

# Tackling Situated Multi-Modal Task-Oriented Dialogs with a Single Transformer Model

Anonymous ACL submission

## Abstract

The Situated Interactive Multi-Modal Conversations (SIMMC) 2.0 aims to create virtual shopping assistants that can accept complex multi-modal inputs, i.e. visual appearances of objects and user utterances. It consists of four subtasks, multi-modal disambiguation (MM-Disamb), multi-modal coreference resolution (MM-Coref), multi-modal dialog state tracking (MM-DST), and response retrieval and generation. While many task-oriented dialog systems usually tackle each subtask separately, we propose a jointly learned encoder-decoder that performs all four subtasks at once for efficiency. Moreover, we handle the multi-modality of the challenge by representing visual objects as special tokens whose joint embedding is learned via auxiliary tasks. This approach won the MM-Coref and response retrieval subtasks and nominated runner-up for the remaining subtasks using a single unified model. In particular, our model achieved 81.5% MRR, 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank in response retrieval task, setting a high bar for the state-of-the-art result in the SIMMC 2.0 track of the Dialog Systems Technology Challenge 10 (DSTC10).

## 1 Introduction

A task-oriented dialog system aims to assist users accomplish certain tasks, such as executing actions or retrieving specific information, with natural language conversations. The traditional approach for building task-oriented dialog systems adopts a pipelined architecture that integrates natural language understanding (NLU) module that identifies user's intent (Liu and Lane, 2016), dialog state tracking (DST) module that extracts values for slots (Henderson et al., 2013; Mrksic et al., 2017), dialog policy management (POL) module that decides system action (Wen et al., 2017), and natural language generation (NLG) module that generates appropriate system utterance according to system action (Wen et al., 2015).

With the rising interest and ubiquity of virtual reality (VR), the next generation of task-oriented virtual assistants is expected to handle conversations in a multi-modal context. For instance, a multi-modal dialog agent may help the user navigate a virtual clothing store and look for an object meeting the user's criteria. In such cases, a successful dialog agent should be able to parse and understand multi-modal contexts. To this end, SIMMC 2.0 (Kottur et al., 2021) proposes a situated multi-modal context in the form of co-observed, realistic scene set in VR stores to incorporate the complexity of multi-modal task-oriented dialogs. The multi-modal subtasks, MM-Disamb and MM-Coref, intend to test the assistant's capability to identify the need for disambiguating reference mentions and to ground them to the scene objects. While challenging, these are all essential to building a successful multi-modal dialog agent.

In this paper, we present our end-to-end, joint-learning approach to address this challenge in SIMMC 2.0. We adopt BART (Lewis et al., 2019) and attach task-specific heads so that the model can make predictions on all subtasks. To be more specific, our model performs MM-Disamb, MM-Coref, and response retrieval by the encoder and MM-DST and response generation in a string format by the decoder. We also integrate multi-modality into the model by treating scene objects as unique object tokens and coreference sentinel tokens. Our model is jointly trained on all subtasks and a few auxiliary objectives to help the model align object tokens to its attributes. For retrieval, we use in-batch negative samples for contrastive metric learning instead of creating a pool of separate training samples.

Our model was ranked at the first place for MM-Coref and response retrieval with 75.8% coreference F1, 81.5% MRR, 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank in the official evaluation of DSTC10. Moreover, our model was nominated runner-up for all other subtasks, in which we

achieved 93.8% disambiguation accuracy, 90.3% slot F1, 95.9% intent F1, and 0.295 BLEU-4. The results were obtained with only a single model and consistent with the results on the devtest (i.e. validation) set, demonstrating a robust, common representation on all subtasks learned by the model.

## 2 Related Work

Recent works on task-oriented dialog systems remove the need for a pipeline composed of NLU, DST, POL, and NLG modules by leveraging pre-trained language models (LM) that integrate all the modules in an end-to-end, auto-regressive manner (Ham et al., 2020; Hosseini-Asl et al., 2020; Yang et al., 2021). Given a dialog context, such systems sequentially generates belief state, system action, and response, making predictions based on decisions made by previous modules in the form of tokens. Some of these systems aim to learn the user preference from dialogs and recommend the object based on external knowledge base (KB) (Zhou et al., 2020).

In a similar context, building cross-modal models has recently gained a lot of attention, especially in the domain of vision and language (VL). Recent works develop VL models on top of the transformer-based (Vaswani et al., 2017) pretrained LM and vision backbones, focusing on pretraining methods to align joint embedding between different modalities. They achieve state-of-the-art performance in downstream tasks such as visual question answering (VQA), as shown in (Chen et al., 2020) and (Li et al., 2020). In this paper, we focus on understanding objects (i.e. shopping items) appearing in a scene, observed by both user and assistant. Based on the objects in a scene, the assistant needs to recommend objects or provide information of objects in the response.

## 3 SIMMC 2.0 Description

### 3.1 Dataset

SIMMC 2.0 (Kottur et al., 2021) follows the setting of SIMMC 1.0 (Moon et al., 2020), which assumed conversations occurring between a user and an assistant in a situated, co-observed VR scene. This newer iteration of the dataset lifts the limitations of SIMMC 1.0 by further capturing the complexity of multi-modal conversations: whereas SIMMC 1.0 had at most three objects in a simple, sanitized scene, SIMMC 2.0 provides a far richer visual context with 19.7 objects on average that are often

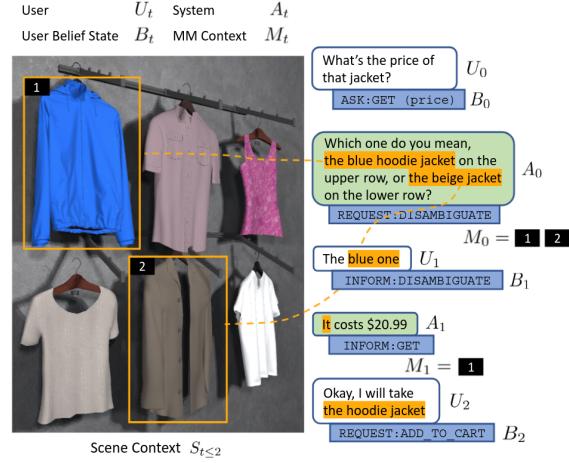


Figure 1: An instance of dialog and the corresponding scene in SIMMC 2.0. Here, the assistant asks the user to disambiguate between *the blue hoodie jacket* (indexed as 1) and *the beige jacket* (indexed as 2), grounding its mentions to the scene via multi-modal context  $M_0 = \{1, 2\}$ . Once the user chooses *the blue one*, the system retrieves the information on the disambiguated object. The multi-modal context in this case would be  $M_1 = \{1\}$ .

occluded, cluttered, or even out of view. An example dialog is shown in Figure 1.

The SIMMC 2.0 dataset consists of 11,244 dialogs split into train (65%), dev (5%), devtest (15%), and teststd (15%) sets. Each dialog includes multiple turns where each turn has grounded multi-modal context and an accompanying scene with referential indices. We shall denote a SIMMC dialog with  $r$  rounds as  $\mathcal{D} := \{(U_t, A_t, M_t, S_t, B_t)\}_{t=1}^r$ , where  $U_t$  is user utterance,  $A_t$  system utterance,  $M_t$  multi-modal context,  $S_t$  scene context, and  $B_t$  user belief state at turn  $t$ . Here,  $M_t$  is a set of object indices mentioned by the system and  $S_t$  contains the corresponding attributes and locations of all the objects in a scene. User belief state  $B_t$  is composed of dialog act (i.e. user intent) and slot (i.e. a tuple of (*slot name*, value), for instance ("price", "\$11.99")). We also define the dialog history at some turn  $T \leq r$  as  $H_T := \{U_0, A_0, M_0, \dots, U_{T-1}, A_{T-1}, M_{T-1}\}$ .

The assistant needs to make predictions conditioned on history  $H_T$ , current user utterance  $U_T$ , and the scenes up to the current turn  $S_{t \leq T}$ . The object set consists of fashion and furniture domain, where each domain has 288 and 57 items respectively. The system is allowed to look up which item is present in a scene at all time. As a side information, the metadata of each object are provided: its

161 non-visual attributes such as brand, size, customer  
162 rating and price are available for both training and  
163 inference, but looking up the visual attribute (e.g.  
164 color, pattern, materials, sleeve length) is prohibited  
165 for inference so as to make the agent reason  
166 with multi-modal information.

### 167 3.2 Subtasks

#### 168 Multi-modal disambiguation (MM-Disamb)

169 The first subtask is to identify whether the assistant  
170 should disambiguate mentions in the next turn  
171 given the dialog and multi-modal context. For in-  
172 stance, given user utterance "*How much is the pair*  
173 *on the left?*", there may be more than two pairs of  
174 pants on the left. In this case, ambiguity in refer-  
175 ence should be resolved. This can be cast into a  
176 binary classification task, and the performance is  
177 measured by accuracy.

#### 178 Multi-modal coreference resolution (MM-Coref)

179 The second subtask is to map the referential men-  
180 tions of the user utterance to the object indices  
181 in the scene. These mentions should be resolved  
182 through the linguistic context and the multi-modal  
183 context. The performance is measured by object  
184 slot F1 score.

