Overview of Situated and Interactive Multimodal Conversations (SIMMC) 2.0 Track at DSTC 10

Satwik Kottur[1]*, **Seungwhan Moon** [2]*, **Alborz Geramifard** [1], **Babak Damavandi** [2]

[1] Meta AI
[2] Meta Reality Labs
{skottur,shanemoon,alborzg,babakd}@fb.com

Abstract

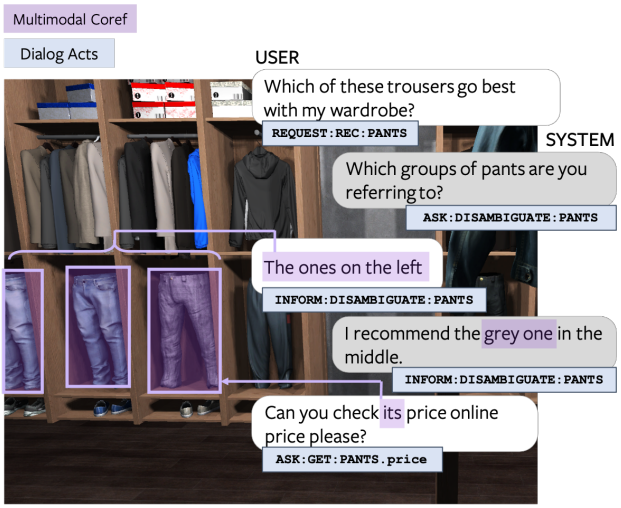
With ever increasing interest in task-oriented dialog systems, the recent work on Situated and Interactive Multimodal Conversations (SIMMC 2.0) aims to develop personal assistants that interact with users, grounded in an immersive and co-observed setting of photo-realistic scenes. The dataset contains $11k$ task-oriented dialogs set in an interactive shopping scenario, spanning more than $117k$ utterances.

To further enable research in this direction, the SIMMC 2.0 challenge[1] was held at the Tenth Dialog System Technology Challenge (DSTC) that saw entries from across the world competing to achieve the state-of-the-art performance in the SIMMC 2.0 task. In this paper, we describe and compare 10 SIMMC 2.0 models to better understand and summarize the current lay of the land for multimodal task-oriented dialog systems. We hope that our analysis throws light on components that showed promise in addition to identifying the gaps for future research towards this grand goal of an immersive multimodal conversational agent.

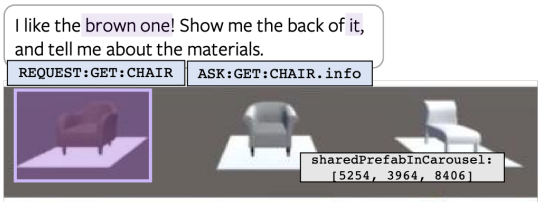
Introduction

Virtual assistants have a massive potential to positively impact and augment user's everyday life. To expand their scope of application, such assistants need to be multimodal and support users' queries grounded in their surroundings. Situated and Interactive Multimodal Conversations (SIMMC) (Moon et al. 2020; Kottur et al. 2021b) are a step towards this end, where the conversational agent is expected to model a multimodal environment (virtual or images), in addition to reasoning over the dialog history and process queries issues by the user.

To encourage research, the Situated and Interactive Multimodal Conversations (SIMMC) challenge (Moon et al. 2020; Kottur et al. 2021a) was held as part of DSTC9 (Gunasekara et al. 2020) in 2019. The first edition of the competition saw participation from over 5 teams (13 model entries) across the world establishing state-of-the-art performance on SIMMC dataset (Kung et al. 2021; Kim et al. 2021; Jeong et al. 2021; Huang et al. 2021; Senese et al. 2021). Propelled by the success of the SIMMC challenge, a second version of the challenge has been organized grounded in an improved SIMMC 2.0 dataset. The newer dataset serves as a new benchmark for studying conversations grounded in an immersive and co-observed virtual environment (through viewpoint screenshots), closer-to-real world context for a fashion or furniture shopping scenario, moving away from the sanitized contexts present in the original SIMMC dataset. See Fig. 1 for an exampe dialog from SIMMC 2.0 dataset, comparing it to the simple and sanitized multimodal contexts present in the original SIMMC dataset.



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



(b) SIMMC 1.0: Controlled and sanitized multimodal contexts

Figure 1: Example from the Situated Interactive Multimodal Conversation (SIMMC)(Kottur et al. 2021b) that shows a task-oriented user↔assistant dialog grounded in a co-observed multimodal context. Figure: Kottur et al. (2021b).

