SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations

Satwik Kottur[*], **Seungwhan Moon**[*], **Alborz Geramifard**, **Babak Damavandi**
Facebook Reality Labs & Facebook AI
✉ {skottur,shanemoon,alborzg,babakd}@fb.com

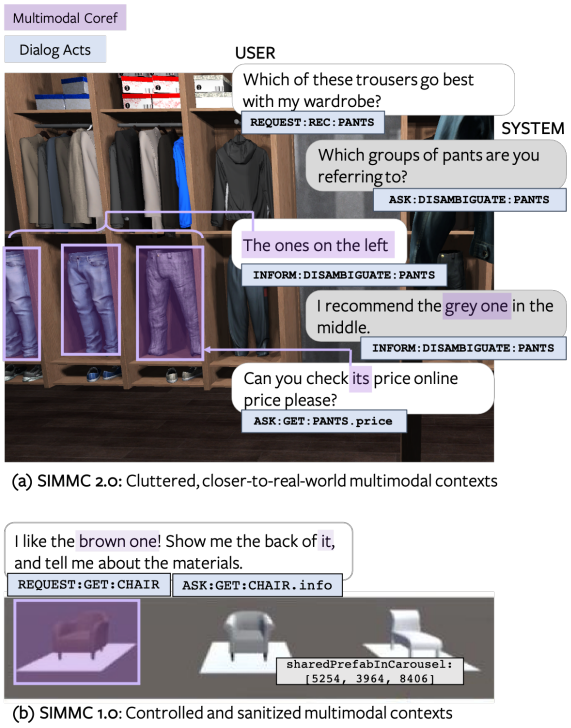
Abstract

Next generation task-oriented dialog systems need to understand conversational contexts with their perceived surroundings, to effectively help users in the real-world multimodal environment. Existing task-oriented dialog datasets aimed towards virtual assistance fall short and do not situate the dialog in the user's multimodal context. To overcome, we present a new dataset for Situated and Interactive Multimodal Conversations, SIMMC 2.0, which includes $11K$ task-oriented user↔assistant dialogs ($117K$ utterances) in the shopping domain, grounded in immersive and photo-realistic scenes.

The dialogs are collected using a two-phase pipeline: (1) A novel multimodal dialog simulator generates simulated dialog flows, with an emphasis on diversity and richness of interactions, (2) Manual paraphrasing of the generated utterances to collect diverse referring expressions. We provide an in-depth analysis of the collected dataset, and describe in detail the four main benchmark tasks we propose. Our baseline model, powered by the state-of-the-art language model, shows promising results, and highlights new challenges and directions for the community to study [1].

1 Introduction

The Situated and Interactive Multimodal Conversational AI (SIMMC) challenge (Moon et al., 2020), held as part of DSTC9 (Gunasekara et al., 2020), aimed to lay the foundations for the real-world assistant agents that can handle multimodal inputs, and perform multimodal actions. Specifically, it provided the SIMMC datasets as new benchmarks for studying task-oriented dialogs that encompass a situated multimodal user context in the form of a co-observed image or virtual reality (VR) environment. Since the introduction of the dataset, a number of follow-up work (Kung et al., 2021; Kim et al., 2021; Jeong et al., 2021; Huang et al., 2021; Senese et al., 2021) have established a new set of state-of-the-art baselines for the multimodal task-oriented dialog systems on SIMMC.



(a) SIMMC 2.0: Cluttered, closer-to-real-world multimodal contexts

(b) SIMMC 1.0: Controlled and sanitized multimodal contexts

Figure 1: Illustration of a Situated Interactive Multimodal Conversation (SIMMC), which presents a task-oriented user↔assistant dialog grounded in a co-observed multimodal context. The newly collected SIMMC 2.0 dataset includes complex and photorealistic multimodal contexts, which poses more challenges for the Multimodal Coreference Resolution task (MM-Coref) and the Multimodal Dialog State Tracking task.

Though SIMMC serves as a step towards building multimodal virtual agents, the dataset falls short (understandably so) in the complexity of the considered multimodal contexts. In particular, the co-observed image or VR environment is simplistic and far from realistic user situations. To bridge this gap, we take inspiration from the first SIMMC

[*] Joint first authors

[1] Code & data are made publicly available: https://github.com/facebookresearch/simmc2

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4903–4912
November 7–11, 2021.

challenge (Moon et al., 2020) and propose a new multimodal dialog dataset (SIMMC 2.0) for the community to tackle and continue the effort towards building a successful multimodal assistant agent. Specifically, SIMMC 2.0 is designed to include a closer-to-real-world context for a fashion or furniture shopping scenario, moving away from the sanitized contexts present in the original SIMMC datasets. To this end, we propose a VR scene generator that allows for controlling and capturing diverse multimodal contexts with ground-truth scene graph information, while serving as a close proxy for real-world scenarios. We then collect $11K$ assistant↔user task-oriented dialogs ($117K$ utterances) grounded on diverse photo-realistic VR renders of commercial stores ($1.5K$ different scenes).