#### 185 Multi-modal dialog state tracking (MM-DST)

186 The third subtask extends the traditional uni-modal  
187 DST to ground user belief state on the multi-modal  
188 objects. This will measure the assistant's under-  
189 standing throughout each dialog, which includes  
190 disambiguation and coreference resolution. The  
191 performance is measured by the F1 score for dialog  
192 act and slots.

193 **Response retrieval & generation** The last sub-  
194 task is to retrieve or generate appropriate system  
195 utterance. Response generation is evaluated with  
196 BLEU-4 (Papineni et al., 2002). For response re-  
197 trieval, the system is expected to choose the most  
198 relevant response from a pool of 100 candidate re-  
199 sponses. Recall@ $k$  ( $k \in \{1, 5, 10\}$ ), mean rank,  
200 and mean reciprocal rank (MRR) are used for re-  
201 trieval evaluation.

## 202 4 Integrated Transformer Model

203 Even though the setting of the dataset is similar to  
204 that of VQA where finetuning the pretrained VL  
205 models are prevalent, we chose to work with LM,  
206 representing objects by tokens. There are several  
207 reasons behind this choice. First, the vision models  
208 are usually pretrained on natural images (Lin et al.,

209 2014; Krishna et al., 2017), so finetuning them re-  
210 quires a relatively large number of training samples  
211 of 3D rendered images that are aligned properly  
212 with text. Second, in a realistic scenario where  
213 the assistant is deployed in a VR environment, the  
214 object metadata and scene graphs would be readily  
215 available as a part of the system. In this case, using  
216 a vision backbone model would be an unnecessary  
217 overhead. Lastly, we can still easily provide addi-  
218 tional supervision signals at train time for modality  
219 alignment by looking up the object metadata. For  
220 this, we represent multi-modal objects as the con-  
221 catenation of their referential indices in the scene  
222 (canonical object ID) and their absolute attribute  
223 (unique object ID).

224 We note that all of the subtasks are related to  
225 each other. For example, if the assistant decides  
226 that the user utterance needs to be disambiguated,  
227 then the appropriate system action is to respond  
228 along the line of "Which one are you referring  
229 to?". We expect that the latent representation of  
230 the multi-modal dialog learned from other subtasks  
231 will translate readily to other subtasks. Hence, we  
232 utilize hard parameter sharing (Caruana, 1993) on  
233 the encoder to jointly learn on all subtasks. This  
234 reduces not only the number of network parameters,  
235 but also the risk of overfitting (Baxter, 1997).

236 Moreover, we decide to view MM-Coref as a  
237 type of set prediction (Zaheer et al., 2017), where  
238 joint learning of set cardinality and state distribu-  
239 tion has been shown effective (Rezatofighi et al.,  
240 2018). Hence, we define an additional empty coref-  
241 erence target prediction (Empty-Coref), a simpli-  
242 fied cardinality prediction task that outputs whether  
243 the current user utterance has no MM-Coref tar-  
244 get. Moreover, we perform a supervised learning  
245 on object attributes to help align object-language  
246 modalities.

247 In order to harness the power of NLU/NLG ca-  
248 pabilities demonstrated by pretrained transformer  
249 encoder-decoder, we adopt BART (Lewis et al.,  
250 2019) as the pretrained language backbone. We  
251 attach classification heads for MM-Disamb and  
252 MM-Coref subtasks at the encoder and LM head  
253 for MM-DST and response generation at the de-  
254 coder. We also perform retrieval by computing the  
255 dot product between representation vectors of re-  
256 sponse candidates and multi-modal dialog context.  
257 The overview of the model is provided in Figure 2.

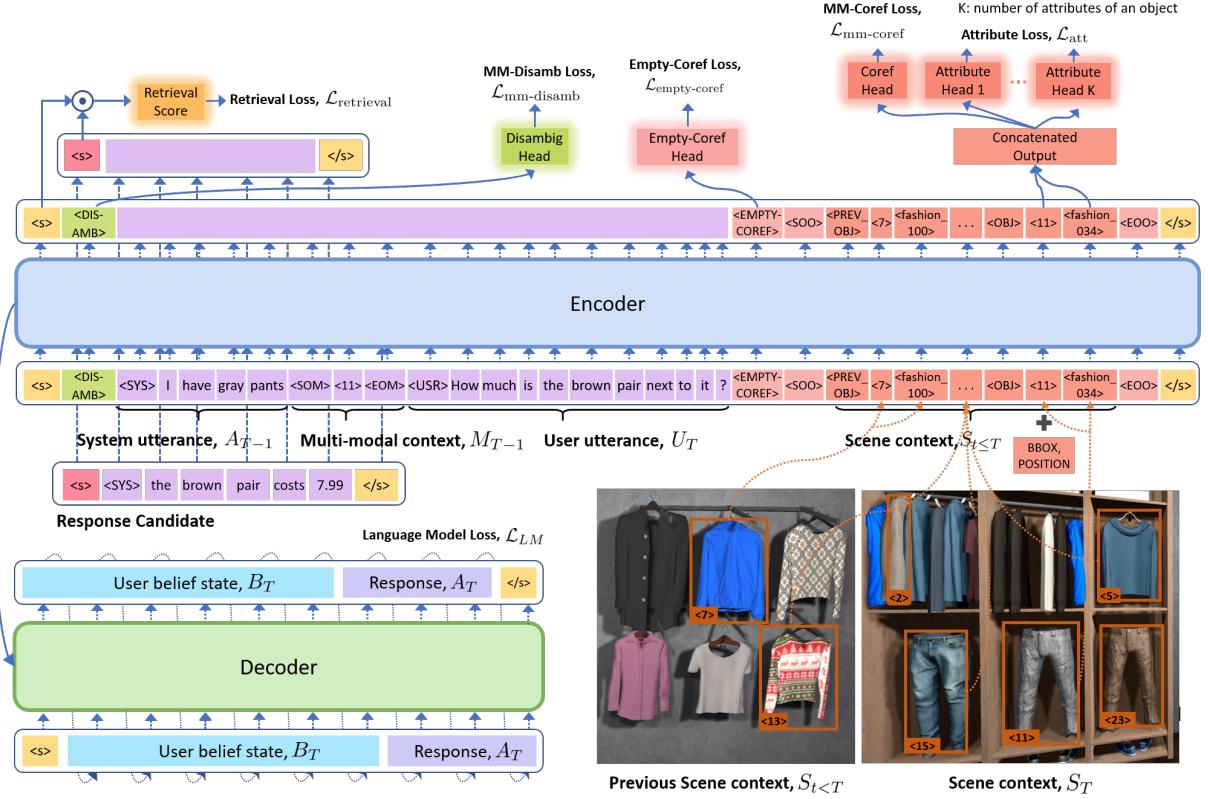


Figure 2: Overview of the jointly learned multi-tasking BART. For  $H_T$ , we show only the last turn without user utterance due to space limit. The details on the loss functions are provided in model specifics. Each scene object is represented by the concatenation of scene canonical object ID token (e.g.  $<11>$ ) and unique object token (e.g.  $<\text{fashion\_123}>$ ). It is then passed through MM-Coref and attribute classification head. MM-DST and response generation subtasks are approached in terms of auto-regressive LM.

## 4.1 Input Representation

For all of the subtasks, we define our input to be a simple concatenation  $x := [H_T; U_T; S_{t \leq T}]$  with separators. We define  $H_T$  to be the dialog history up to 2 turns to limit the length of input, i.e.  $\{U_{T-2}, A_{T-2}, M_{T-2}, U_{T-1}, A_{T-1}, M_{T-1}\}$ . SIMMC 2.0 assumes that utterances may mention objects that are not in the current scene  $S_T$  but in the previously observed scene  $S_{t < T} \neq S_T$ . Hence, our model integrates the objects from the previous scene that are not in the current scene. We find that our scene representation by enumerating all objects is a simple yet effective method for the model to understand the multi-modal context. An exemplar input is provided in Table 1.

### 4.1.1 Canonical object ID token

A canonical object ID token takes the form of  $<\backslash d+>$  (e.g.  $<32>$ ). This provides a relational context of the object within the scene, grounding each object to its scene object index provided in the dataset. This scheme was also used in the base-

line code for SIMMC 2.0 (Kottur et al., 2021), but without any association to object attributes. In our method, this token intends to provide contextual information of the object alongside its absolute attributes (unique object token), allowing the assistant to make connections between different modalities.

### 4.1.2 Unique object ID token

Unique object ID token takes the form of  $<\{\text{domain}\}_\backslash d+>$  (e.g.  $<\text{fashion\_123}>$ ,  $<\text{furniture\_028}>$ ). The digits following the domain specifier denote index of the unique object in that domain. This token intends to provide an embedding which encodes the visual (e.g. type, color, material) and non-visual (e.g. price, customer rating) attributes unique to each object.

