of multimodal recommendation agents. Based on the SURE, we propose a baseline model[†], powered by a state-of-the-art multimodal model, for these tasks.

1 Introduction

Building conversational agents that can communicate with people in multimodal situations is an attractive goal for the AI community. Many different tasks and datasets for the multimodal dialog have been proposed in recent years. Among them, Moon et al. (Moon et al., 2020) provided a multimodal task-oriented dialog dataset SIMMC 1.0 in two shopping domains. It aims to train interactive assistants which can handle multimodal inputs in a co-observed environment. The SIMMC challenge based on SIMMC 1.0 was held as part of DSTC9 (Gunasekara et al., 2021). Many works (Kung et al., 2021; Kim et al., 2021; Jeong et al., 2021; Huang et al., 2021; Senese et al., 2021) have been done following SIMMC 1.0. Since SIMMC 1.0 environment is simple and far from realistic stores, Kottur et al (Kottur et al., 2021) proposed SIMMC 2.0 with closer-to-real-world shop-

Figure 1: **Customer** *(left)* expresses subjective preferences. **Salesperson** *(right)* adopts different acts to clarify subjective preferences by multimodal context and recommend clothes from the situated store. Subjective preferences are highlighted by underlining while corresponding categorization concepts and candidate attribute values of scene items are shown in beneath.

ping scenarios, which was then used in DSTC10 challenge (Kottur1 et al., 2022).

Though these datasets facilitate research of conversational agents, they simplify some crucial problems in the real-life shopping dialog, which should be addressed for building multimodal recommendation agents.

In previous TOD datasets, most of the user descriptions for items (e.g., clothes or furniture) are referring expressions in the domains, such as *"the white couch chair"*, and *"the black hat in the middle of the long rack"*, which can be mapped to a

---

slot value without ambiguity. While in practice, customers are not experts in the domains. They use lots of different words to describe what they want, such as *"clothes designed for young ladies"*, and *"color makes me feel quiet and calm"* in Fig. 1. These words (or phrases) usually reflect customers' subjective cognition and preference for items they want. We call this kind of expression as **subjective preference**. To understand such expressions, the salesperson needs to map subjective preferences to standardized categorization concepts in the domains, and then use the concepts to filter candidate attribute values (slot values) of scene items. (Fig. 1 shows the two-step mappings under the subjective preferences). As we can see, such subjective preferences often correspond to a set of slot values instead of a unique one, which is very different from that in previous datasets. Facing this kind of customer requirement, a salesperson needs to communicate with the customers, utilize suitable strategies to narrow candidates progressively and give sound recommendations through multimodal context (Fig. 1 shows some recommendation acts, such as *Ask Preference*). In a word, understanding subjective preferences, finding a way to clarify the subjective preferences, and finally giving good recommendations is the challenge not depicted in both SIMMC 1.0 and 2.0. None of the existing multimodal dialogs research focuses on subjective preference and item recommendation.

To facilitate building conversational agents that can handle subjective preference and make shopping recommendations, we introduce a dataset for *Multimodal Recommendation Dialog with SUbjective PREference* (SURE). Specifically, we collect 12K salesperson ↔ customer goal-oriented recommendation dialogs in complex store scenes in two phases. Dialog flows were first generated by self-playing between the carefully designed customer and salesperson simulators, which helps to ensure the flows are reasonable. Crowd-sourcing is then employed for rewriting categorization concepts in dialog flows to diverse subjective preferences. The dataset contains well-annotated subjective preferences and diverse dialog acts proposed by experienced sales experts, which provides rich resources for evaluating subjective preferences understanding and dialog policies.

We then propose three benchmark tasks for evaluating multimodal recommendation agents' capability on subjective preference understanding and

item recommendation: **Subjective Preference Disambiguation**, **Referred Region Understanding**, and **Multimodal Recommendation**. We provide a baseline model for these tasks and highlight the key challenges and future research directions.

Our main contributions are:

- We introduce a large-scale multimodal dialogs data in two domains including $12K$ dialogs in complex scenes. The data is built in two phases with human annotations to ensure both dialog quality and language diversity.

- The data is well-annotated with subjective preferences and recommendation acts. Diverse acts and transition probabilities are obtained from the survey for sales experts.

- Three tasks are designed to evaluate the capability of multimodal recommendation agents. A strong baseline model MRA is proposed for these tasks.

2 Related Work

Textual Conversational Recommendation. The rise of e-books, music, and video websites has witnessed the development of conversational recommendation agents. Existing agents all operate in the textual modal, which elicits user preferences via conversation and recommends items based on dialog history and structured attribute data. The ReDial (Li et al., 2018) agent immediately recommends movies obtained from DBpedia (Auer et al., 2007) after the user expresses their preferences. TG-ReDial (Zhou et al., 2020) is developed by walking along ConceptNet (Speer et al., 2016) threads containing movies to collect user preferences and movie recommendations. GoRecDial (Kang et al., 2019) dialogs are collected by game-play to recommend target movies from candidate sets. DuRecDial (Liu et al., 2020) include Chinese dialogs between movie seeker and conversational bot based on the knowledge graph. INSPIRED (Hayati et al., 2020) focuses on how social strategies adopted by the agent influence the final success rate of recommendation. Although these works study conversational recommendation problems from different aspects including topic, strategy, and language, they are all based on textual modal. Therefore, conversational agents built on these datasets cannot equip the abilities required
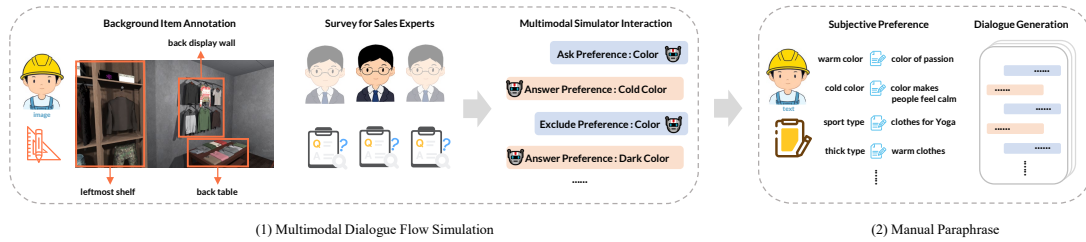
Figure 2: The annotation process of SURE dataset. (1) **Multimodal Dialog Flow Simulation**: we first hire annotators to label background items in the scene and do a survey for sales experts about attribute categorization and act. Then, we generate dialog flows by simulator interaction based on survey results. (2) **Manual Paraphrase**: annotators write subjective preferences based on attribute categorization concepts and paraphrase dialog flows.

in multimodal scenarios. Besides, all of these researches ignore developing agents to respond to subjective preferences frequently appearing in real recommendation dialogs.

Multimodal Shopping Dialogs. Developing multimodal conversational agent for shopping scenarios is significant for improving the quality of commercial service quality. MMD (Saha et al., 2018) establishes the first large-scale multimodal dialog dataset between shoppers and sales agents, which empowers conversational agents with abilities of multimodal understanding and querying. SIMMC (Moon et al., 2020) serves as a first step towards building task-oriented multimodal conversational agents with simple acts like informing information, confirming, and prompting. SIMMC 2.0 (Kottur et al., 2021) constructs more complex multimodal context with closer-to-real-world store scenes and introduces new challenges like multimodal co-reference resolution and multimodal dialog state tracking. The SIMMC challenge based on SIMMC 1.0 was held as part of DSTC9 (Gunasekara et al., 2021), there are many follows-up work (Kung et al., 2021; Kim et al., 2021; Jeong et al., 2021; Huang et al., 2021; Senese et al., 2021; Long et al., 2023). SIMMC 2.0 was used in the DSTC10 challenge (Kottur1 et al., 2022). These datasets facilitate research on multimodal conversational agents greatly. However, they still lack diverse expressions of user subjective preferences and recommendation acts.