The incorporation of the complex and cluttered multimodal contexts introduces several interesting challenges, such as understanding visual *and* dialog coreferences ('*the one directly behind it*', '*the one I mentioned*'), tracking dialog states along with multimodal objects, etc. In addition, the use of photo-realistic scenes surfaces practical limitations of CV models that need to be addressed, such as the detection of partially observed or obstructed referent objects, the visual texture recognition, *etc*.

To this end, we propose four main benchmark tasks that are essential in building a multimodal task assistant: Multimodal Disambiguation, Multimodal Coreference Resolution (MM-Coref), Multimodal Dialog State Tracking (MM-DST), and Response Generation. We then provide a baseline model trained for these tasks, and highlight the key challenges and future research directions.

2 Related Work

Problem Setup: The SIMMC 2.0 dataset addresses the conversational scenarios where the virtual assistant shares a co-observed scene with a user in addition to the traditional communication that takes place in the form of natural language. Specifically, we choose the shopping experience as the domain for this study, as it often induces rich multimodal interactions around browsing visually grounded items. We assume that the assistant agent has ground-truth meta information of every object in the scene, while users only observe those objects through the visual modality to describe and compose a request. In addition, we allow users to physically navigate within each scene, which we simulate as multiple viewpoints updated at different time steps throughout each dialog. Thus, models for SIMMC 2.0 would need to understand the user utterance using both the dialog history and the state of the environment as multimodal context.

Figure 2: Example snapshots from random camera viewpoints generated from a rearranged scene. Refer to Sec. 3.1.1 for more details.

Multimodal Dialog Datasets: Note that our problem setup for co-observing assistant scenarios allows for more natural multimodal coreferences to be used as part of user-assistant conversations. The existing literature in multimodal dialogs (Hori et al., 2018; Das et al., 2017; Kottur et al., 2019; de Vries et al., 2017, 2018) often posits the roles of a primary and secondary observer, *i.e. questioner* and *answerer* similar to the Visual Question Answering (Antol et al., 2015) tasks, hence showing a different distribution of language.

Task-oriented Dialog Systems: Many datasets have been developed in the past to support various assistant scenarios (*e.g.* booking hotels, reserving hotels) (Henderson et al., 2014; Rastogi et al., 2019; Budzianowski et al., 2018; Eric et al., 2019), defining many challenges in handling user requests under the unimodal dialog setting. Our setup extends many of these challenges studied in the previous literature on task-oriented dialog systems (*e.g.* DST, slot carryovers) to the unique multimodal settings.

The most recent thread in building a task-oriented dialog system is to fine-tune an end-to-end system on a large pre-trained causal language model, which achieves the state-of-the-art performance in many metrics (Hosseini-Asl et al., 2020; Peng et al., 2020; Chao and Lane, 2019; Gao et al., 2019; Crook et al., 2021). We follow this line of work and provide a baseline which extends it to accommodate for the multimodal input.

3 SIMMC 2.0 Dataset

SIMMC 2.0 assumes the scenario where a user is interacting with a conversational assistant to ob-

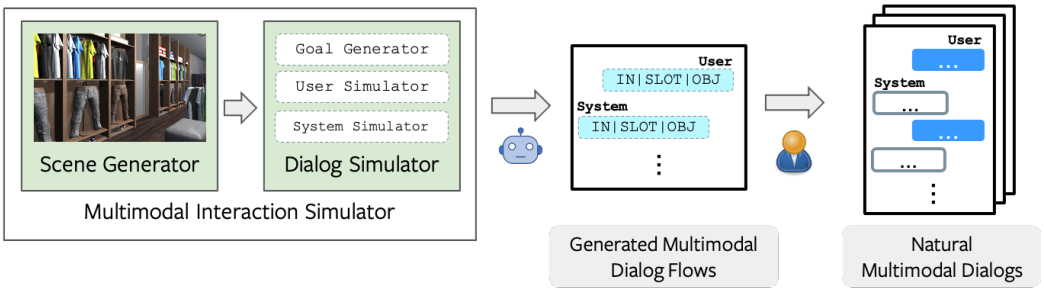


Figure 3: Illustration of the two-stage data collection pipeline. Phase 1: Simulated Multimodal Dialog Self-Play (Sec. 3.1) & Phase 2: Manual Human Paraphrase (Sec. 3.2)

tain recommendations for a piece of furniture or a clothing item. The dialogs were collected through a two-phase pipeline (Fig. 3), minimizing the annotation overheads (time and cost). This approach extends the popular machine↔human collaborative dialog collection approaches (Rastogi et al., 2019; Shah et al., 2018) to the multimodal settings.