* Joint first authors

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

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 1: Proposed tasks and descriptions on SIMMC 2.0 dataset. Please see text for more details. Table: Kottur et al. (2021b)

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 2: **SIMMC 2.0 Dataset Statistics**

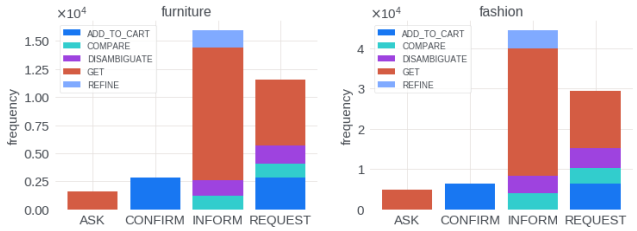


Figure 2: Distribution of dialog acts and activities for SIMMC 2.0. Figure from (Kottur et al. 2021b).

SIMMC 2.0 Challenge Details

Datasets

For this challenge, we use the SIMMC 2.0 dataset (Kottur et al. 2021b). The dataset contains $11.2k$ dialogs totalling $117k+$ utterances, grounded in $1.6k$ scenes. This is then split into 4 sets: `train` (65%), `dev` (5%), `dev-test` (15%), and `test-std` (15%). Annotations have been publicly released for the first three splits, while those for the last split are hidden and used for challenge purposes to compare performances of different entries.

Annotations

Due to its synthetic nature, the SIMMC 2.0 dataset has a wide-range of annotations including dialog and utterance-level acts, multimodal coreferences, disambiguation, and dialog state tracking. These annotations are used as golden standard for several tasks in the SIMMC 2.0 challenge.

Data and Annotation Analysis.

The dataset contains 4 dialog acts (`INFORM`, `CONFIRM`, `REQUEST`, `ASK`) and 5 activities (`GET`, `DISAMBIGUATE`, `REFINE`, `ADD_TO_CART`, `COMPARE`), whose distributions are given in Fig. 2. Tab. 3 contains examples illustrating these dialog acts and activities. Please see (Kottur et al. 2021b) for further analysis of the data along with the available annotations.

Tasks and Evaluation

SIMMC 2.0 proposes four different tasks to measure and benchmark performance of a task-oriented dialog agent that can interest with users in an immersive and situated environments. We briefly describe these tasks, see Tab. 1 for a summary.

Dialog Act	Activity	Example
`ASK`	`GET`	*U: I'd like to know the brand and customer rating of that please.*
`CONFIRM`	`ADD_TO_CART`	*A: Got it! You will have them in your cart in a moment.*
`INFORM`	`GET`	*A: It's made of leather and has a rating of 3.1.*
	`DISAMBIGUATION`	*U: I'm talking about the brown chair in the back and the black chair just behind the divider.*
	`REFINE`	*U: How about anything in size M?*
	`COMPARE`	*A: Of course! The black jacket is shown in a large size, and the grey and white one in XS.*
`REQUEST`	`GET`	*U: Do you have a nice black dress here?*
	`DISAMBIGUATION`	*A: Sorry, which one would you like to know about?*
	`ADD_TO_CART`	*U: Let's put the pink one in my cart along with the black one up above it.*
	`COMPARE`	*U: What's the difference in ratings on the light grey coat on the left up front and the black on the right wall?*

Table 3: Examples for valid combinations of dialog acts and activities in SIMMC 2.0. Please see (Moon et al. 2020) for definitions.

***(Subtask 1)* Multimodal Disambiguation** This task focuses on identifying whether a given user turn contains ambiguity in referencing to objects in the scene. Based on this, the assistant can trigger a disambiguation request to elicit further details about the object of user's choice. As defined in (Kottur et al. 2021b), 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. We use accuracy to measure and compare model performances for this task.

***(Subtask 2)* Multimodal Coreference Resolution** As the name suggests, this task requires the dialog system 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.

***(Subtask 3)* Multimodal Dialog State Tracking** Kottur et al. (2021b) 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.

***(Subtask 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.*".

There are two ways to evaluate the performance of systems for response generation: (a) 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. (b) As a **generation** task, where 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.

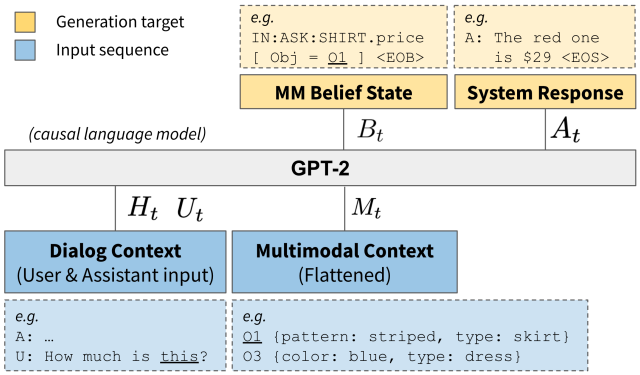


Figure 3: Illustration of the GPT-2 based baseline, which takes as input the dialog context and the flattened multimodal context, and outputs the belief states as well as the system response.

Baselines

There are two baselines, adopted from (Kottur et al. 2021b): (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 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).

Survey of Submitted Systems

We now provide brief description of the entries submitted to the SIMMC 2.0 challenge. A summary is provided in Tab. 4.