3 SURE Dataset

We build SURE (*Multimodal Recommendation Dialog with SUbjective PREference*) dataset to facilitate research on multimodal recommendation agents. In SURE dialogs, customers express their preferences subjectively. To effectively recommend items, the salespersons perform: ❶ Actively elicit customer preferences about item attributes;

❷ Disambiguate subjective preferences according to the multimodal context; ❸ Narrow the candidate set of items based on dialog history and scene; ❹ Recommend target item from the situated scene.

To collect SURE dialogs, we design a two phrases pipeline (Fig. 2) following popular machine ↔ human collaborative dialog collection approaches (Kottur et al., 2021; Rastogi et al., 2020; Shah et al., 2018). In this section, we will introduce the SURE two phrases collection process in order and then analyze the distinguishing features of SURE.

3.1 Multimodal Dialog Flow Simulation

To generate dialog flows between salesperson and customer, we collect real-life shopping information first and then construct simulators based on this information.

3.1.1 Real-life Information Collection

We invite human annotators to label scene background items and experienced sales experts to complete questionnaires about attribute categorization and multimodal recommendation strategy.

Store Scene & Background Item Annotation We develop SURE based on 1566 scene snapshots in SIMMC 2.0 (Kottur et al., 2021). These snapshots come from 140 fashion stores and 20 furniture stores generated by Unity 3D. To utilize spatial relations between commodity items and background items to facilitate recommendations, we invite Amazon Mechanical Turk (AMT) annotators with a higher than 95% HIT approval rate to label bounding boxes of background items in fashion scenes (Appendix A.1.1). These background items include *display table*, *wardrobe*, *floor rack* and *display wall*. Each bounding box of background item covers all clothes in it and is labeled as *"absolute position + name"* like *"back leftmost closet"*. To

ATTRIBUTE TYPE	CATEGORIZATION CONCEPT	SUBJECTIVE PREFERENCE
Color	warm color (*red, brown, yellow, light pink...*)	*"color of passion", "lively color", "color for outgoing people", "color for happy activity"...*
	cold color (*green, blue, light purple, olive...*)	*"cool color", "color which makes people feel calm", "refreshing color", "color of quietness"...*
	powerful color (*red, orange, light red...*)	*"color full of energy", "color inspires the fighting spirit", "color makes people feel excited"...*
	mysterious color (*violet, black, dark blue...*)	*"elusive color", "unfathomable color", "color inspires exploration", "color arouses curiosity"...*
Pattern	lively pattern (*floral pattern, leopard print...*)	*"vibrant pattern", "pattern closer to nature", "pattern that is popular among conservationists"...*
	dazzling pattern (*star design, diamond style...*)	*"pattern welcomed by superstar", "eyes-catching pattern", "pattern suitable for a dancing party"...*
	modest pattern (*stripes, checkered and plain...*)	*"simply decorated pattern", "pattern for the middle-aged", "pattern selected by humble people"...*
Material	soft material (*natural fibers, wool, leather...*)	*"easily bent material", "gentle material", "material suitable for a child", "comfortable material"...*
	gorgeous material (*leather, wool, silk...*)	*"luxury material", "material indicating social status", "material welcomed by rich people"...*
	reliable material (*metal, marble, plastic...*)	*"durable material", "material for restaurant furniture", "material for outdoor furniture"...*

Table 1: Examples of some categorization concepts and corresponding subjective preferences.

ensure unambiguity, there are no repeated background item labels in the same snapshot.

Attribute Categorization There are 290 and 110 different digital items in clothes and furniture domains. Nine attributes (e.g., *type, color, pattern, material, price, brand, size and customer review*) are used for describing the items in metadata (the database of all digit items).

Customers usually are not experts in those domains. They tend to express their requirements with what we call subjective preferences. On the one hand, a subjective preference can normally be mapped to a set of attribute values. For example, *"color for happy activity"* corresponds to *"red, brown, yellow, ..."*. On the other hand, there are lots of subjective preferences with the same meaning. For example, customers also say "color for the welcome ceremony" and "color of passion", which is similar to "color for happy activity" in attribute reference. The relationship between subjective preferences and attribute values is a many-to-many mapping. To bridge them, we collect a set of categorization concepts from the survey for domain experts(Appendix A.1.2), each of which is a synonym of corresponding subjective preferences. Every subjective preference can be mapped to a categorization concept. Therefore, we transform the many-to-many mapping between subjective preferences and attribute values into two stages: ❶ Many-to-one mapping from subjective preferences to categorization concepts, ❷ One-to-many mapping from concepts to attribute values.

The introduction of categorization concepts is necessary and convenient for the two phases TOD data building: In the dialog flow simulation, the customer simulator expresses requirements by categorization concepts as the slots. In the manual paraphrase, all categorization concepts are paraphrased to subjective preferences by human annotators. In this way, we can simulate dialog flow and guarantee language diversity at the same time.

Tab. 1 gives some examples of subjective preferences, categorization concepts, and attribute values. Fig. 22 in Appendix show more details about them.

Salesperson Act. To collect common salesperson acts, we invite 238 sales experts with more than three years of work experience to complete questionnaires (Appendix A.1.2). The maximum entry number is limited to 1 to guarantee result diversity. We summarize 8 different dialog acts of salespersons (Tab. 2) and introduce them in detail.

- *Ask Preference* refers to asking customer's preference about one attribute type. Overlapping values referred to by all subjective preferences responded from the customer are new candidates.

- *Exclude Preference* is asking the customer what he dislike. The responded subjective preference from customer is utilized to exclude unfavored attribute values from candidates.

- *Prompt Preference* is actively providing subjective "preference" for the customer to confirm.

- *Guess Attribute Value* is predicting one concrete value from candidate attribute values based on multimodal context.

- *Revise Attribute Value* is revising the previous prediction of concrete attribute value following the customer's feedback.

- *Display Candidate Values* refers to listing all candidate attribute values based on multimodal context for the customer to choose from.

- *Refer Region* is an act that salesperson points out one region in the store like *"front floor rack"* to ask the customer whether the region contains the item he wants.