3.1 Multimodal Dialog Self-Play

The first phase entails generating synthetic dialog flows between the user and assistant using a multimodal dialog simulator (Sec. 3.1.2). The simulator conditions the flow generation on various VR scenes snapshots (produced by scene simulator) for both fashion and furniture domains.

3.1.1 Scene Simulator

In our work, we use photo-realistic, virtual renderings of cluttered shopping environments for fashion and furniture domains, to replicate real-world settings. To this end, we develop a scene simulator to generate a diverse set of snapshots that serve as the multimodal context for the conversations.

Scene Generation. Re-arranging objects semantically in a 3D environment to create novel arrangements is a long standing research problem (Fisher et al., 2012, 2015). To avoid the challenges in a completely automatic approach, we design the following pipeline in a VR environment (Unity Technologies, 2019). We begin by manually constructing photo-realistic shopping scenarios as 'seed scenes', using publicly available digital assets like, (a) fashion: shirts, dresses, trousers, and shoes, (b) furniture: sofas, chairs, dining table, and lamps. We then programmatically re-arrange these assets at random for each of these seed scenes to create a larger pool of scenes (Tab. 1), while keeping the semantics of the scene intact. For example, a shirt in the seed scene is replaced only by another asset from either the same (shirt) or semantic related asset category (*e.g.*, T-shirt, jacket). This ensures that re-arranged scenes continue to be photo-realistic and avoids object collisions and hallucinations, trading off with fixed arrangement of semantic asset types within each seed scene.

Finally, we capture multiple views from random camera positions within each scene, as shown in Fig. 2. The height of the camera is mostly held constant with a small jitter, whereas the camera position (when projected onto the floor plane) is randomly chosen and is constrained to be within 75% of the floor bounds. These settings give us a good view of the scene objects without the risk of being either: (a) too close, such that the entire snapshot is taken by partially visible 1–2 objects resulting in a poor and uninteresting scene view, or, (b) too far away, where objects are small and hard to differentiate from one another. Randomly sampled camera viewpoints also encourage the diversity of referring expressions (*e.g.*, *'shirt closest to the changing rooms'*, *'cap at the farthest end of the table'*), useful for successful coreferences or disambiguation within the dialog utterances.

Annotation Extraction. The synthetic nature of our scenes facilitates an easy extraction of complete scene graph information for any given snapshot, without any additional human annotations. This is particularly beneficial as it enables the generation of rich and interesting dialog flows (Sec. 3.1.2), and allows for a tighter control over the distribution of objects and attributes within the conversation, which is nearly impossible with real world multimodal contexts. The annotations we extract for each scene snapshot consists of all the assets that appear in the snapshot, their image 2D bounding box, and an index to cross reference additional metadata from the catalog (*e.g.*, price, available sizes, color, pattern). After extracting these anno-

tations, we filter out snapshots with less than 5 objects in the field of view, and input the remaining scenes into the dialog simulator. See Sec. 3.3 for a detailed analysis of the generated scene snapshots and the underlying assets used in our work.

3.1.2 Multimodal Dialog Simulator

The multimodal dialog simulator takes generated scenes along with the meta information (objects, locations, and attributes) to create user↔assistant dialog flows, following an agenda-based dialog simulator approach (Schatzmann et al., 2007).

Multimodal Dialog Self-play. The dialog simulator consists of three main components: the *goal generator*, the *user simulator*, and the *assistant simulator*. The goal generator randomly selects an agenda for each dialog, which describes a high-level sequence of *goals* within shopping scenarios (*e.g.*, `BROWSE` → `GET_INFO` → `REFINE`; see Fig. 5). Given a goal, the user simulator draws a suitable dialog action following a probability distribution, which consists of natural language understanding (NLU) intents (*e.g.*, `REQUEST:GET`, `CONFIRM:ADD_TO_CART`), slots (*e.g.*, color, pattern), and object references. The assistant simulator then reads the user request, interacts with the multimodal contexts via the simulated API (*e.g.* for looking up the information of an item from the catalog, recommending items from the scene), and responds with natural language generation (NLG) intents, slots and object references. This dialog self-play repeats until each goal in the agenda is successfully met, or when the dialog reaches the maximum number of turns.