### 4.1.3 Separator tokens

We define several separator tokens to delimit different components of the multi-modal dialogs. We use  $<\text{SOM}>$ ,  $<\text{EOM}>$  for the start and the end of multi-modal context and  $<\text{SOO}>$ ,  $<\text{EOO}>$  for the

Common Input ( $x$ )	
$U_{T-1}$	<USR> What are the good hoodies around here?
$A_{T-1}$	<SYS> I advise you consider the solid green one.
$M_{T-1}$	<SOM> <56> <EOM>
$U_T$	<USR> I do like solid colors, but I'm looking for something with excellent ratings.
$S_{t < T}$	<SOO> <PREV_OBJ> <12> <fashion_142> <PREV_OBJ> <13> <fashion_058>
$S_T$	<OBJ> <56> <fashion_269> <OBJ> <85> <fashion_007> <EOO>
Generation Target	
$B_T$	<SOB> INFORM:GET <customerReview> good <pattern> plain <type> hoodie <EOB>
$A_T$	In fact, that green hoodie is very highly rated.
Response Candidate	<SYS> In fact, that green hoodie is very highly rated.

Table 1: Example input representations for our model. We show only up to last 1 turn due to space limit. Thus, the common input  $x$  is a concatenation  $[H_T; U_T; S_{t \leq T}]$  where  $H_T = \{U_{T-1}, A_{T-1}, M_{T-1}\}$ . Here, we separate the previous scene history  $S_{t < T}$  to show how we handle out-of-view objects. The generation target is a concatenation  $[B_T; A_T]$ , which is used by the decoder. The response candidate is  $A_T$  with speaker identifier <SYS> prepended.

start and the end of scene objects. Within the scene context, <OBJ> token is used as a separator token between objects, which are represented by the concatenation of a canonical object ID token and a unique object ID token. We also mark the objects from the previous scene with <PREV\_OBJ> instead of <OBJ>. For generation target, we mark the start and the end of the user belief state with <SOB>, <EOB>.

#### 4.1.4 Encoding object locations

For the assistant to understand the spatial relation among objects within the scene, we must incorporate encoded representation of location of each object. We follow the commonly used techniques in VL models (Li et al., 2020; Chen et al., 2020; Zhang et al., 2021) for encoding object locations with the bounding box information. Given a bounding box represented by its upper-left and lower-right vertices,  $(x_1, y_1)$  and  $(x_2, y_2)$ , with height  $h$  and width  $w$ , we encode its location as tuple  $(x_1/w - 0.5, y_1/h - 0.5, x_2/w - 0.5, y_2/h - 0.5, (x_2 - x_1)(y_2 - y_1)/(h \cdot w))$ . This is passed through a location embedding layer (a fully-connected layer followed by layer norm) to be added with the canonical object ID token encoding.

## 4.2 Model Specifics

### 4.2.1 Binary prediction for MM-Disamb and MM-Coref

We formulate MM-Disamb as a binary classification on the pooled output of the encoder from the pooling token <DISAMB>. The binary head for MM-Disamb should predict true if the current user utterance  $U_T$  needs to be disambiguated and false otherwise.

For MM-Coref, we make binary predictions on all objects in  $S_{t \leq T}$ . We do so by passing the concatenated canonical object (e.g. <11>) and unique object ID (e.g. <fashion\_001>) encoder output of each object through a binary classification head. The MM-Coref head will predict true if the current user utterance mentions that object and false otherwise. We use a simple cross-entropy loss for both MM-Disamb and MM-Coref, denoted  $\mathcal{L}_{\text{mm-disamb}}$  and  $\mathcal{L}_{\text{mm-coref}}$ .

### 4.2.2 Auto-regressive LM for MM-DST and generation

We also approach MM-DST and response generation subtasks with auto-regressive LM following the recent approaches in end-to-end dialog systems. For MM-DST and response generation, we use the standard left-to-right LM loss (Bengio et al., 2003).

$$\mathcal{L}_{\text{LM}} = \sum_{i=1}^L -\log P(\omega_i | \omega_1, \dots, \omega_{i-1}),$$

where  $\omega_i$  is the  $i$ -th target token and  $L$  the total length of the target.

### 4.2.3 In-batch negative samples for retrieval

For response retrieval task, we make use of in-batch negative samples for contrastive learning on similarity metrics. We treat the system responses of the other samples in the batch formatted according to Table 1 as in-batch negatives. We then pool the encoder outputs of the input and the response candidates with BART bos token, i.e. <s>, to compute their dot product, so that the correct scene-response candidate pair stays close and the incorrect pairs stay apart. We use multi-class cross-entropy loss

300  
301  
302  
303  
304  
305  
306  
307  
308

309

310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325

326

327  
328

329  
330  
331  
332  
333  
334

335  
336  
337  
338  
339  
340  
341  
342  
343  
344

345  
346  
347  
348  
349  
350  
351

352

353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365

366 applied to dot-product similarities, i.e.,

$$367 \quad \mathcal{L}_{\text{retrieval}} = -\log \frac{\exp(\mathbf{x} \cdot \mathbf{a}^+)}{\sum_{\mathbf{a}^- \in B^-(\mathbf{x}) \cup \{\mathbf{a}^+\}} \exp(\mathbf{x} \cdot \mathbf{a}^-)},$$

368 where  $\mathbf{a}^+$  is the positive response sample of the  
369 input  $\mathbf{x}$  and  $B^-(\mathbf{x})$  the set of in-batch negative  
370 responses (assume  $\mathbf{x}$ ,  $\mathbf{a}^+$ , and  $\mathbf{a}^-$  are pooled rep-  
371 resentations from the encoder). We formulate the  
372 task loss  $\mathcal{L}_{\text{task}}$  as a linear combination of losses  
373 from each subtask.

$$374 \quad \mathcal{L}_{\text{task}} = \lambda_{\text{LM}} \mathcal{L}_{\text{LM}} + \lambda_{\text{mm-disamb}} \mathcal{L}_{\text{mm-disamb}} \\ + \lambda_{\text{mm-coref}} \mathcal{L}_{\text{mm-coref}} + \lambda_{\text{retrieval}} \mathcal{L}_{\text{retrieval}} \quad (1)$$

### 375 4.3 Auxiliary Tasks

#### 376 4.3.1 Binary prediction for Empty-Coref

377 We define an additional Empty-Coref task, in which  
378 the assistant predicts whether the current dialog  
379 turn has MM-Coref targets. This can be seen as a  
380 simpler version of set cardinality prediction. We  
381 find this additional signal for coreference resolu-  
382 tion, denoted  $\mathcal{L}_{\text{empty-coref}}$ , is advantageous in boost-  
383 ing MM-Coref performance, a type of set predic-  
384 tion task. Moreover, MM-Coref sometimes pre-  
385 dicted targets when there is actually none, so we  
386 override any MM-Coref predictions if the Empty-  
387 Coref prediction is true (i.e. there is no coreference  
388 target). For this, we use <EMPTY\_COREF> for  
389 pooling. At inference time, . We use a binary  
390 cross-entropy loss for  $\mathcal{L}_{\text{empty-coref}}$ .

#### 391 4.3.2 Encoding object attributes

392 We encode object attributes by providing additional  
393 supervision signal during training. We do so by  
394 simply training to classify each object to its cor-  
395 responding visual and non-visual attributes such  
396 as color, price, and customer ratings. Each object  
397 is represented as a concatenation of its canonical  
398 object ID and unique object token as in MM-Coref  
399 (refer to Figure 2). Each attribute head predicts a  
400 categorical class for each corresponding object, for  
401 example, if <fashion\_001> is a grey jacket, the  
402 color-attribute head predicts the class of grey and  
403 the type-attribute head predicts the class of jacket.

404 Let  $\mathcal{O}_{t \leq T}$  be the set of objects in the scene his-  
405 tory,  $S_{t \leq T}$ . We denote attribute multi-class classi-  
406 fication loss  $\mathcal{L}_{\text{att}}$  for all objects in  $\mathcal{O}_{t \leq T}$ ,

$$407 \quad \mathcal{L}_{\text{att}} = \sum_{j \in \mathcal{O}_{t \leq T}} \sum_{k=1}^K \sum_{c \in \mathcal{C}_k} -\mathbb{1}\{c = y_{jk}\} \log P(c),$$

408 where  $K$  is the number of attributes,  $\mathcal{C}_k$  the set  
409 of all classes of the  $k$ -th attribute,  $y_{jk}$  the label of  
410 the  $k$ -th attribute of the  $j$ -th object, and  $\mathbb{1}\{\cdot\}$  is an  
411 indicator function.

412 As a result, the auxiliary loss  $\mathcal{L}_{\text{aux}}$  is defined  
413 as the weighted sum of attribute loss and empty-  
414 coreference prediction loss:

$$415 \quad \mathcal{L}_{\text{aux}} = \lambda_{\text{att}} \mathcal{L}_{\text{att}} + \lambda_{\text{empty-coref}} \mathcal{L}_{\text{empty-coref}} \quad (2)$$

416 In summary, we minimize the total loss  $\mathcal{L}_{\text{total}}$ ,  
417 which is the sum of the task loss  $\mathcal{L}_{\text{task}}$  from Equa-  
418 tion 1 and the auxiliary loss  $\mathcal{L}_{\text{aux}}$  from Equation 2.

$$419 \quad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{aux}}$$

## 5 Experiments

### 420 5.1 Experimental Setup

421 Our model is built on top of 24-layer BART from  
422 HuggingFace (facebook/bart-large) (Wolf  
423 et al., 2019).<sup>1</sup> We finetune the model for 10 epochs  
424 with an initial learning rate of 5e-5 and a batch  
425 size of 16 with AdamW optimizer (Loshchilov and  
426 Hutter, 2018). We also use linear warmup schedule  
427 with 8000 warmup steps and clip gradient norms at  
428 1.0. For decoding, we use top- $p$  sampling (Holtz-  
429 man et al., 2020) with  $p = 0.9$  to generate the user  
430 belief state and system response. We choose the  
431 best checkpoint evaluated at every 1000 steps on  
432 the devtest set. For joint learning coefficients, see  
433 Appendix A.