SALESPERSON ACT	CUSTOMER ACT	EXAMPLES
Ask Preference	Answer Preference	A:*"Could you please tell me your preference on clothes pattern?"* → U:*"I would like to choose clothes with **pattern that is suitable for my dancing party**."*
Exclude Preference	Negate Preference	A:*"Is there any kind of furniture material you don't like? I can avoid recommending it to you."* → U:*"I will not choose **materials popular among rich people**. I am used to a simple life."*
Prompt Preference	Respond Prompt	A:*"In this case, what is your opinion on the **color that makes people feel cool**?"* → U:*"Cool color sounds Great! It is suitable for my summer vacation. Just follow your suggestion."*
Guess Attribute Value	Respond Attribute Value	A:*"Can you tell me how do you think about <u>leather material</u>?"* → U:*"I like this kind of material. Just recommend <u>leather</u> furniture to me later."*
Revise Attribute Value	Respond Attribute Value	A:*"How about brand <u>Art News Today</u>?"* → U:*"I don't like **brand with a long history**."* (next round) A:*"I see .. would you like trademark <u>Coats & More</u>?"* → U:*"Oh! You know my taste."*
Display Candidate Values	Choose Attribute Value	A:*"I have jackets with blue color, green color and white color. Which color do you prefer?"* → U:*"Let me have a think. Oh, as for me, I would like <u>white color</u> for jacket."*
Refer Region	Judge Region	A:*"Have a look at the **back left wardrobe near the mirror**. Is there anything you like in this region?"* → U:*"Sorry, there seem to be no suitable clothes in this closet."*
Recommend Item	Respond Recommendation	A:*"What is your idea of the black and white t-shirt with stripes hanging on the **rightmost floor rack**?"* → U:*"Yeah, it looks very beautiful! Please help me add this t-shirt to my cart."*

Table 2: Sixteen kinds of salesperson acts and customer acts in SURE dialogs. Corresponding dialog examples for each pair of acts are displayed in the last column. All subjective preferences and referred regions are highlighted in bold while concrete attribute values are underlined.

- ***Recommend Item*** is trying to recommend items from the candidate item set based on multimodal context.

Different combinations of the above acts in one dialog can form diverse recommendation strategies. For example, a salesperson can continually ask customer preferences to narrow the candidate set of attribute values, and then display all possible values for the customer to choose when several attribute values are in the candidate set. Using reasonable strategies for different situations can effectively improve recommendation efficiency and accuracy.

3.1.2 Multimodal Dialog Simulator

The multimodal dialog simulator takes store scenes along with the meta information to create salesperson-customer dialog flows following (Kottur et al., 2021).

Multimodal Dialog Flow Generation. The dialog flow simulator is composed of the *goal generator*, the *customer simulator* and the *salesperson simulator*. The goal generator randomly selects an item from the given store scene and takes it as a target item that is invisible to others. The salesperson simulator is aware of the scene snapshot (commodity and background items' position) and metadata of all items, which actively elicits customer preferences and recommends items following a probability distribution. Each act (e.g., *Prompt_Reference, Refer_Region*) is companied by some slots (e.g., attribute, region). The customer simulator is assigned attribute values of the target item and responds to salespersons by customer acts with slot values limited to categorization concepts or simple yes/no (except *Display_Candidate_Values*), which simulates

the user requirements. Take one-round interaction as an example. After the salesperson simulator generates *"Ask_Preference*:{Color}*"*, the customer simulator chooses its act *"Answer_Preference"* and a categorization concept *"warm color"* based on assigned *"red"* color as the preference slot. The simulation repeats until the salesperson simulator successfully recommends the target item.

3.2 Manual Paraphrase

Based on dialog flows obtained from simulator interaction, we design a manual paraphrase process to rewrite subjective preferences and then paraphrase dialog flows.

Subjective Preference. In simulated dialog flows, subjective preferences are expressed by categorization concepts for attribute values. Every categorization concept can be paraphrased to many different subjective preferences. Take *"warm color"* as an example. It can be rewritten to *"color of passion"* by customers' subjective feeling, to *"color welcomed by outgoing people"* by suitable persons, to *"color for the happy ceremony"* by applicable scenarios. Human annotators are required to paraphrase categorization concepts to subjective preferences in any of the cases to increase language diversity.

Dialog Generation. To make dialogs closer to language distribution of real shopping dialogs, we invite AMT annotators with a higher than 90% HIT approval rate to paraphrase dialog flow following these instructions: ❶ Write salesperson ↔ customer utterances based on dialog flows. All subjective preferences and concrete attribute values are
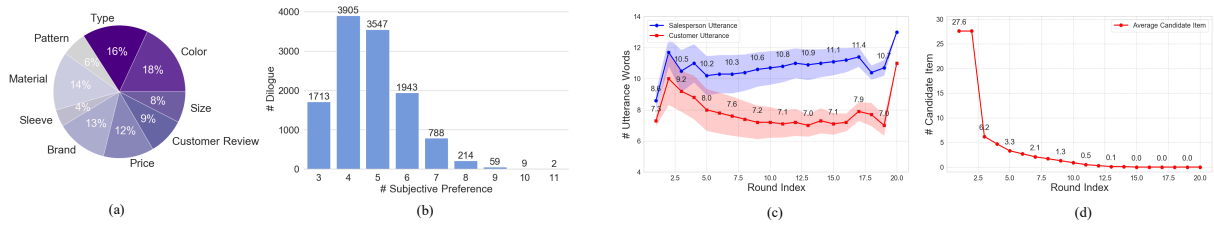
Figure 3: (a) Percentage of subjective preferences in attribute types, (b) distribution of subjective preferences among dialogs, (c) number of utterance words with dialog rounds, (d) number of candidate items with dialog rounds.
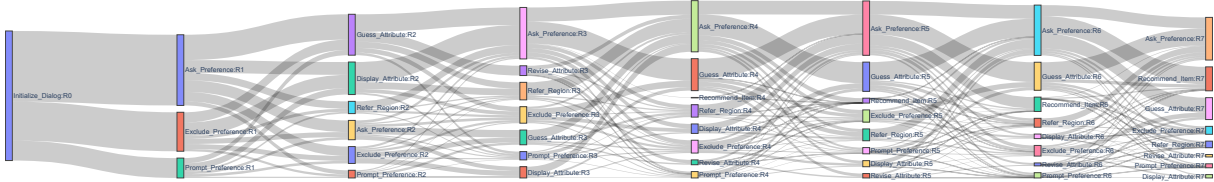

Figure 4: Salesperson act transitions in the first eight rounds. Different combinations of salesperson acts form diverse strategies to recommend the target item from complex store scene based on subjective preferences.

reserved; ❷ Add the visual descriptions and spatial relations of scene items; ❸ Rewrite repeated nouns or phrases to co-reference; ❹ Add polite expressions and modal particles into utterances. The detailed instructions with scene snapshot and dialog flow can be checked in Appendix A.1.3. An example of SURE dialog with annotations is shown in Appendix Fig. 21.

3.3 SURE Dataset Analysis

To the end, we build the SURE dataset, Tab. 3 gives some statistics of the data. We highlight the information on subjective preference and dialog policy in the following subsections.

Total # dialogs	12180
Total # utterances	223492
Total # scene snapshots	1566
Total # subjective preferences	3043
Avg # words per customer turns	7.75
Avg # words per salesperson turns	10.49
Avg # utterances per dialog	18.35
Avg # objects per scene in dialog	27.6
Avg # subjective preferences per dialog	4.48
Avg # salesperson acts per dialog	8.17

Table 3: SURE Dataset Statistics.

Subjective Preference. As shown in Tab. 3, there are $3K$ different subjective preferences in SURE. The percentage of subjective preferences in different attribute types is shown in Fig. 3 (a), from which we can observe the richness and diversity of subjective preferences. On average, each dialog contains 4.48 subjective preferences. The distribution of subjective preferences among dialogs is displayed in Fig. 3 (b). It is clear that subjective preferences are widely distributed in SURE, which brings a new challenge for conversational agents.