Multimodal Dialog Ontology. The dialog annotations for SIMMC 2.0 include the NLU and NLG intent and slot labels, following the conventional approaches for task-oriented dialog systems (Eric et al., 2019; Rastogi et al., 2019; Moon et al., 2020). Extending the dialog ontology for the complex multimodal settings, we also annotate the object references with their corresponding IDs as defined by the bounding boxes in each scene, allowing for a seamless annotation of multimodal contexts and language (*e.g. 'Do you have anything similar to the two middle jackets on the table?'* → `INFORM:GET_SIMILAR`, *slots: {type: jacket}, objects: [0,8]*). Note also that the same notation is employed to refer to object mentions that are carried over in the dialog context (*e.g. 'How much is the jacket I mentioned earlier?'* → `INFORM:GET.price`, *objects: [8]*). This fine-grained and unified ontology allows for the systematic study of the diverse referring expressions (*i.e.*, object mentions) in multimodal dialogs.

Total # dialogs	11,244
Total # utterances	117,236
Total # scene snapshots	1566
Avg # words per user turns	12
Avg # words per assistant turns	13.7
Avg # utterances per dialog	10.4
Avg # objects mentioned per dialog	4.7
Avg # objects in scene per dialog	19.7

Table 1: **SIMMC 2.0 Dataset Statistics**

3.2 Manual Paraphrase

The simulated dialog flows are then paraphrased by human annotators. This helps us draw utterances from the natural language distribution, as expected in a real world application. For this annotation effort, we designed a tool that displays a multimodal scene (generated VR scene screenshot) and a simulated dialog flow, and asked the human annotators to paraphrase the utterance ensuring that critical information such as objects and attributes is retained. An example dialog is shown in Appendix.

Advantages of the two-stage approach: Since paraphrasing synthetic utterances is much faster and less demanding, our approach requires reduced annotation effort. Further, the simulator in Phase 1 provides all annotations for the dialog state and coreferences for free, *i.e.*, without any additional human annotations.

3.3 SIMMC 2.0 Dataset Analysis

We analyze our dataset that contains a total of $11.2k$ dialogs (about $117k$ utterances), split into $7.2k$ and $4k$ dialogs from fashion and furniture domains respectively, along with the rich annotations from both scene simulator and multimodal dialog simulator that are extracted automatically without any additional human annotators. Tab. 1 shows the overall statistics of the dataset.

Analyzing Assets & Scene Snapshots. In our work, we use around 290 digital assets for fashion[2], and 110 assets for furniture[3], across several asset categories shown in Tab. 2. From these assets, we construct 7 seed scenes for fashion and 1 seed scene for furniture. We then rearrange assets within each seed scenes 20 times to result in a pool of

[2] https://www.turbosquid.com/
[3] https://www.wayfair.com/

Fashion	hat, tshirt, jacket, hoodie, sweater, shirt, suit, vest, coat, trousers, jeans, joggers, skirt, blouse, tank top, dress, shoes
Furniture	area rug, bed, chair, couch chair, dining table, coffee table, end table, lamp, shelves, sofa

Table 2: Digital Asset Categories used in SIMMC 2.0 for both fashion and furniture domains.