### 435 5.2 Baselines

436 The challenge organizers provided two baseline  
437 models: an end-to-end GPT-2 (Radford et al., 2019)  
438 and multi-modal transformer networks (MTN) (Le  
439 et al., 2019). The baseline models do not explic-  
440 itly use object attributes and model each subtask  
441 separately, except for MM-Coref, MM-DST, and  
442 response generation. GPT-2 baseline generates the  
443 user belief state, coreference objects (in the form  
444 of canonical object IDs), and response in an end-  
445 to-end manner. MTN baseline conditions on the  
446 scene image and dialog history then generate the  
447 user belief state and response using a multi-model  
448 transformer. The MTN baseline only implements  
449 MM-DST and response generation.

<sup>1</sup><https://github.com/huggingface/transformers>

Models	#1 Disamb.	#2 MM-Coref	#3 MM-DST		#4-1 Res. Retrieval					#4-2 Res. Gen.
	Accuracy ( $\uparrow$ )	Obj. F1 ( $\uparrow$ )	Slot F1 ( $\uparrow$ )	Act. F1 ( $\uparrow$ )	MRR ( $\uparrow$ )	R@1 ( $\uparrow$ )	R@5 ( $\uparrow$ )	R@10 ( $\uparrow$ )	M. Rank ( $\downarrow$ )	BLEU-4 ( $\uparrow$ )
GPT-2 Baseline	73.8%	36.6%	81.7%	94.5%	8.8%	2.6%	10.7%	18.4%	38.0	0.192
MTN Baseline	-	-	74.8%	93.4%	-	-	-	-	-	0.217
bart-large	92.7%	74.3%	89.2%	96.2%	80.7%	71.1%	94.4%	98.3%	1.93	0.314
- (1)	92.6%	68.3%	87.3%	96.0%	80.7%	70.7%	94.3%	98.0%	1.98	0.304
- (2)	92.6%	74.6%	89.0%	96.0%	80.6%	70.1%	94.4%	98.4%	1.92	0.305
- (1), (2)	93.0%	48.7%	87.6%	96.1%	81.1%	70.6%	94.8%	98.6%	1.88	0.302

Table 2: Overall and ablation study results on the devtest set. GPT-2 and MTN are the baselines provided by the organizers, which are separately trained on each subtask. The MTN baseline performs only MM-DST and response generation. For the ablation study results, - (1) represents removing attribute classification auxiliary loss, - (2) represents removing Empty-Coref prediction auxiliary loss, and - (1),(2) represents removing both.

Entry ID	#1 Disamb.	#2 MM-Coref	#3 MM-DST		#4-1 Res. Retrieval					#4-2 Res. Gen.
	Accuracy ( $\uparrow$ )	Obj. F1 ( $\uparrow$ )	Slot F1 ( $\uparrow$ )	Act. F1 ( $\uparrow$ )	MRR ( $\uparrow$ )	R@1 ( $\uparrow$ )	R@5 ( $\uparrow$ )	R@10 ( $\uparrow$ )	M. Rank ( $\downarrow$ )	BLEU-4 ( $\uparrow$ )
1	-	52.1%	89.1%	96.3%	53.5%	42.8%	65.4%	74.9%	11.9	0.285
2	89.5%	42.2%	87.8%	96.2%	61.2% <sup>†</sup>	49.6% <sup>†</sup>	74.7% <sup>†</sup>	84.5% <sup>†</sup>	6.6 <sup>†</sup>	0.256
3 (Ours)	93.9% <sup>†</sup>	<b>75.8%</b>	90.3% <sup>†</sup>	95.9% <sup>†</sup>	<b>81.5%</b>	<b>71.2%</b>	<b>95.0%</b>	<b>98.2%</b>	<b>1.9</b>	0.295 <sup>†</sup>
4	93.8% <sup>†</sup>	56.4%	89.3%	96.4%	32.0%	19.9%	41.8%	61.2%	12.9	<b>0.322</b>
5	<b>94.7%</b>	59.5%	<b>91.5%</b>	<b>96.0%</b>	-	-	-	-	-	-
6	93.1%	57.3%	-	-	-	-	-	-	-	-
7	93.1%	68.2%	4.0%	41.4%	-	-	-	-	-	0.297 <sup>†</sup>
8	-	73.3% <sup>†</sup>	-	-	-	-	-	-	-	-
9	93.6% <sup>†</sup>	68.2%	87.7%	95.8%	-	-	-	-	-	<b>0.327</b>

Table 3: The official leaderboard of DSTC10 on the teststd set. The subtask winners are bold-faced and runner-ups are marked with  $\dagger$ . “-” means that the entry did not participate in that subtask.

### 5.3 Results

The results on the devtest (validation) and teststd (test) splits are shown in Table 2 and 3, respectively. On devtest set, our proposed model outperforms the baselines by a large margin. Our proposed model based on bart-large was ranked at the first place with **75.8%** coreference F1 in MM-Coref. This demonstrates that our method of injecting object attributes to the model was effective, providing a richer context about the scene and its objects to the assistant. Furthermore, our model was declared winner in the response retrieval subtask with 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank. This is a remarkable performance compared to existing methods such as bi- and poly-encoders (Humeau et al., 2020), despite the fact that we only used a single encoder built into the model to encode both the dialog context and candidates.

Our method of representing scene and learning joint embedding between dialog and scene successfully captured fine-grained information on the scene objects. This allows for the model to attend and focus on objects that are being mentioned in the conversation, learning to choose the right response most of the time. Moreover, our model

showed competitive performance and was declared runner-up in all remaining sub-tasks, in which we achieved 93.8% disambiguation accuracy, 90.3% slot F1, 95.9% intent F1, and 0.295 BLEU-4 with a single model.

### 5.4 Ablation Studies

We conducted ablation studies on auxiliary objectives, namely removing (1) attribute classification and (2) Empty-Coref target set prediction during training, to observe their effectiveness in the assistant’s understanding of multi-modality and overall performance in the four subtasks. All ablation models are trained in the same setting as in the earlier part of this section. The results are shown in Table 2.

#### 5.4.1 Attribute classification

We remove the attribute classification loss  $\mathcal{L}_{att}$  from the main loss. We observe that removing attribute classification results in a significant drop in the MM-Coref performance by 6.0%. The performance degradation demonstrates the effectiveness of the attribute classification objective. Furthermore, we observe noticeable drop in performance in other subtasks, especially the slot prediction of MM-DST subtasks. Here, understanding and dis-

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

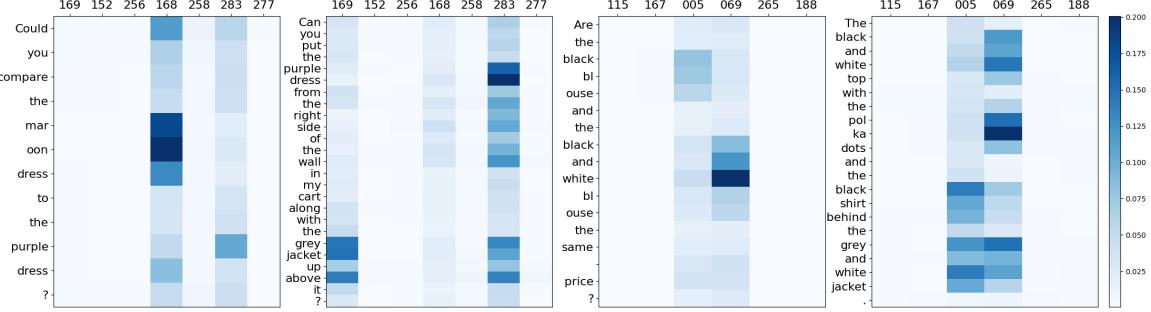


Figure 3: Attention maps between utterance and fashion unique IDs. The object attributes are given in Table 4.

tinguishing different objects by their attributes are crucial in predicting correct slot values.

#### 5.4.2 Empty-Coref prediction

We remove the Empty-Coref loss  $\mathcal{L}_{\text{empty-coref}}$  from the main loss. We observe no significant difference from the full model. In fact, we observe a better performance in MM-coref possibly because there is no interference in subtask losses from the Empty-Coref objective. However, Empty-Coref prediction becomes important when the attribute classification objective is removed. The model only achieves 48.7% coreference F1 as opposed to 68.3% with Empty-Coref. This suggests that this auxiliary sub-task provides a useful signal for MM-Coref. We also see overall improvements in other related sub-tasks such as MM-DST and response retrieval.