Dialog Policy. As shown in Tab. 3, the SURE dataset collects $12K$ shopping recommendation dialogs. The utterance length with dialog turn is displayed in Fig. 3 (c). On average, each dialog contains 8.17 salesperson acts to recommend the target item from 27.6 candidate scene items. We visualize the salesperson act transactions for the first eight rounds in Fig. 4. It can be observed that different act combinations form diverse recommendation strategies in SURE. From the stream width, we can find that salespersons have a higher probability of asking for preference than excluding and prompting preferences. Besides, salespersons don't directly guess or display concrete attribute values at the very beginning round. They are prone to conduct these acts after they collect at least one customer preference. As the dialogs go on, salespersons also try to reduce candidate items by referring to specific region. We display the number of candidate items over rounds in Fig. 3 (d). It can be seen that salespersons begin to recommend items when the candidate range is small enough. The recommendation strategies in SURE are close to real-life shopping.

4 Task Formulation

We propose three tasks on SURE dataset to evaluate multimodal recommendation agents. The tasks of Subjective Preference Disambiguation and Referred Region Understanding evaluate the side of multimodal understanding, while Multimodal Recommendation evaluates the side of policy learning.

Subjective Preference Disambiguation (SPD). After the customer expresses subjective prefer-
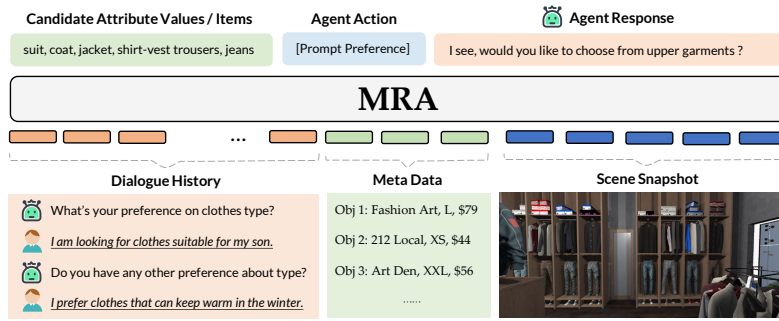
Figure 5: The MRA model for **Subjective Preference Disambiguation**, **Referred Region Understanding** and **Multimodal Recommendation** tasks. The flattened dialog history and metadata are concatenated as textual input while the scene snapshot is split into patches. Note that the green area in the output will predict candidate attribute values if the latest round relates to eliciting preference. When the latest round refers to specific region, the green area predicts object IDs in the region.

ences, the agent needs to determine candidate attribute values based on preferences in dialog history and scene items in the store. We denote this process as subjective preference disambiguation, which establishes the connection between subjective preferences and concrete attribute values. This task is important because the correct recommendation is closely dependent on clear attribute requirements. It requires the agent to abstract subjective preferences to categorization concepts and then filter grounded attribute values by these concepts. From cognition research, this task involves visual perception, language conceptualization and attribute categorization.

The input of this task includes dialog history, current customer utterance, and scene snapshot. With this information, the agent predicts all possible attribute values (e.g., U:*"I prefer the color of happiness."* → *"yellow, brown, red"*). Attribute values that meet customer requirements but do not exist in the scene snapshot should be excluded. The main evaluation metric can be F1, precision, and recall performance. Note that the evaluation is only implemented on rounds for eliciting preferences.

Referred Region Understanding (RRU). Referring to region is an essential act for narrowing the candidate item range. This task aims to update the candidate item set after the customer responds to referred region. To complete this task, the agent needs to locate the regional referring expression (e.g., *"far back middle floor rack"*) and then filter the previous candidate set by items in the region. It requires the agent to correctly understand the referred region, visual attributes and spatial relations in the scene.

The input of this task is dialog history with the latest round containing referred region and scene

snapshot. Based on this information, the agent needs to predict all object IDs in the region (e.g., A:*"Come with me to look at the shelf on the right. Are there any clothes that you like?"* U:*Sorry, there is no garment that I am looking for in this region.* → *"12, 13, 16, 22, 31"*). Objects in the same scene but not in the referred region should be excluded. The agent performance can be measured by F1 score, precision, and recall metrics on object ID prediction. Note that the evaluation is only implemented on rounds for referring region.

Multimodal Recommendation (MR). When customers seek recommendations in the store, they hope the salesperson can recommend items accurately and efficiently. Therefore, recommendation strategy, recommendation success rate, and language quality all influence customer shopping experience. We define **Act Prediction** sub-task and **Response Generation** sub-task at turn level and define **Item Recommendation** sub-task at dialog level to evaluate the agent's multimodal recommendation performance comprehensively.

The input of these three sub-tasks is dialog history, current customer utterance, and scene snapshot. The **Act Prediction** sub-task requires the agent to predict the next salesperson act (e.g., U:*"The price $299 is too expensive for me to afford"* → *"Revise_Attribute"*). The F1 score, precision, and recall can be calculated for cumulative act predictions to measure performance. The **Response Generation** sub-task requires the agent to generate the next salesperson utterance (e.g., U:*"I'd like to buy a sofa made by materials obtained from nature"* → *"You can consider the leather material, which is natural and smooth."*). The generated utterance can be evaluated by BLEU-4 (Papineni et al., 2002) or ROUGE (Lin and Hovy, 2003). The

MODELS	Task 1 SPD	Task 2 RRU	Task 3 MR		
	Disam. F1↑	Refer F1↑	Act F1↑	BLUE-4↑	Recom. F1↑
MRA	**36.78 / 37.49**	**13.77 / 14.77**	**10.80 / 10.90**	**18.28 / 18.45**	**23.49 / 23.14**
- Meta Data	32.34 / 31.98	9.57 / 10.15	10.01 / 10.21	16.42 / 16.83	19.78 / 19.23
- Scene Snapshot	31.44 / 31.83	7.36 / 7.09	9.45 / 9.77	17.72 / 17.63	21.54 / 21.26

Table 4: The performance of baseline model MRA on three SURE benchmarks. MRA's results on dev-test and test-std are displayed by **"(dev-test / test-std)"** format. We respectively ablate metadata and scene snapshot to observe MRA's ability to utilize multimodal context to recommend item via subjective preferences.

Item Recommendation sub-task requires the agent to predict the target item ID (e.g., *"<@1132>"*) in the last round utterance, which can be extracted by regex and evaluated by F1 score, precision, and recall.

5 Modeling & Empirical Analysis

Dataset Split. SURE is randomly divided into 4 parts: train (65%), dev (5%), dev-test (15%), and test-std (15%). We leave test-std as a held-out hidden set for performing a fair comparison of models in future potential competition.

Baseline. We proposed MRA(Multimodal Recommendation Agent) model as the baseline model for subjective preference disambiguation, referred region understanding, and multimodal recommendation tasks. The backbone of the MRA model is encoder-decoder based single-stream Visual-Language Pre-training Model, which are stacks of Transformer (Vaswani et al., 2017) layers. The scene image is split into P patches. And each patch is projected to the visual embedding of the model's hidden size. The flattened dialog history and non-visual metadata are converted to subword sequences by Byte-Pair Encoding (BPE) and then embedded into textual embedding. All visual embedding and textual embedding are concatenated as model input. The MRA model completes three benchmarks at the same time by generating *next salesperson act, candidate attribute values / referred region items* and *agent response* auto-regressively as Fig. 5 shows. Note that MRA will predict candidate attribute value if the latest round relates to eliciting preference. When the latest round relates to referring region, MRA predicts object IDs in the region.