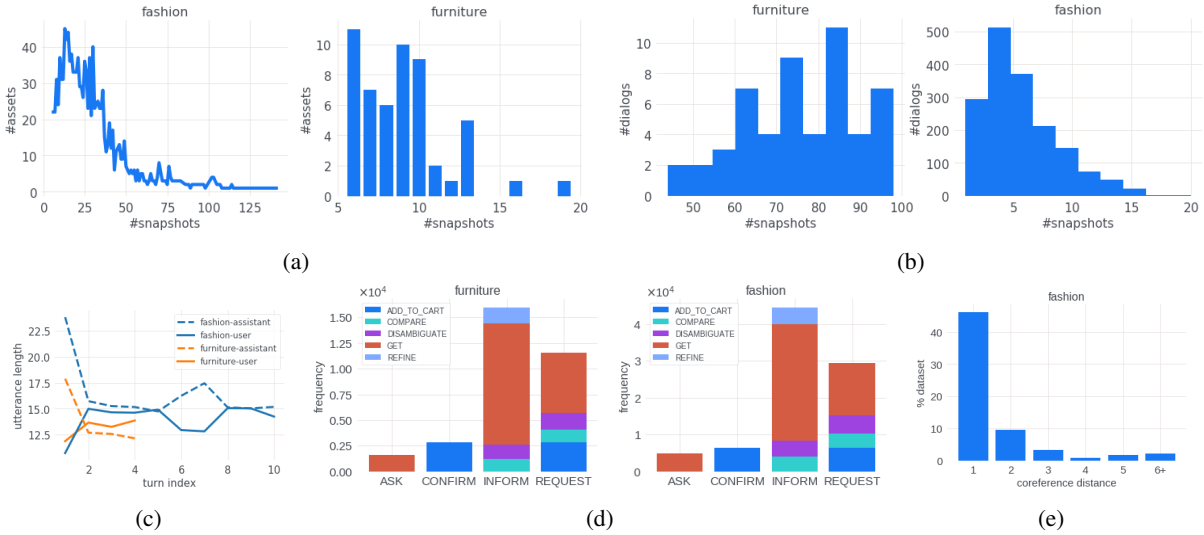


Figure 4: Distribution of (a) number of assets in each snapshot, (b) number of dialogs for each snapshot, (c) utterance lengths with dialog turns, (d) acts and activities, and (e) coreference distance between object mentions.

140 and 20 scenes for respective domains. Finally, we extract 10 snapshots from random camera viewpoints for each scene in the pool giving us a total of about 1566 unique scene snapshots to choose from post filtering. The number of objects in each snapshot is distributed as shown in Fig. 4a. This huge variance presents a good opportunity for the dialog simulator to ground the conversational flow in a varied number of objects.

Analyzing Dialog Annotations. The dialog simulator generates dialog flows by randomly sampling from the pool of 1566 filtered snapshots, which are later manually paraphrased. While most of the dialogs ($9.3k$) are grounded in a single snapshot, we include few dialog flows that spans over two snapshots ($1.9k$) with overlapping set of objects. This allows for modeling interesting conversations that require a context carry-over across the two viewpoints, thus moving closer to the real-world scenario. Each snapshot corresponds to about 7.1 dialogs on an average with the distribution shown in Fig. 4b. Further, each dialog contains around 5.2 utterance pairs (user↔assistant), where the utterances are 12.0 and 13.7 tokens long respectively (see Fig. 4c for distribution over different turns).

Following prior work (Moon et al., 2020), our dialog annotations also comprise dialog acts (4: `INFORM`, `CONFIRM`, `REQUEST`, `ASK`) and activities (5: `GET`, `DISAMBIGUATE`, `REFINE`, `ADD_TO_CART`, `COMPARE`). Fig. 4d shows their frequency breakdown. We also visualize the dialog act transitions for furniture in Fig. 5 for the first four rounds of the dialog. The presence of wide branch-offs and inter-connectivity suggests that our simulator is able to generate a diverse set of flows, useful to train a robust conversational system. It is interesting to note that the user utterances (marked with :U) almost always are more varied than assistant counterparts (marked with :A). This is probably due to `INFORM:GET` (fetching information) being a reasonable assistant response to a large number of user utterance queries. Fig. 4e shows the challenging nature of coreferences within our dialogs, where we measure the distance to the latest mention of an object, and requires models to reason across the utterances. Finally, Tab. 3 lists the various types of referring expressions in SIMMC 2.0, as implicitly controlled by the dialog simulator.

3.4 Comparison: SIMMC 2.0 vs SIMMC 1.0

The key differences between SIMMC 2.0 (ours) and SIMMC 1.0 (Moon et al., 2020) are (Fig. 1):

(a) The multimodal context in SIMMC 1.0 consists of either co-observed images or VR environment, which are simplistic and sanitized in comparison to real-world scenarios. For instance, the VR environment in SIMMC-Furniture comprises three slots (left, center, right) to populate the catalog items,