## 6 Visualizing attention

We visualize the learned attention between the two different modalities. Figure 3 shows attention heatmaps from the fifth head in last encoder layer. The rows indicate extracted utterance from  $[H_T; U_T]$  and the columns unique object IDs in  $S_{t \leq T}$ . Table 4 lists the visual-metadata of these objects. According to the visualization, the model was able to make a connection between natural language attributes mentioned in the dialog and the corresponding unique object ID token.

## 7 Conclusion

In this paper, we propose a multi-modal task-oriented dialog system based on BART that can perform all SIMMC 2.0 subtasks at once. Our model overcomes the challenge of adopting severely occluded, 3D rendered artificial images to vision models by integrating multi-modal objects as special tokens. In addition to joint learning of all subtasks, we introduce Empty-Coref and attribute classification as auxiliary tasks to directly align objects

fashion unique ID	color	type	pattern
169	light grey	jacket	plain
152	black, white	blouse	vertical
256	black	sweater	knit
168	maroon	dress	plain
258	brown	dress	plain
283	purple	dress	plain
277	grey	trousers	heavy stripes
115	grey, white	jacket	twin colors
167	blue	jacket	plain
005	black	blouse	velvet
069	black, white	blouse	spots
265	blue	jeans	denim
188	blue	trousers	plain

Table 4: Visual metadata of unique object IDs shown in Figure 3.

to their corresponding attributes. We observe that these additional subtasks are crucial in building a successful multi-modal assistant for SIMMC 2.0. Our model is able to perform competitively in all of the subtasks with a single model, ranking first place for MM-Coref and response retrieval and runner-up for the remaining subtasks in DSTC10.

Despite the success in SIMMC 2.0, our approach has a few limitations. Most notably, our approach cannot be applied to cases with novel objects at inference, i.e. the objects that don't appear in the database at training. As such, it relies on latent object features learned from linguistic description for retrieving the requested object attributes. Our method also does not fully capture the semantic locality of objects within the scene (e.g. on the table, in the closet, etc.). We believe that these limitations can be addressed by training with a larger amount of data and including visual features in the multi-modal context as part of the input to the transformer.

## References

- 559 Jonathan Baxter. 1997. A bayesian/information the-  
560oretic model of learning to learn via multiple task  
561 sampling. *Mach. Learn.*, 28(1):7–39.  
562
- 563 Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and  
564 Christian Janvin. 2003. A neural probabilistic lan-  
565 guage model. *J. Mach. Learn. Res.*, 3:1137–1155.  
566
- 567 Rich Caruana. 1993. Multitask learning: A knowledge-  
568 based source of inductive bias. In *Machine Learning,*  
569 *Proceedings of the Tenth International Conference,*  
570 *University of Massachusetts, Amherst, MA, USA,* June 27-29, 1993, pages 41–48. Morgan Kaufmann.  
571
- 572 Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed  
573 El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and  
574 Jingjing Liu. 2020. Uniter: Universal image-text  
575 representation learning. In *European conference on*  
*computer vision*, pages 104–120. Springer.  
576
- 577 Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and  
578 Kee-Eung Kim. 2020. End-to-end neural pipeline  
579 for goal-oriented dialogue systems using gpt-2. In  
580 *Proceedings of the 58th Annual Meeting of the Association*  
*for Computational Linguistics*, pages 583–592.  
581
- 582 Matthew Henderson, Blaise Thomson, and Steve J.  
583 Young. 2013. Deep neural network approach for  
584 the dialog state tracking challenge. In *Proceedings*  
585 *of the SIGDIAL 2013 Conference, The 14th Annual*  
586 *Meeting of the Special Interest Group on Discourse*  
587 *and Dialogue, 22-24 August 2013, SUPELEC, Metz,*  
588 *France*, pages 467–471. The Association for Computer  
589 Linguistics.  
590
- 591 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and  
592 Yejin Choi. 2020. The curious case of neural text  
593 degeneration. In *8th International Conference on*  
*Learning Representations, ICLR 2020, Addis Ababa,*  
*Ethiopia, April 26-30, 2020.* OpenReview.net.  
594
- 595 Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu,  
596 Semih Yavuz, and Richard Socher. 2020. A simple  
597 language model for task-oriented dialogue. In *Ad-*  
598 *vances in Neural Information Processing Systems 33:*  
599 *Annual Conference on Neural Information Process-*  
600 *ing Systems 2020, NeurIPS 2020, December 6-12,*  
*2020, virtual.*  
601
- 602 Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux,  
603 and Jason Weston. 2020. Poly-encoders: Architec-  
604 tures and pre-training strategies for fast and accurate  
605 multi-sentence scoring. In *8th International Confer-  
606 ence on Learning Representations, ICLR 2020, Addis*  
*Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net.  
607
- 608 Satwik Kottur, Seungwhan Moon, Alborz Geramifard,  
609 and Babak Damavandi. 2021. SIMMC 2.0: A task-  
610 oriented dialog dataset for immersive multimodal  
611 conversations. In *Proceedings of the 2021 Confer-  
612 ence on Empirical Methods in Natural Language Pro-  
613 cessing, EMNLP 2021, Virtual Event / Punta Cana,*  
614 *Dominican Republic, 7-11 November, 2021*, pages  
615 4903–4912. Association for Computational Linguis-  
tics.  
616
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson,  
617 Kenji Hata, Joshua Kravitz, Stephanie Chen,  
618 Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.  
2017. Visual genome: Connecting language and vi-  
619 sion using crowdsourced dense image annotations.  
*International journal of computer vision*, 123(1):32–  
620 73.  
621
- Hung Le, Doyen Sahoo, Nancy F. Chen, and Steven  
C. H. Hoi. 2019. Multimodal transformer networks  
622 for end-to-end video-grounded dialogue systems. In  
623 *Proceedings of the 57th Conference of the Associa-  
624 tion for Computational Linguistics, ACL 2019, Flo-  
625 rence, Italy, July 28- August 2, 2019, Volume 1: Long  
626 Papers*, pages 5612–5623. Association for Computa-  
627 tional Linguistics.  
628
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan  
629 Ghazvininejad, Abdelrahman Mohamed, Omer Levy,  
630 Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: De-  
noising sequence-to-sequence pre-training for natural  
631 language generation, translation, and comprehension.  
*arXiv preprint arXiv:1910.13461.*  
632
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang,  
633 Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong  
634 Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-  
635 semantics aligned pre-training for vision-language  
636 tasks. In *European Conference on Computer Vision*,  
637 pages 121–137. Springer.  
638
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James  
639 Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,  
640 and C Lawrence Zitnick. 2014. Microsoft coco:  
641 Common objects in context. In *European confer-  
642 ence on computer vision*, pages 740–755. Springer.  
643
- Bing Liu and Ian R. Lane. 2016. Attention-based recur-  
644 rent neural network models for joint intent detection  
645 and slot filling. In *Interspeech 2016, 17th Annual*  
646 *Conference of the International Speech Communica-  
647 tion Association, San Francisco, CA, USA, September*  
8-12, 2016, pages 685–689. ISCA.  
648
- Ilya Loshchilov and Frank Hutter. 2018. Fixing weight  
649 decay regularization in adam.  
650
- Seungwhan Moon, Satwik Kottur, Paul A. Crook,  
651 Ankita De, Shivani Poddar, Theodore Levin, David  
652 Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon  
653 Cho, Rajen Subba, and Alborz Geramifard. 2020.  
654 Situated and interactive multimodal conversa-  
655 tions. In *Proceedings of the 28th International Confer-  
656 ence on Computational Linguistics, COLING 2020,*  
657 *Barcelona, Spain (Online), December 8-13, 2020*,  
658 pages 1103–1121. International Committee on Com-  
659 putational Linguistics.  
660
- Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien  
661 Wen, Blaise Thomson, and Steve J. Young. 2017.  
662 Neural belief tracker: Data-driven dialogue state  
663 tracking. In *Proceedings of the 55th Annual Meet-  
664 ing of the Association for Computational Linguistics,*  
665 *ACL 2017, Vancouver, Canada, July 30 - August 4,*  
666 *Volume 1: Long Papers*, pages 1777–1788. Associa-  
667 tion for Computational Linguistics.  
668

674	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.	In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	730
675			731
676			732
677			733
678			734
679	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.	<i>OpenAI blog</i> , 1(8):9.	735
680			
681			
682			
683	Seyed Hamid Rezatofighi, Anton Milan, Qinfeng Shi, Anthony R. Dick, and Ian D. Reid. 2018. Joint learning of set cardinality and state distribution.	In <i>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</i> , pages 3968–3975. AAAI Press.	736
684			737
685			738
686			739
687			740
688			741
689			742
690			
691			
692			
693	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.	In <i>Advances in neural information processing systems</i> , pages 5998–6008.	743
694			744
695			745
696			746
697			
698	Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems.	<i>arXiv preprint arXiv:1508.01745</i> .	747
699			748
700			749
701			750
702			751
703	Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve J. Young. 2017. Latent intention dialogue models.	In <i>Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017</i> , volume 70 of <i>Proceedings of Machine Learning Research</i> , pages 3732–3741. PMLR.	752
704			
705			
706			
707			
708			
709			
710	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing.	<i>CoRR</i> , abs/1910.03771.	753
711			754
712			755
713			
714			
715			
716	Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2.	In <i>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</i> , pages 14230–14238. AAAI Press.	756
717			757
718			758
719			759
720			760
721			761
722			762
723			763
724			764
725	Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep sets.	In <i>Advances in Neural Information Processing Systems</i> , volume 30. Curran Associates, Inc.	765
726			766
727			
728			
729			
730	Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models.	In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 5579–5588.	735
731			
732			
733			
734			
735			
736	Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion.	In <i>Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining</i> , pages 1006–1014.	736
737			737
738			738
739			739
740			740
741			741
742			742
743			
744			
745			
746			
747			
748			
749			
750			
751			
752			
753			
754			
755			

## A Implementation Details

### A.1 Joint Learning Coefficients

We train the model jointly for all subtasks with two more auxiliary tasks. The total loss is the sum of the subtask loss from Equation 1 and the auxiliary loss from Equation 2. We found the optimal combination of coefficients via grid search with the following choice of coefficient, while fixing  $\mathcal{L}_{LM}$  to 1.0. Table 5 shows the final choice of coefficients used for training.