Implementation Details. MRA model is based on Transformer (Vaswani et al., 2017) structure with 12 layers, where every Transformer block has 768 hidden units and 12 attention heads. Textual and visual embedding are projected to features the same size as the hidden units. We initialize MRA

parameters from pretrained OFA-base (Wang et al., 2022) model. During training, MRA model is trained for 20 epochs with 18 batch sizes to optimize language modeling loss. At the end of every epoch, the model is evaluated on dev split to save the best model parameters. The hyperparameters are determined by area search. Adam (Kingma and Ba, 2015) is adopted as the optimizer with a 4e-4 learning rate while the dropout rate is set to 0.2 to prevent over-fitting. The whole training costs around 36 Tesla-V100 GPU hours. Note that the BLEU-4 score is calculated by NLTK (Bird et al., 2009) package in the evaluation.

Analysis and Future Work As Tab. 4 shows, model MRA, powered by an advanced multimodal backbone, fails to perform well on three SURE benchmarks. From the case study (Fig. 20), we can observe that it is difficult for MRA to accurately understand subjective preferences (Task 1) and referred region (Task 2), which further hinders the model from adopting suitable acts to make correct recommendations (Task 3). Ablation of metadata and scene snapshot greatly weakens MRA on the first two tasks. It indicates that effective utilization of metadata and scene snapshot plays an essential role in model performance. Future work can be done by designing modules or proposing multimodal pretraining tasks to facilitate model's understanding of subjective preferences, perception of referred region, and ability to take suitable acts.

6 Conclusion

We introduce *Multimodal Recommendation Dialog with SUbjective PREference* (SURE) dataset with $12K$ salesperson ↔ customer dialogs and $3K$ subjective preferences to study how to recommend item from complex scene based on subjective preferences. Our proposed three benchmarks and strong baseline model MRA address the new challenges and directions in the multimodal recommendation dialog.

Limitations

The annotation of attribute categorization and subjective preferences may vary from person to person, influencing preference disambiguation results in the real world. We have tried to reduce bias by choosing categorization concepts and subjective preferences agreed upon by more than three annotators. Besides, owing to time and funds constraints, we only manually paraphrase dialog flow in English. For this reason, the agent built on SURE can just communicate in English. To overcome this limitation, we plan to annotate SURE in multilanguage in the next stage.

Ethics Statement

Our work strictly complies ACL Code of Ethics. We respect CC-BY-NC-SA-4.0 license required by SIMMC 2.0 and only use its scene snapshots for academic research. We will also release our dataset to the community with the same license. As for human annotation, we anonymously recruit human annotators on the Amazon Mechanical Turk (AMT) platform to protect their personal privacy. In the task instructions, we have informed participators that any annotations related to personal attacks, racial or sexism discrimination will lead to HIT rejection. Besides, we also demonstrate that their annotations will be used for academic purposes. The payment of our released tasks is competitive on the AMT platform compared with similar tasks (Appendix A). After completing human annotations, we manually check the collected information to exclude any potential offensive information. Our annotation process and data content got approval from an ethics review board by an anonymous IT company. We can guarantee the trustworthiness of our technologies, limitations, and ethics statement.

Acknowledgements

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, and Yang Liu ... 2021. Overview of the ninth dialog system technology challenge: DSTC9. In *DSTC9 challenge wokrshop at AAAI*.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.

Xin Huang, Chor Seng Tan, Yan Bin Ng, Wei Shi, Kheng Hui Yeo, Ridong Jiang, and Jung Jae Kim. 2021. Joint generation and bi-encoder for situated interactive multimodal conversations. In *DSTC9 challenge wokrshop at AAAI*.

Younghoon Jeong, Se Jin Lee, Youngjoong Ko, and Jungyun Seo. 2021. Tom: End-to-end task-oriented multimodal dialog system with gpt-2. In *DSTC9 challenge wokrshop at AAAI*.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A. Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. *CoRR*, abs/1909.03922.

Byoungjae Kim, Inkwon Lee, Yeonseok Jeong, Ko Youngjoong, Myoung-Wan Koo, and Jungyun Seo. 2021. Improving multimodal api prediction via adding dialog state and various multimodal gates. In *DSTC9 challenge wokrshop at AAAI*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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)*.

Satwik Kottur1, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2022. Overview of situated and interactive multimodal conversations (simmc) 2.0 track at dstc 10. In *DSTC10 challenge wokrshop at AAAI*.

Po-Nien Kung, Tse-Hsuan Yang, Chung-Cheng Chang, Hsin-Kai Hsu, Yu-Jia Liou, and Yun-Nung Chen. 2021. Multi-task learning for situated multi-domain end-to-end dialogue systems. In *DSTC9 challenge wokrshop at AAAI*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. *CoRR*, abs/2005.03954.

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.

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)*.

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 (ACL)*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. 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 8689–8696. AAAI Press.

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

Matteo Antonio Senese, Giuseppe Rizzo, Alberto Benincasa, and Barbara Caputo. 2021. A response retrieval approach for dialogue using a multi-attentive transformer. In *DSTC9 challenge wokrshop at AAAI*.

Pararth Shah, Dilek Hakkani-Tür, Gökhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry P. Heck. 2018. Building a conversational agent overnight with dialogue self-play. *CoRR*, abs/1801.04871.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

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

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain, December 8-11, 2020*.

A Appendix

A.1 Human Annotation

A.1.1 Background Item Annotation.

We release "Background Item Annotation" task on Amazon Mechanical Turk (AMT) platform to hire workers to draw bounding box around background item and annotate corresponding label. To guarantee the quality, we require workers have greater than 95% HIT approval rate. For payment, we pay $0.25 for every store scene snapshots that contains about 5 potential background items, which is competitive compared with similar tasks. The detailed task instruction and scene snapshot are displayed in the following.

Background Item Annotation

This HIT is a part of scientific research, whose results may be presented at scientific meetings or published in scientific journals. In this task, you are invited to draw all bounding boxes of background items, like floor rack, display table and shelf, in the given scene snapshot and annotate the corresponding label. The annotated bounding box has to cover all clothes bounding boxes contained in it and the label should be in the "absolute position + name" format such as "back leftmost closet". The hit will

be rejected if the annotated bounding box does not cover all clothes bounding boxes contained in it or the label has racism, sexism and privacy information. If you are fully aware of and agree with above information, you can begin to work on the following scene snapshot.

Figure 6: Scene snapshot of background item annotation. Existing red bounding boxes are for clothes items.

A.1.2 Questionnaire for Sales Expert

We release "Clothes Recommendation Survey" task and "Furniture Recommendation Survey" task on Amazon Mechanical Turk (AMT) platform to invite fashion sales experts and furniture sales experts to complete questionnaire. To guarantee the quality, we require answers have greater than 90% HIT approval rate. For payment, we pay for $2.0 for every carefully completed questionnaires, which is high compared with other survey tasks in the same period. We display the instruction and questionnaire for "Clothes Recommendation Survey" task in the following.

Clothes Recommendation Survey

Our project is aimed at studying clothes recommendation in the store scene. To collect real-life data, we sincerely invite experienced sales expert to complete the following questionnaire with 12 questions. *We guarantee that all questionnaires will be conducted anonymously, and the survey results will only be used for academic research rather than commercial purposes. The hit will be rejected if the comments contain any racism, sexism and privacy information. If you are fully aware of and agree with above information, you are welcome to accept the survey.

1. How long work experience do you have in clothing sales? (Single Choice Question)
A. I don't have work experience on clothing sales.
B. Less than 1 year.
C. About 1 - 3 years.
D. About 3 - 5 years.
E. More than 5 years.