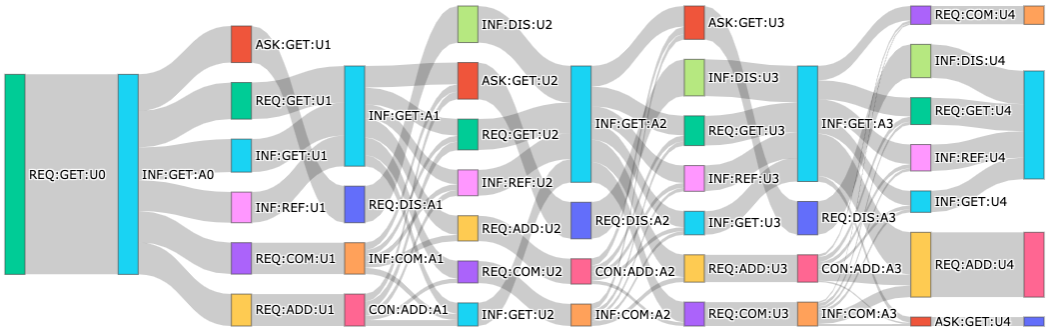


Figure 5: Dialog act transitions for the first four rounds of dialogs in the fashion domain. Label `ACT:ACTIVITY:[A|U][turn]` denotes the act and activity (shortened for brevity; see Fig. 3.3 for full names), either **A**ssistant or **U**ser utterance with turn index. The wide branch-offs and inter-connectivity demonstrates the diversity of dialog flows generated by our dialog simulator.

Referring Expression Type		**Examples**
Visual	Spatial (Absolute)	*'Have you got anything in the same size as the black top in the middle of the long rack?'* *'Add the white couch chair sitting on the red rug.'*
	Spatial (Relative)	*'Is there an armchair like the white one closer to us but from Modern Arts?'* *'I like the light grey jacket that is closer to us on that shelf, so please put it into my cart.*
	Adjectival	(color) *'Anything else like that blue sweater?'* (shape) *'There's the brown z-shaped end table you might enjoy.'* (multiple) *'I have that gray wooden table.'*
Textual	Noun Phrase (Paraphrase)	U: *'I'm kinda liking that brown one in the middle row*. Anything similar?'* → (*after 2 turns*) U: *'I mean the one on the left I told you I liked*.'*
	Noun Phrase (Copy)	A: *'Check out that brown table in the back right*.'* → (*after 2 turns*) U: *'OK. I've decided. Add the brown table*.'*
	Slot Carryover	*'Do you have anything similar at an affordable price in black and white?'*
	Pronoun	A: *What do you think of the black sweater on the right wall*?* → U: *'I like it*. Add it* to my cart.'*
Ambiguous	Dialog Context	*'What's the price?'*
	Visual Context	*'What's the price on the blue one?'*

Table 3: Referring expression types in SIMMC 2.0 with examples. (* for each row: referring the same object)

thus limiting the complexity of referential or disambiguation language. In contrast, conversations in SIMMC 2.0 are grounded in photo-realistic scene renderings of commercial stores that are cluttered and thus closely represent the real-world contexts.

(b) The number of objects in the multimodal context for SIMMC 1.0 is capped at 3 compared to 19.7 on average for SIMMC 2.0. This allows for richer coreferences, referential expressions, and disambiguation scenarios, elevating the role of dialog.

(c) Many of the objects in each scene are only partially observed (*e.g.*, blocked by different items or shelves, out of POV frame), which reflects real-world scenes but poses a more challenging problem for the computer vision module.

4 Task Formulation

The aim of the SIMMC 2.0 dataset is to emulate futuristic, real-world shopping scenarios where humans converse with a dialog agent in natural language grounded in a situated multimodal context. As a step towards this intelligent conversational agent, we leverage the dialogs and annotations in our dataset and propose four benchmark tasks (summarized in Tab. 4) along with evaluation metrics. These tasks capture several multimodal conversational reasoning challenges, as elaborated next.

4.1 Multimodal Disambiguation

In a real-world conversation, humans often use shorthands (coreferences) in order to refer to objects / events that have already been mentioned in the dialog. While we reserve modeling coreference resolution as a challenging task in Sec. 4.2, it is important for the system to recognize ambiguous uses of such coreferences even before attempting to resolve them. For example, consider *'A: The blue trousers are priced at $45. U: What about those?'*, where the phrase those could be ambiguous in the

Task Name	Goal	Evaluation
1. Multimodal Disambiguation	Given user utterances, classify if the assistant should disambiguate in the next turn.	Binary classification accuracy
2. Multimodal Coreference Resolution (MM-Coref)	Given user utterances with object mentions, resolve referent objects to their canonical ID(s) as defined by the catalog.	Coref Precision / Recall / F1
3. Multimodal Dialog State Tracking (MM-DST)	Given user utterances, track user belief states across multiple turns.	Intent Accuracy, Slot Precision / Recall / F1
4. Response Generation	Given user utterances, ground-truth APIs and ground-truth object IDs, generate Assistant responses or retrieve from a candidate pool.	Generation: BLEU; Retrieval: Accuracy@k, mean reciprocal rank, mean rank

Table 4: Proposed tasks and descriptions on our SIMMC 2.0 dataset. Please see Sec. 4 for more details.

following situations: (a) The user refers to a group of trousers without specifying the exact one they have in mind, (b) The user incorrectly uses a shorthand for a novel pair of trousers not mentioned in the dialog due to conversational brevity. In either cases, identifying the need for disambiguation and responding with *'Which ones are you talking about? The red or the green pair?'* is a desirable trait for a robust assistant system. The multimodal disambiguation task tests this ability of the agent.