- $\lambda_{mm\text{-disamb}}, \lambda_{att}, \lambda_{empty\text{-coref}} \in \{0.1, 0.2, 0.3\}$
- $\lambda_{mm\text{-coref}} \in \{0.8, 1.0\}$
- $\lambda_{retrieval} \in \{0.2, 0.4, 0.6, 0.8\}$

$\lambda_{LM}$	$\lambda_{mm\text{-disamb}}$	$\lambda_{mm\text{-coref}}$
1.0	0.1	0.8
$\lambda_{retrieval}$	$\lambda_{att}$	$\lambda_{empty\text{-coref}}$
0.4	0.1	0.1

Table 5: Optimal choice of joint learning coefficients used for training.

### A.2 Task-Specific Heads

Object related classification heads (MM-Coref, attribute classification) have input dimension of twice the model dimension of BART (i.e. 2048 for bart-large). For MM-Disamb and Empty-Coref classification head, we use a single linear layer with softmax activation. For MM-Coref and attribute classification, we use an intermediate layer with the same hidden size as the input dimension, which is followed by a linear layer with softmax activation.

$U_T$	<i>Can you show me a jacket that goes with my wardrobe?</i>
Generated $A_T$	<i>There are several jackets which match your criteria, including a black jacket, a grey and white jacket, and a light grey jacket.</i>
GT $A_T$	<i>There are a couple of jackets throughout the store that you might like. What do you think?</i>
$U_T$	<i>Can you recommend a jacket to match my clothes?</i>
Generated $A_T$	<i>Do you like the blue one on the bottom row of the wall? It is a jacket.</i>
GT $A_T$	<i>Do you like the brown one on the top left of the wall or the blue one directly under it? They are jackets.</i>
$U_T$	<i>Can you help me find a cheap blouse?</i>
Generated $A_T$	<i>What do you think of the grey and white blouse <b>on the wall</b>, the maroon, white, and blue <b>on the rack</b>, or the white and grey <b>on the back wall</b>?</i>
GT $A_T$	<i>There are three options <b>on the left wall</b>: a grey and white, a white and grey, and a maroon, white, and blue?</i>

Table 6: Comparison between Generated  $A_T$  and ground truth (GT) in the case of recommendation. The differences between the generated and the ground-truth (in terms of semantics) are bolded.

$U_T$	<i>What is the rating and price of that gray and white shirt?</i>
Generated $A_T$	<i>Which one?</i>
GT $A_T$	<i>Which one?</i>
$U_T$	<i>The gray and white one hanging on the left.</i>
Generated $A_T$	<i>The grey one is \$239.99 and the brown one is only \$59.99.</i>
GT $A_T$	<i>Sure! The brown one is \$59.99 and the grey one is \$239.99.</i>
$U_T$	<i>The green and white blouse in the closet and the red, white, and yellow blouse on the rack.</i>
Generated $A_T$	<i>The former is from <b>Downtown Consignment</b> and is available in S, XL, and XXL. The latter is from <b>Pedals &amp; Gears</b> and is in stock in XS, S and XL.</i>
GT $A_T$	<i>The first is available in XS, S, and XL and is from <b>The Vegan Baker</b>. The other is in XS, S, and XL and is from <b>Downtown Consignment</b>.</i>

Table 7: Comparison between Generated and GT  $A_T$  in the case of disambiguation and informing object attributes. The differences between the generated and the ground-truth (in terms of semantics) are bolded.

## B Qualitative analysis

A successful multi-modal agent should be able to recommend objects that fit the user’s requested criteria within the scene context, understand the locations of the objects, and provide the requested information on the object such as ratings and price. We qualitatively analyze the generated system utterances to check whether our model can capture the object attributes along with spatial information.

### B.1 Recommending objects from scene

Refer to Table 6 for examples. Upon inspecting generated samples, we observe that our model is often able to recommend appropriate objects that fall under the user’s criteria. The first example take place in a scene with jackets with the

color attributes mentioned by the system Generated  $A_T$ , demonstrating the ability to capture object attributes. The second example demonstrates the case where the system correctly recommend and ground jacket to the correct location.

However, it is not hard to find cases where the system is able to recommend the correct objects but in a wrong location. The third example demonstrates such case. All of the three recommended objects match those in the ground-truth response, but the system believes that they are all at a different location when in fact they are all on the left wall. We conjecture that our method of encoding object locations did not provide enough spatial information especially because we do not integrate the store structure itself. The retrieved  $A_T$  with

798 the same dialog yield the correct response since  
799 all negative samples in the candidate pool did not  
800 contain all of the three objects mentioned in the  
801 ground truth.

802 **B.2 Predicting coreference object and**  
803 **attributes**

804 Refer to Table 7 for examples. We see that the  
805 model successfully identifies which objects and  
806 slots are being queried. In most cases, the model  
807 outputs the exact corresponding object information  
808 without having to lookup the object metadata di-  
809 rectly. Furthermore, the model correctly identifies  
810 the turn for disambiguation. However, for more  
811 complicated instances such as the third example,  
812 the model mixes up the reference mentions and  
813 identifies the wrong value for the attribute. We  
814 also provide examples of all subtasks results (MM-  
815 Disamb, MM-Coref, MM-DST, Response Gener-  
816 ated Retrieval) with the corresponding VR scene  
817 in Figure 4, 5, 6, 7, and 8



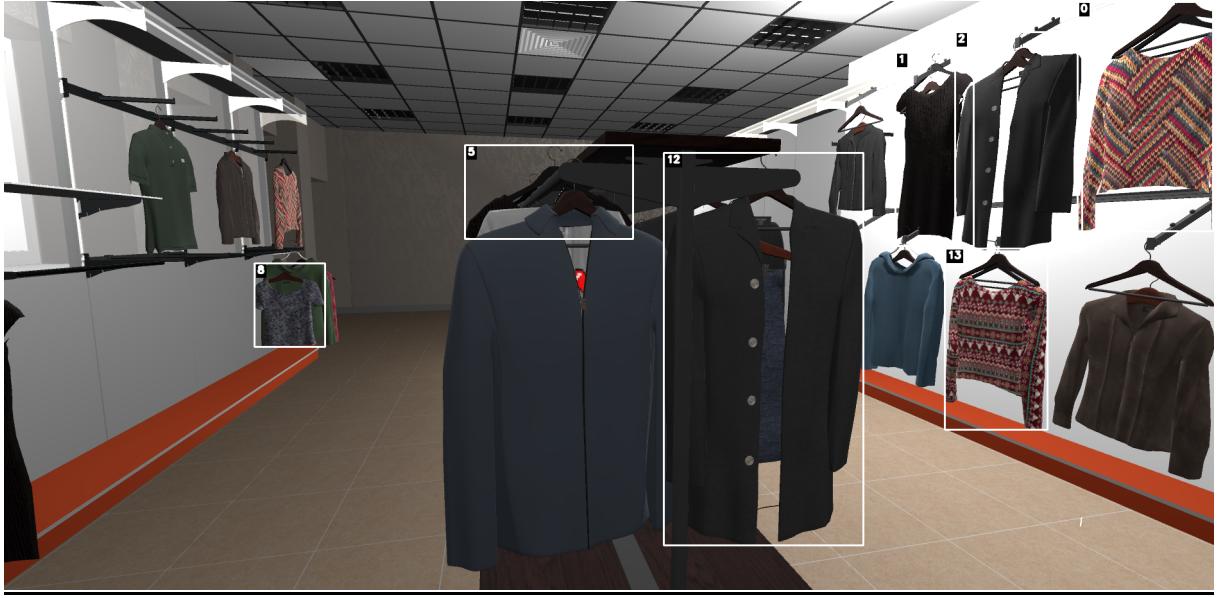
Utterance		MM-Disamb		MM-Coref	
		GT	Pred	GT	Pred
$U_0$	Are there any blouses that are cheap and in red and white that you can show me?	-	-	-	-
GT $B_0$	REQUEST:GET, slots: { type: blouse, price: cheap, color: red, white }				
Pred $B_0$	REQUEST:GET, slots: { type: blouse, price: cheap, color: red }				
GT $A_0$	There is one on the left rack. What do you think?				
Generated $A_0$	There is a red and white blouse which matches your search criteria.				
Retrieved $A_0$	Sorry but no matches were found. How else may I be of assistance?				
$U_1$	What are the available sizes and the brands of the red, white, and yellow blouse and the green and white blouse?	1	1	21, 10	21
GT $B_1$	ASK:GET, slots: { }				
Pred $B_1$	ASK:GET, slots: { }				
GT $A_1$	Which items?				
Generated $A_1$	Which items are you referring to?				
Retrieved $A_1$	Sorry, but no other matches were found.				
$U_2$	The green and white blouse in the closet and the red, white, and yellow blouse on the rack.	-	-	10, 21	10, 21
GT $B_2$	INFORM:DISAMBIGUATE, slots: { }				
Pred $B_2$	INFORM:DISAMBIGUATE, slots: { }				
GT $A_2$	The first is available in XS, S, and XL and is from The Vegan Baker. The other is in XS, S, and XL and is from Downtown Consignment.				
Generated $A_2$	Which items are you referring to?				
Retrieved $A_2$	The grey and white blouse is by Nature Photographers and is available in XS, M, L, and XL. The maroon and white blouse is by Global Voyager and is available in S, L, XL, and XXL.				
$U_3$	What else do you have that has good ratings?	-	-	-	-
GT $B_3$	INFORM:REFINE, slots: { customerReview: good, type: blouse, price: cheap, color: red, white }				
Pred $B_3$	INFORM:REFINE, slots: { customerReview: good, type: joggers }				
GT $A_T$	Sorry, but no other matches were found.				
Generated $A_3$	Sorry, but no item matches your search criteria.				
Retrieved $A_3$	Sorry, but no other matches were found.				
$U_4$	Now tell me how the green and white blouse compares with the other red and white blouse.	-	-	10, 27	10, 27
GT $B_4$	REQUEST:COMPARE, slots: { }				
Pred $B_4$	REQUEST:COMPARE, slots: { }				
GT $A_4$	The green blouse is 24.99. The red blouse is 24.99.				
Generated $A_4$	The green and white blouse has a light spots pattern and is available in XS, S, and XL and is in size XL. The other has a plaid pattern, is available only in XL, and has a red and white pattern.				
Retrieved $A_4$	The first one is from North Lodge and is a size L, the second one is from Modern Arts in an XXL.				