2. What order of priority do you usually follow when eliciting customer preferences on attributes? (Sorting Question)
- *Type (e.g., jacket, t-shirt, dress)*
- *Color (e.g., red, blue, white)*
- *Pattern (e.g., plain, stripe, spot)*
- *Size (e.g., XS, M, XXL)*
- *Sleeve Length (e.g., short, full, half)*
- *Brand*
- *Price*

3. Do you feel how often customer express their requirements on attribute value by subjective preference such as "color of passion", "lively pattern" and "formal fashion type"? (Single Choice Question)
A. More than 60% of time.
B. About 40% - 60% of time.
C. About 20% - 40% of time.
D. Less 20% of time.

4. Please categorize the given attribute values by your domain knowledge and subjective feeling. For example, "red, yellow, brown..." can be categorized to "warm" color while "floral pattern, leopard print..." can be categorized to "lively pattern". You can just select several values from candidates to define categorization concept. Then, write down some subjective preference your customer expressed if you have met such case. (Leave Comments)
Given Attribute Values: red, yellow, brown, orange, pink, blue, black, grey, dark olive, light red. (These provided attribute values vary from questionnaire to questionnaire.)

5. Do you think knowing about what kinds of concrete attribute values are most important for successful recommendation? (Multiple Choice)
A. Type (e.g., jacket, t-shirt, dress)
B. Color (e.g., red, blue, white)
C. Pattern (e.g., plain, stripe, spot)
D. Size (e.g., XS, M, XXL)
E. Sleeve Length (e.g., short, full, half)
F. Brand
G. Price

6. When customer express subjective preferences, what strategy will you utilize to

disambiguate their requirements? For example, "color of passion"→"red, orange, yellow...". (Multiple Choice)
A. Continually ask customers more preferences on this attribute type.
B. Invite customers to describe some information about their unfavored attribute values.
C. Actively prompt some subjective preferences on attribute type to inspire customers.
D. Directly display all attribute values in the candidate set for customers to choose.
E. Guess one concrete attribute value based on customer expressed preferences and revise prediction following feedback.
F. You are welcome to tell us your personal strategy! (Leave Comments)

7. At the time you elicit customer preference on one particular clothes attribute, how many attribute values remain in the candidate attribute set when you list them for customer to choose? (Single Choice Question)
A. More than 8 candidate attribute values.
B. About 5 - 8 candidate attribute values.
C. About 3 - 5 candidate attribute values.
D. About 1 - 3 candidate attribute values.
E. Only 1 candidate attribute value.

8. Collecting information will promote recommendation accuracy but may consume customer patience. In this case, how do you balance collecting customer preference and trying recommending items? (Single Choice Question)
A. I will learn about customer preference about all clothes attributes in detail first, and recommend item when having full confidence.
B. I will learn about customer preference roughly first, and recommend item when candidate range is small enough.
C. I will learn about detailed customer preference on 1 - 2 clothes attributes, and then try to recommend item until customer accept.
D. I will learn about detailed customer preference on 3 - 4 clothes attributes, and then try to recommend item until customer accept.

9. When will you guide customers to a particular region and invite them to see if there are suitable clothes? For example, "Please follow me to have a look at the left display wall. Are there

anything you like?". (Single Choice Question)
A. After learning about detailed customer preferences on 1 - 2 clothes attributes, I will select one rack in the store and show it to the customer.
B. After learning about detailed customer preferences on 3 - 4 clothes attributes, I will select one rack in the store and show it to the customer.
C. Only when the range of candidate items is small enough, I will select one rack in the store and show it to the customer.
D. I don't do like what the question says. I always directly point out clothes which customer may be interested in.

10. How many clothes items remain in the candidate item set when you recommend concrete item for customer? (Single Choice Question)
A. More than 8 candidate items.
B. About 5 - 8 candidate items.
C. About 3 - 5 candidate items.
D. About 1 - 3 candidate items.
E. Only 1 candidate item.

11. How do you unambiguously refer the item in the store when you recommend item to customer? (Single Choice Question)
A. Refer the item by its color.
B. Refer the item by its pattern,
C. Refer the item by its color and pattern.
D. Refer the item by spatial relation.
E. Combine all above methods to refer item although its description is complex.

12. What do you think is the most important when shopping recommendation dialogs? (Single Choice Question)
A. Recommend customer desired item through as few conversations as possible. (i.g. recommendation efficiency)
B. Keep recommendation strategies be consistent with those in real-life shopping conversations.
C. Balance the recommendation efficiency and customer shopping experience.
E: Guarantee dialog language polite and warm.
F: You are welcome to write your advice! (Leave Comments)

For clothes recommendation questionnaire, we totally receive 765 complete survey results. To guarantee the reliability of servery, we exclude questionnaires from salesperson with less than

three years work experience and remain 238 results for statistics. In the following, we visualize survey results in Fig. 7 to Fig. 17 by bar charts and pie charts.
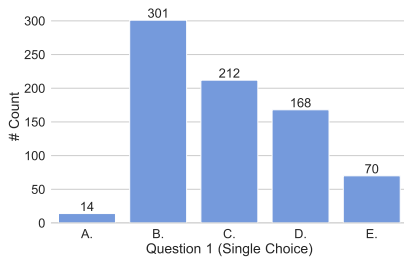

Figure 7: The bar chart of Question 1 results. **Choice D** and **Choice E** are selected by 238 sales expert with more than three years work experience. Only these 238 experienced sales expert are invited to complete the following Question 2 - Question 12.
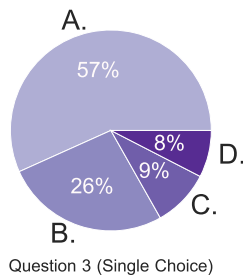

Figure 8: The pie chart of Question 3 results. From **Choice A** selection (57%), we can find that over half of sales expert often meet customers who express their requirements on attribute value by subjective preferences.
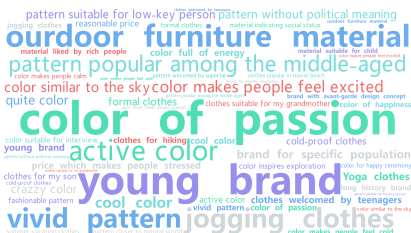
Figure 9: The word cloud of Question 4 comments. We can find that customers usually express subjective preferences by visual perception and commonsense.

A.1.3 Manual Paraphrase

We release "Dialog Writing with Subjective Preference" task on on Amazon Mechanical Turk (AMT) platform to hire workers to write subjective preferences according to attribute categorization concepts and paraphrase dialog flows to natural language utterances with subjective preferences. To guarantee the quality, we require answers have greater than 90% HIT approval rate. For payment, we pay for $1 for every carefully paraphrased
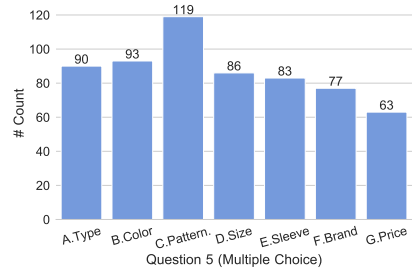

Figure 10: The bar chart of Question 5 results. From the choice distribution, we can see that **Pattern** plays the most important role in successful recommendation while **Price** is relatively less essential.
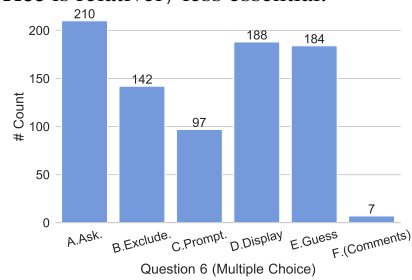