More concretely, given the dialog history and the current user utterance, multimodal disambiguation requires the agent to predict a binary label conditioned on the multimodal context, to indicate the presence of a referential ambiguity in the user utterance. This label could also be useful for other downstream tasks like assistant response generation (Sec. 4.4) in order to continue the conversation in a meaningful way. We use accuracy to measure and compare model performances for this task.

4.2 Multimodal Coreference Resolution

For this task, we aim to resolve referential mentions in user utterances to their canonical object IDs as defined for each scene. These mentions can be resolved through (1) the dialog context (*e.g.* A: *'This shirt comes in XL and is $29.'* → U: *'Please add it to cart.'*, or (2) the multimodal context (*e.g.* U: *'How much is that red shirt?'*), or (3) both (*e.g.* U: *'How much is the one next to the one you mentioned?'*).

The input for this task includes the ground-truth bounding boxes defining each object ID, to avoid the performance bottleneck by the object detection algorithms. The main evaluation metric includes F1, precision and recall performance. Note that we exclude from evaluation the object mentions that are immediately followed by a disambiguation request (*e.g.*, *'How much is the one over there?'* ↔ *'Which one do you mean?'*, as they provide insufficient descriptions for resolving those coreferences.

4.3 Multimodal Dialog State Tracking

Following Moon et al. (2020), we extend the traditional notion of the unimodal dialog state tracking (DST) and propose multimodal dialog state tracking (MM-DST) as a main sub-task where slots are grounded on the coexisting multimodal context, which requires handling of multimodal objects (as opposed to textual tokens) as part of dialog states.

The performance is measured by the joint F1, recall and precision performance for the cumulative intent, slot and object reference predictions. The underlying reasoning behind this task is that the MM-DST labels will be able to provide sufficient information for a multimodal dialog system to carry out dialog policies and actions, given the detected and resolved items in each multimodal scene. Therefore, the MM-DST task measures the model's holistic understanding of user requests throughout each dialog, including the disambiguation needs as well as the coreferences.

4.4 Assistant Response Generation

The goal of this task is to generate assistant responses or retrieve from a candidate pool, given user utterances, ground-truth belief state, and object IDs. While we assume the assistant agent has the ground-truth meta information on each object, each response needs to naturally describe the referent objects *as observed and understood* by the user through the co-observed scene or the dialog context (*e.g.* `INFORM:RECOMMEND (OBJ_ID: 3)` → A: "*I recommend the blue shirt directly behind the brown jacket.*".

Similar to (Moon et al., 2020), we propose two ways to evaluate the performance of systems for response generation: (a) As a **generation** task, where

1. Disamb.	2. MM-Coref	3. DST		4. Gen.
Acc↑	Coref F1↑	Slot F1↑	Intent F1↑	BLEU↑
73.9±1.2	36.64±0.58	81.72±0.51	94.53±0.36	0.192±0.002
-	-	74.75±0.42	93.40±0.26	0.217±0.002

Table 5: Baseline performances: Moon et al. (2020) (top), Le et al. (2019) (bottom). **(1) Multimodal Disambiguation (Disamb.)**, via classification accuracy, **(2) Multimodal Coreference Resolution (MM-Coref)**, via coref prediction F1, **(3) Dialog State Tracking (DST)**, via slot and intent F1, **(4) Response Generation** via BLEU. ↑: higher is better.

the agent is seen as conditional language model. Performance is measured using BLEU-4 score (Papineni et al., 2002) between the generated response and the ground truth response provided with the dataset. (b) As a **retrieval** task, where the agent has to pick the ground truth response from a list of candidate responses (generated randomly; unique to each utterance). We use traditional information retrieval metrics like recall@k ($k = \{1, 5, 10\}$), mean rank, and mean reciprocal rank for comparing model performances.

5 Modeling & Empirical Analysis

In this section, we perform preliminary empirical analysis and train baselines. We leave more detailed modeling work for the future.

Dataset split. We randomly split the dataset into 4 sets: train (65%), dev (5%), dev-test (15%), and test-std (15%), which we leave as a held-out hidden set for performing a fair comparison of models.