Figure 4: Comparison between the model prediction and the ground-truth of a SIMMC 2.0 dialog instance.



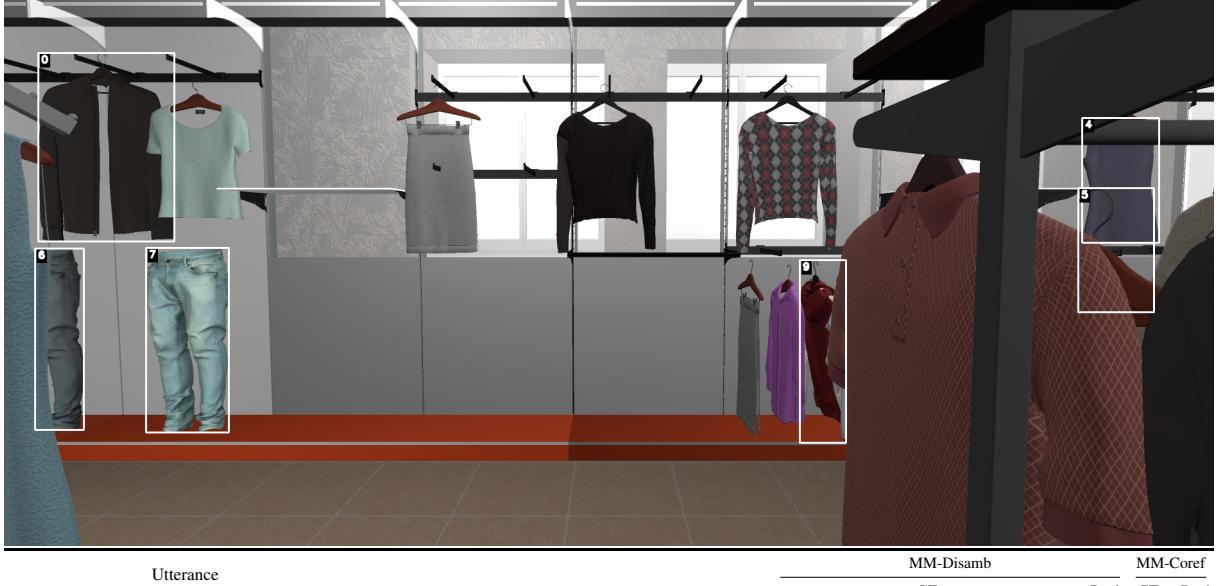
Utterance		MM-Disamb	MM-Coref		
		GT	Pred	GT	Pred
<i>U<sub>0</sub></i>	Can you help me find a cheap blouse?	-	-	-	-
GT <i>B<sub>0</sub></i>	REQUEST:GET, slots: {type: blouse, price: cheap }				
Pred <i>B<sub>0</sub></i>	REQUEST:GET, slots: {type: blouse, price: cheap }				
GT <i>A<sub>0</sub></i>	There are three options on the left wall: a grey and white, a white and grey, and a maroon, white, and blue.				
Generated <i>A<sub>0</sub></i>	What do you think of the grey and white blouse on the wall, the maroon, white, and blue on the rack, or the white and grey on the back wall?				
Retrieved <i>A<sub>0</sub></i>	There are three options on the left wall: a grey and white, a white and grey, and a maroon, white, and blue.				
<i>U<sub>1</sub></i>	Are any from Global Voyager and more expensive?	-	-	-	-
GT <i>B<sub>1</sub></i>	INFORM:REFINE, slots: { brand: Global Voyager, price: cheap, type: blouse }				
Pred <i>B<sub>1</sub></i>	INFORM:REFINE, slots: { brand: Global Voyager, price: cheap, type: blouse }				
GT <i>A<sub>1</sub></i>	None of the options match that description.				
Generated <i>A<sub>1</sub></i>	Sorry, but no.				
Retrieved <i>A<sub>1</sub></i>	None that match your search terms.				
<i>U<sub>2</sub></i>	What can you tell me about the white and grey blouse and the maroon, white, and blue?	0	0	7, 3	7, 3
GT <i>B<sub>2</sub></i>	REQUEST:COMPARE, slots: {}				
Pred <i>B<sub>2</sub></i>	REQUEST:COMPARE, slots: {}				
GT <i>A<sub>2</sub></i>	Both have long sleeves.				
Generated <i>A<sub>2</sub></i>	The white and grey is a hanging blouse with long sleeves. It is available in XS, S, M, and XL. It has a rating of 2.9 and is from StyleNow Feed. It costs 54.99 and is in size S. The maroon, white, and blue is a long-sleeved blouse from Global Voyager. It also has long sleeves and a 3.8 rating. It comes in S and L and costs 39.99.				
Retrieved <i>A<sub>2</sub></i>	Both have long sleeves.				
<i>U<sub>3</sub></i>	Will you add the maroon sweater from the display and the black and olive dress from the rack to my cart?	0	0	14, 30	28, 30
GT <i>B<sub>3</sub></i>	REQUEST:ADDTOCART, slots: {}				
GT <i>A<sub>3</sub></i>	No problem, both items have been added for you.				
Generated <i>A<sub>3</sub></i>	Of course. Both have been added.				
Retrieved <i>A<sub>3</sub></i>	No problem, both items have been added for you.				
<i>U<sub>4</sub></i>	Will you also add the black dress and the maroon one from the rack to my cart.	-	-	27, 20	20
GT <i>B<sub>4</sub></i>	REQUEST:ADDTOCART, slots: {}				
Pred <i>B<sub>4</sub></i>	REQUEST:ADDTOCART, slots: {}				
GT <i>A<sub>4</sub></i>	Sure, your cart has been updated.				
Generated <i>A<sub>4</sub></i>	Of course, both have been added to your cart.				
Retrieved <i>A<sub>4</sub></i>	Of course, both are being added to your cart now.				

Figure 5: Comparison between the model prediction and the ground-truth of a SIMMC 2.0 dialog instance.