Figure 11: The bar chart of Question 6 results. From the choice distribution, we can see that salespersons are prone to ask preference, display candidate attribute values and guess one concrete attribute value. Sometimes, they also try to exclude customer unlikeness or actively prompt.
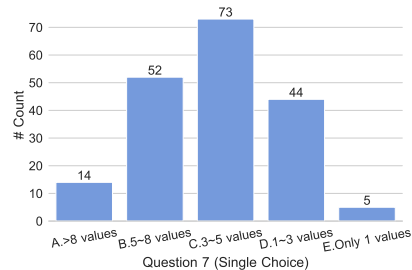

Figure 12: The bar chart of Question 7 results. Only 188 sales experts who choose Choice D in Question 6 participate in answering this question. We can find that most sales experts prefer to display all candidate attribute values when there only exists 3~5 values in the candidate set.
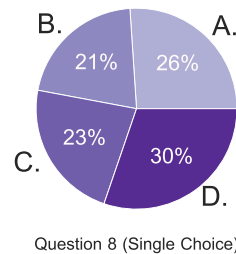

Figure 13: The pie chart of Question 8 results. Most sales experts are prone to learn about some aspects of customer preferences in detail as **Choice B**, **Choice C** and **Choice D** show. Only 21% of sales experts just roughly learn about customer preferences before recommending items.
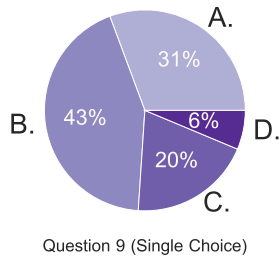
Question 9 (Single Choice)

Figure 14: The pie chart of Question 9 results. We can find that most sales experts (**Choice A** and **Choice B**) narrow candidate items range by referring one specific region in the store scene after they elicit several aspects of customer preference.
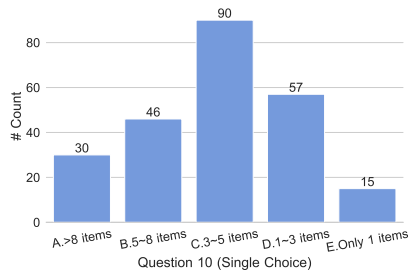

Question 10 (Single Choice)

Figure 15: The bar chart of Question 10 results. We can find that the distribution mainly falls into **Choice C** and **Choice D**, which indicates that sales experts prefer to recommend items when there are about 1~5 items left in the candidate set.
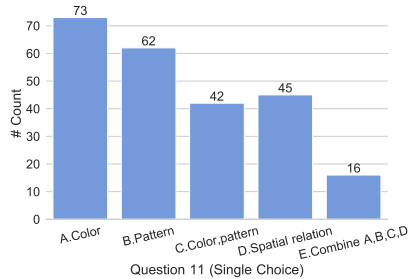

Question 11 (Single Choice)

Figure 16: The bar chart of Question 11 results. Most sales experts are prone to refer scene items solely by **Color**, **Pattern** or **Spatial Relation** instead of their combinations. Clear and simple referring expression is more consistent with real-life shopping conversation.
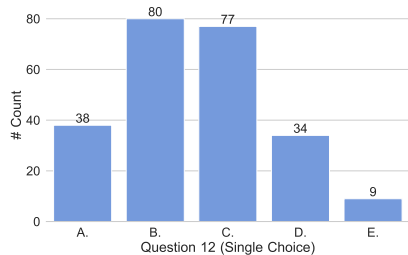

Question 12 (Single Choice)

Figure 17: The bar chart of Question 12 results. We find that sales experts pay more attention to reality and diversity of recommendation strategies (**Choice B**) and customer shopping experience (**Choice C**) rather than recommendation efficiency (**Choice A**).
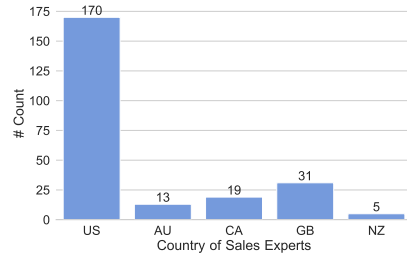

Figure 18: The bar chart displays the demographic and geographic characteristics of the 238 surveyed sales experts. Most experts come from United States (US) while other experts come from United Kingdom (GB), Australia (AU), Canada (CA) and New Zealand (NZ).

dialog, which is competitive with similar tasks in the same period. We display the instruction and scene snapshot for this task in the following.

Dialog Writing with Subjective Preference

*This task is to collect multimodal recommendation dialogs between salesperson and customer with subjective preferences. You need to write subjective preferences based on given attribute categorization concepts by visual perception or commonsense firstly. For example, **"warm color"** can be written to "color of passion", "color of happiness" and "color for welcome ceremony" while **"lively pattern"** can be written to "vibrant pattern", "pattern closer to nature" and "pattern that is popular among conservationists". Then, you need to paraphrase provided dialog flow to dialog with written subjective preferences. We introduce every kind of dialog act in the following.*

- *Ask Preference (Salesperson) refers to asking customer preference about one attribute type.*

- *Answer Preference (Customer) refers to answering subjective preference about asked attribute type.*

- *Exclude Preference (Salesperson) is asking customer what he unlike.*

- *Negate Preference (Customer) is answering unlike attribute by subjective preference.*

- *Prompt Preference (Salesperson) is actively providing subjective "preference" to inspire customer.*

- *Respond Prompt (Customer) is responding to salesperson's prompt.*

3528

- *Guess Attribute Value (Salesperson)* is predicting one concrete value from candidate attribute value.

- *Revise Attribute Value (Salesperson)* is revising prediction of concrete attribute value following customer's feedback.

- *Respond Attribute Value (Customer)* is responding to attribute value guessed or revised by salerperson.

- *Display Candidate Values (Salesperson)* refers to listing all candidate attribute values for customer to choose.

- *Choose Value (Customer)* refers to choosing target value from listed values.

- *Refer Region (Salesperson)* is an act that salesperson points out one region in the store to ask customer whether the region contains item he wants.

- *Judge Region (Customer)* is judging whether the referred region contains the item that customer is looking for.

- *Recommend Item (Salesperson)* is trying to recommend items from candidate item set based on multimodal context.

- *Respond Recommend (Customer)* is responding to item recommendation.

Please paraphrase dialog act according to above information. The written should keep the similar meaning as original dialog act and reserve the corresponding attribute type or subjective preference following given slot. For example, "Answer_Preference:{Color:warm color}" can be written to "I am looking for clothes with color that makes me feel happy." while "Negate_Preference:{Pattern:dazzling pattern}" can be written to "I am interested in eyes-catching pattern." You are welcome to add visual descriptions and spatial relations with background items to refer commodity items (orange bounding boxes in the scene snapshot annotate all background items). You are encouraged to add polite expressions and modal particles into utterances. Note that the hit will be rejected if the utterances contain any racism, sexism and privacy information. The annotated dialogs may be presented at scientific meetings or

published in scientific journals for academic research. If you are fully aware of and agree with above information, you are welcome to accept the task.

Figure 19: Scene snapshot of manual paraphrase. Red bounding boxes are for clothes items while orange boxes are for background items.

A.2 Case Study for MRA

Dialogue History

🤖 What's your preference on clothes type?