Notations. We denote a SIMMC dialog with N_r rounds: $\mathcal{D} = \{(U_i, A_i, M_i, B_i)\}_{i=1}^{N_r}$, where U_i and A_i are the user and assistant utterances, M_i is the domain-specific multimodal context, and B_i is a multimodal belief state represented as a semantic parse of user-side dialog (*i.e.* intent, slot, object references, disambiguation labels), respectively. At each round t, given the current user utterance U_t, the dialog history $H_t = (U_i, A_i)_{i=1}^{t-1}$, and the multimodal context M_t, the task is to predict the user belief state B_t, as well as the natural language assistant response A_t.

Baselines. We benchmark the dataset by adopting: (a) **MM-DST model** by Moon et al. (2020), where we train a multi-task GPT-2 (Radford et al., 2019) based Transformer model using the joint supervision signals for the Disambiguation, MM-Coref, DST, and Response Generation tasks. Specifically, the model takes as input the dialog context and the

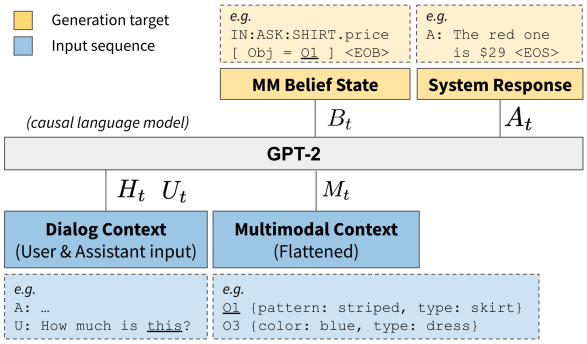


Figure 6: Illustration of the GPT-2 based baseline, which takes as input the dialog context and the flattened multimodal context, and outputs the belief states as well as the system response.

flattened multimodal contexts (as structurally formatted strings) to predict the belief states and the responses, following the popular causal language model approach (Peng et al., 2020; Hosseini-Asl et al., 2020). We use the 12-layer GPT-2 (117M parameters) as the pre-trained language model and fine-tune for ten epochs. Note that this baseline uses the ground-truth multimodal contexts provided from the scene generator, instead of consuming raw images as input, and thus serves as a soft oracle on the proposed dataset. (b) **Multimodal Transformer Network (MTN)** (Le et al., 2019) for the DST and Response Generation tasks. In particular, MTN uses image features extracted from scene snapshots and attends to relevant parts as guided by the dialog. We use the same training setting and hyperparameters as Le et al. (2019).

Analysis. The results are summarized in Tab. 5. Note that the F1 performance on the multimodal object coreference resolution task on SIMMC 2.0 is only at 36.6%, whereas the best model on SIMMC 1.0 (Moon et al., 2020) achieved 85.9% on the similar task. This demonstrates that SIMMC 2.0 presents more complex and cluttered scenes, thus requires more rigorous visual grounding of multimodal contexts (19.7 objects per dialog on average).

Conclusions. We present a novel dataset for the Situated and Interactive Multimodal Conversations, SIMMC 2.0, with $11K$ user↔assistant dialogs ($117K$ utterances) on shopping domain (fashion and furniture), grounded in situated and photo-realistic VR scenes. We then present a novel multimodal dialog simulator, which generates simulated dialogs grounded on diverse multimodal contexts that are automatically configured. Our empirical analysis with a baseline model demonstrates many new challenges that our SIMMC 2.0 dataset brings, highlighting new directions of research in this area.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Paul A. Crook, Satwik Kottur, Seungwhan Moon, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2021. Situated interactive multimodal conversations (simmc) track at dstc9. *AAAI DSTC9 Workshop*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. 2012. Example-based synthesis of 3d object arrangements. In *ACM SIGGRAPH Asia 2012 papers*, SIGGRAPH Asia '12.

Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. 2015. Activity-centric scene synthesis for functional 3d scene modeling. *ACM Transactions on Graphics (TOG)*, 34(6).

Shuyang Gao, Sanchit Agarwal Abhishek Seth and, Tagyoung Chun, and Dilek Hakkani-Ture. 2019. Dialog state tracking: A neural reading comprehension approach. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Chiori Hori, Anoop Cherian, Tim K. Marks, and Florian Metze. 2018. Audio visual scene-aware dialog track in dstc8. *DSTC Track Proposal*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

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. *AAAI 2021 DSTC9 Workshop*.

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

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. *AAAI 2021 DSTC9 Workshop*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

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. *AAAI 2021 DSTC9 Workshop*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623.

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. *arXiv preprint arXiv:2006.01460*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

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

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Unity Technologies. 2019. Unity. https://unity.com/.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.