Utterance	MM-Disamb		MM-Coref	
	GT	Pred	GT	Pred
<i>U<sub>0</sub></i>	I want a coat with good reviews.		-	-
GT <i>B<sub>0</sub></i>	REQUEST:GET, slots: {type: coat, customerReview: good}			
Pred <i>B<sub>0</sub></i>	REQUEST:GET, slots: {type: coat, customerReview: good}			
GT <i>A<sub>0</sub></i>	How is this black one in the top middle on the wall, or the other black one on the right side of the rack?			
Generated <i>A<sub>0</sub></i>	What do you think of the black coat on the rack?			
Retrieved <i>A<sub>0</sub></i>	How is this black one in the top middle on the wall, or the other black one on the right side of the rack?			
<i>U<sub>1</sub></i>	How does the red, white and yellow blouse on the top right compare to the white red and violet one on the bottom of the left wall?	-	-	0, 8 0, 8
GT <i>B<sub>1</sub></i>	REQUEST:COMPARE, slots: {}			
Pred <i>B<sub>1</sub></i>	REQUEST:COMPARE, slots: {}			
GT <i>A<sub>T</sub></i>	The first one is rated 3.4, and the other one is rated 4.7			
Generated <i>A<sub>1</sub></i>	The one on the right is rated 3.4, and the other one is rated higher at 4.7.			
Retrieved <i>A<sub>1</sub></i>	The left is light grey in M with short sleeves and the right is short sleeve in XXL with colors green, violet, and pick.			
<i>U<sub>2</sub></i>	Do you have anything similar to those black coats, but made by Downtown Consignment?	-	-	2, 12 2, 12
GT <i>B<sub>2</sub></i>	INFORM:GET, slots: {sleeveLength: full, brand: Downtown Consignment, size: L, price: 139.99 type: coat}			
Pred <i>B<sub>2</sub></i>	INFORM:GET, slots: {brand: Downtown Consignment, type: coat}			
GT <i>A<sub>2</sub></i>	I don't have anything like that.			
Generated <i>A<sub>2</sub></i>	I'm sorry, I don't have anything like that. Can I get you something different?			
Retrieved <i>A<sub>2</sub></i>	I don't have anything like that.			
<i>U<sub>3</sub></i>	What is the size on that dress?	1	1	1 -
GT <i>B<sub>3</sub></i>	ASK:GET, slots: {}			
Pred <i>B<sub>3</sub></i>	ASK:GET, slots: {}			
GT <i>A<sub>3</sub></i>	Which dress are you asking about?			
Generated <i>A<sub>3</sub></i>	Which dress?			
Retrieved <i>A<sub>3</sub></i>	Which dress are you asking about?			
<i>U<sub>4</sub></i>	The black one on the top middle of the wall display.	-	-	1 -
GT <i>B<sub>4</sub></i> .	INFORM:DISAMBIGUATE, slots: {}			
Pred <i>B<sub>4</sub></i> .	INFORM:DISAMBIGUATE, slots: {}			
GT <i>A<sub>4</sub></i>	That dress is a size M.			
Generated <i>A<sub>4</sub></i>	That dress is a size M.			
Retrieved <i>A<sub>4</sub></i>	This dress is available in S, M, L, XL, and XXL.			

Figure 6: Comparison between the model prediction and the ground-truth of a SIMMC 2.0 dialog instance.



Utterance	MM-Disamb	MM-Coref	
		GT	Pred
$U_0$ I'm looking to buy some luxury jeans that have reviewers falling head over heels.	-	-	-
GT $B_0$ REQUEST:GET, slots: {type: jeans, customerReview: good, price: expensive}	-	-	-
Pred $B_0$ REQUEST:GET, slots: {type: jeans, customerReview: good, price: expensive}	-	-	-
GT $A_0$ Those two jeans in the back to the left are a match! Do either of those suit your preferences?	-	-	-
Generated $A_0$ The blue and black pair on the left and the blue pair next to them are both in the expensive price range and have high ratings.	-	-	-
Retrieved $A_0$ Those two jeans in the back to the left are a match! Do either of those suit your preferences?	-	-	-
$U_1$ Yeah, I could get on board with those. Which one is more popular with buyers?	0	0	6, 7 6, 7
GT $B_1$ REQUEST:COMPARE, slots: {}	-	-	-
Pred $B_1$ REQUEST:COMPARE, slots: {}	-	-	-
GT $A_1$ The lighter pair is more highly rated, with a nearly unbeatable average customer rating of 4.9. The darker pair also comes highly recommended, with an average score of 4.6.	-	-	-
Generated $A_1$ Both of those jeans have an average customer rating of 4.6.	-	-	-
Retrieved $A_1$ The lighter pair is more highly rated, with a nearly unbeatable average customer rating of 4.9. The darker pair also comes highly recommended, with an average score of 4.6.	-	-	-
$U_2$ Good to know. Can I also get you to look for something in the same vein as that violet blouse hanging up to the right or the green, violet, and pink one near it, but by the brand Downtown Consignment?	-	-	4, 5 4, 5
GT $B_2$ INFORM:GET, slots: {brand: Downtown Consignment, type: blouse}	-	-	-
Pred $B_2$ INFORM:GET, slots: {brand: Downtown Consignment, type: blouse}	-	-	-
GT $A_2$ That brand makes the green and white shirt hanging above the light-wash jeans to the left.	-	-	-
Generated $A_2$ I'm so sorry, but there are no similar blouses by that brand.	-	-	-
Retrieved $A_2$ That brand makes the green and white shirt hanging above the light-wash jeans to the left.	-	-	-
$U_3$ Hm okay and what are the rating figures on that red jacket and the grey one?	1	1	9, 0 9, 0
GT $B_3$ ASK:GET, slots: {}	-	-	-
Pred $B_3$ ASK:GET, slots: {}	-	-	-
GT $A_3$ Woops, I'm not certain which jackets you're talking about. Could you please be more specific?	-	-	-
Generated $A_3$ Which ones are you referring to?	-	-	-
Retrieved $A_3$ Woops, I'm not certain which jackets you're talking about. Could you please be more specific?	-	-	-
$U_4$ I mean that red and white one to the right by the purple shirt and the grey one hanging near the green shirt to the left.	-	-	0, 9 0, 9
GT $B_4$ INFORM:DISAMBIGUATE, slots: {} Pred $B_4$	INFORM:DISAMBIGUATE, slots: {}	-	-
GT $A_4$ The red and white jacket is rated a 3.3, while the grey jacket is rated a 3.6.	-	-	-
Generated $A_4$ The red and white jacket has a customer rating of 3.3, and the grey jacket has an average customer rating for 3.6.	-	-	-
Retrieved $A_4$ The red and white jacket is rated a 3.3, while the grey jacket is rated a 3.6.	-	-	-

Figure 7: Comparison between the model prediction and the ground-truth of a SIMMC 2.0 dialog instance.



Models	Utterance	MM-Disamb		MM-Coref	
		GT	Pred	GT	Pred
$U_0$	Are there any highly rated pants around here?	-	-	-	-
GT $B_0$	REQUEST:GET, slots: { type: trousers, customerReview: good }				
Pred $B_0$	REQUEST:GET, slots: { type: trousers, customerReview: good }				
GT $A_0$	The first two pairs on the left have great reviews!				
Generated $A_0$	The blue pants on the left and the blue pants next to them are highly rated.				
Retrieved $A_0$	The first two pairs on the left have great reviews!				
$U_1$	Okay the last thing I need right now is more tshirts, but I can't help myself. I'm totally eyeing that white one in the back to the left and the black one hanging on the top row. Which of those is less expensive, and what sizes are those?	0	0	1, 11	1, 10, 11
GT $B_1$	REQUEST:COMPARE, slots: {}				
Pred $B_1$	REQUEST:COMPARE, slots: {}				
GT $A_1$	The black shirt pictured here is an extra large, and it sells for \$24.99. The white shirt pictured here is a large, and it is priced slightly higher, at \$29.99.				
Generated $A_1$	The white tshirt is less expensive, at \$29.99. It is a small, while the black tshirt costs \$24.99 and is an extra large.				
Retrieved $A_1$	The black shirt pictured here is an extra large, and it sells for \$24.99. The white shirt pictured here is a large, and it is priced slightly higher, at \$29.99.				
$U_2$	Gotcha. Got anything else for me to see?	-	-	-	-
GT $B_2$	INFORM:REFINE, slots: {type: trousers, customerReview: good}				
Pred $B_2$	INFORM:REFINE, slots: {type: tshirt}				
GT $A_2$	My apologies, but there are no more highly rated trousers for you to view at this time.				
Generated $A_2$	I'm sorry, but those are the only tshirts I have in stock at the moment.				
Retrieved $A_2$	Unfortunately, those are the only ones I have. Can I show you something different?				
$U_3$	Hm okay well can you try to dig up something with the same sizes as this black blouse on the first rack but with a polka dot pattern of some kind?	0	0	40	-
GT $B_3$	INFORM:GET, slots: {pattern: spots, availableSizes: [S, XS, M], type: blouse}				
Pred $B_3$	INFORM:GET, slots: {pattern: checkered, availableSizes: [XS], type: blouse}				
GT $A_3$	Oh no, it looks like there are no similar blouses in that pattern. What else can I do for you today?				
Generated $A_3$	I'm so sorry, but I can't find any similar items. Is there something else I might be able to help you with?				
Retrieved $A_3$	Oh no, it looks like there are no similar blouses in that pattern. What else can I do for you today?				
$U_4$	Oh snap that black blouse back there might be just the thing actually. What's that cost, and who makes it?	-	-	3	-
GT $B_4$	ASK:GET, slots: {}				
Pred $B_4$	ASK:GET, slots: {}				
GT $A_4$	Which blouse are you inquiring about?				
Generated $A_4$	Sorry, which black blouse are you interested in learning more about?				
Retrieved $A_4$	Which blouse are you inquiring about?				

Figure 8: Comparison between the model prediction and the ground-truth of a SIMMC 2.0 dialog instance.