🧑 *I am looking for clothes suitable for my grandpa.*

Candidate Values Prediction

trousers, blouse, coat, shirt

Figure 20: Case study for MRA.

Role	Utterance	Act	Candidate Items	Attribute Values
Salesperson	Good afternoon! Welcome to our store. What can I do for you?	Initialize_Dialog	<@1004>, <@1259>, <@1062>, <@1227>, <@1169>, <@1055>, <@1119>, <@1072> …	
Customer	I am looking for some **spring summer** clothes.	Express_Preference		
Salesperson	Could you please give me some examples about clothes type which you will not consider?	Exclude_Preference	<@1004>, <@1259>, <@1062>, <@1227>, <@1169>, <@1055>, <@1119>, <@1072> …	
Customer	It has no impact if you ignore **garments designed for running or Yoga**. Actually, I already have such kind.	Negate_Preference		
Salesperson	Please consider blouse and dress. You can choose from them.	Display_Attribute	<@1004>, <@1240>, <@1062>, <@1009>, <@1119>	Clothes Type: blouse, dress
Customer	Oh, I would like to buy a blouse for my daughter.	Choose_Attribute		
Salesperson	You can come with me to rightmost floor rack to take a look at these items in this area.	Refer_Region	<@1004>, <@1240>, <@1062>, <@1009>, <@1119>	
Customer	Sorry, I cannot find items that I prefer in your referred region.	Judge_Region		
Salesperson	I see.. what is your preference on sleeve length?	Ask_Preference	<@1004>, <@1062>, <@1009>, <@1119>	
Customer	I am looking for clothes with **sleeve keeps people feel cool**.	Answer_Preference		
Salesperson	Could you please tell me some kinds of brand which you don't want to consider?	Exclude_Preference	<@1062>, <@1009>, <@1119>	Sleeve Length: short
Customer	It has no impact if you ignore **ground-breaking brand**. Actually, I am worried about their quality.	Negate_Preference		
Salesperson	Are there any other kinds of you unliked brand?	Exclude_Preference	<@1062>, <@1009>, <@1119>	Brand: StyleNow Feed, Brain Puzzles, Art Den
Customer	I am also not interested in **brand with long history**. My daughter dislike to wear clothes with the sane brand as mine.	Negate_Preference		
Salesperson	OK, maybe you have other kind of unfavored brand?	Exclude_Preference	<@1062>, <@1119>	Brand: Brain Puzzles, Art Den
Customer	I remember that my daughter also told me not to choose **widely-known brand**.	Negate_Preference		
Salesperson	Would you like to tell me your preferred pattern? I can make recommendations aimly.	Ask_Preference	<@1062>	Brand: Brain Puzzles
Customer	**Simple pattern** is the best choice for me.	Answer_Preference		
Salesperson	Maybe you can consider the light grey plain blouse <@1062>.	Recommend_Item	<@1062>	Pattern: plain
Customer	Thank you! I am very satisfied about your recommendation.	Respond_Item		

Figure 21: Example of one complete dialog section in SURE dataset. Candidate Items and candidate attribute values in the last two columns are derived from previous rounds and situated scene. We can observe that salesperson adopt acts including *"Ask_Preference"*, *"Exclude_Preference"*, *"Display_Attribute"* to clarify customer subjective preferences highlighted by underline and narrow candidate range of items by *"Refer_Region"*.
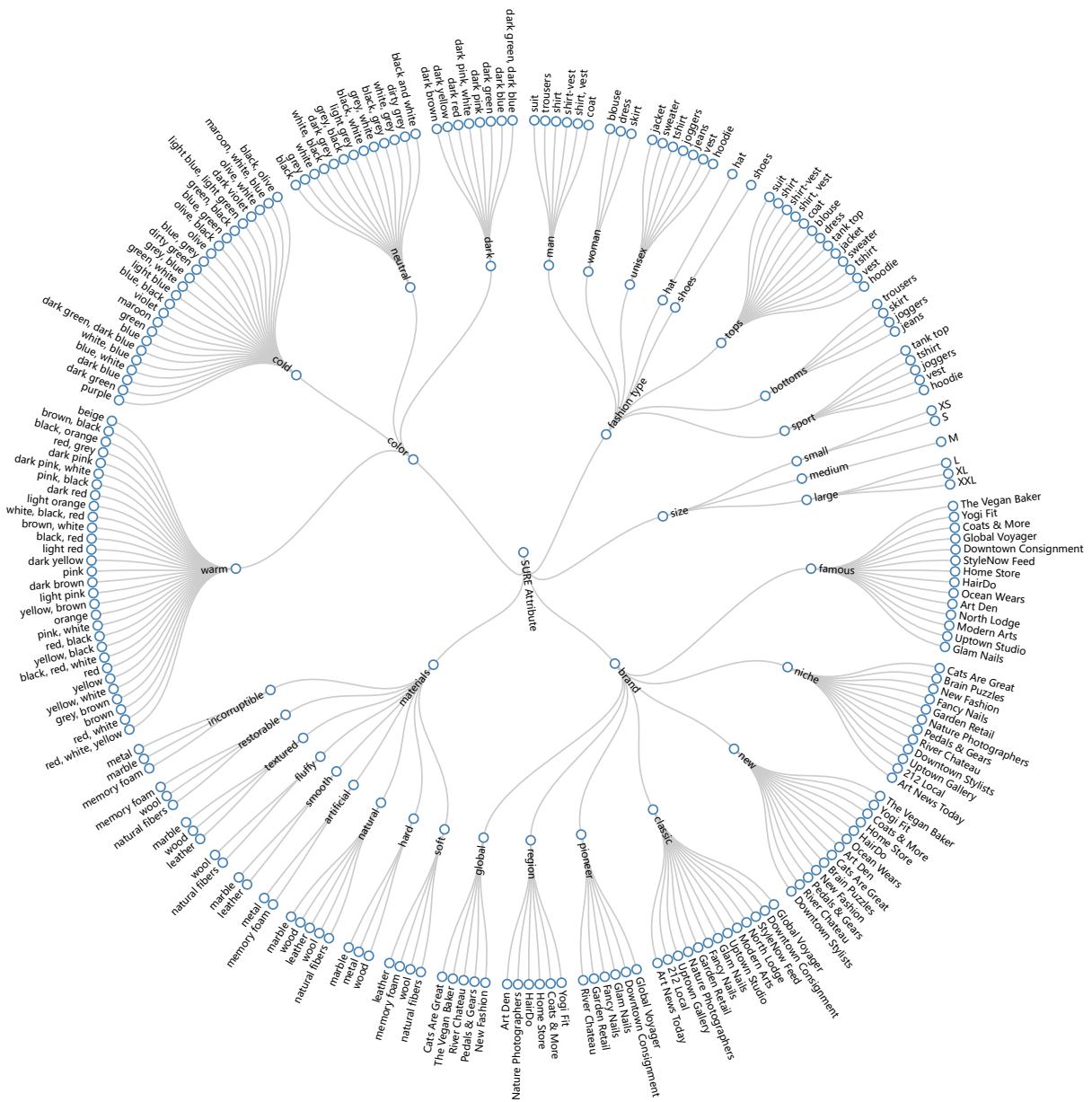
Figure 22: Some attribute categorization concepts and corresponding concrete attribute values in SURE dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section Ethics Statement

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section Ethics Statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D   ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3 and Appendix A.1

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.1

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.1

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A.1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section Ethics Statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A.1