Team 1 uses a larger GPT-2 model with a similar architecture to the baseline, trained using multi-task learning on all the subtasks. Further, beamsearch of size 2 and 3 are used to improve generation capabilities. Finally, retrieval is performed using cosine similarity between the context query vector and candidate encoding.

Team 2 did not release their implementation, thus details about their approach are not public.

Team 3 uses a BART-based encoder-decoder framework. Specifically, the multimodal context of scene context descriptions (name ID, bounding box & position of objects, etc.) are encoded through the Bi-encoder & the Poly-encoder. The model is then trained to output all of the dis-

Team	Models	Joint (Pre)Train subtasks	Ens.	Language Model	MM Rep.	Subtask Ranks 1	2	3	4a	4b
Team 1	GPT-2 + Beam Search	2, 3, 4a, 4b	no	GPT-2 (large)	stringified	·	9	4	3	3
	GPT-2 + Beam Search	2, 3, 4a, 4b	no	GPT-2 (large)	stringified	·	9	4	4	3
Team 2	did not open-source	·	·	·	·	·	(1)	·	·	·
Team 3	BART+Poly-Encoder	1, 2, 3, 4b	no	BART	stringified	4	11	4	2	4
Team 4	BART	1, 2, 3, 4b	no	BART	object token	2	1	2	1	2
Team 5	BERT + ELECTRA	2, 3, 4a, 4b	no	BERT	stringified	2	8	3	5	1
	BERT + ELECTRA	2, 3, 4a, 4b	yes	BERT	stringified	·	·	·	·	1
Team 6	BART+ResNet (Ensemble)	2, 3, 4b	yes	BART	stringified (ResNet)	1	6	1	·	(2)
	BART+ResNet	2, 3, 4b	no	BART	stringified (ResNet)	1	·	·	·	·
Team 7	TOD-BERT	·	no	TOD-BERT	-	3	·	·	·	·
	LXMERT	·	no	LXMERT	LXMERT	·	7	·	·	·
Team 8	1: (UC+SM+MVM)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	3	5	5	·	2
	2: (SM)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	5	·	·	·
	3: (UC+SM+MVM w/o MI)	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	4	·	·	·
	1+3	1, 2, 3, 4b	no	RoBERTa/GPT-2	DeIT	·	3	·	·	·
Team 9	UNITER + Scene Graph	2b	no	UNITER	UNITER	·	2	·	·	·
Team 10	GLIMMeR	1, 2, 3, 4b	yes	GPT-2	stringified	·	10	·	·	·
	GLIMMeR (Ensemble)	1, 2, 3, 4b	yes	GPT-2	stringified	2	3	4	·	1

Table 4: **Summary of the developed models** for Subtasks 1, 2, 3, 4a and 4b.

ambiguation label, belief state and the assistant response in a single string.

Team 4 proposes a universal transformer model that utilizes the dialog context as well as the multimodal context (object knowledge base and visual scenes) to determine whether an object is mentioned in the current utterance. Specifically, object token embeddings are represented by taking as input the object index, object locations, etc. The model is then fine-tuned for each of the task.

Team 5 uses ELECTRA finetuning approach (Clark et al. 2020), which uses discriminative signals instead of generative signals on BERT architecture. Specifically, the model is finetuned by discriminating between a 'real' and a 'fake' input token.

Team 6 proposes a RoBERTa model for sub-task1, and a BART model approach for sub-tasks 1 through 4 with the ResNet fine-tuned to predict visual metadata. The predicted visual metadata is then converted into a sequence of word tokens representing the multimodal context. The dialog history and the multimodal context are then given as input to the causal language model (BART) that sequentially predicts the labels for subtasks 2, 3 and 4.

Team 7 proposes a TOD-BERT model for sub-task 1 fine-tuned for the binary classification, and a LXMERT-based model for sub-task 2. For each of the objects, visual feature is extracted by combining bounding boxes, RoI features and object positional embeddings through several linear and normalisation layers. These visual features are then combined with the BERT sentence embeddings. The model then is trained to predict the likelihood of a sentence referencing a particular object.

Team 8 proposes a multimodal model composed of a RoBERTa model that encodes the textual representation of a multimodal object, and a DeIT model (Touvron et al. 2012) that encodes the visual representation. The multimodal model is first pre-trained to learn the relationship between the two modalities, with object candidates pooled from 'mention_inform' (MI). Total three main objectives were proposed for loss propagation: utterance classification (UC), system matching (SM), and meta-visual matching (MVM). The model is then fine-tuned for each sub-task.

Team 9 proposes a UNITER-based model (Chen et al. 2020) that utilizes the dialog context as well as the multimodal context (object knowledge base and visual scenes) to determine whether an object is mentioned in the current utterance. Specifically, object token embeddings are represented by taking as input the object index, image pixels, KB entities, and the indicators of whether an object has been mentioned before.

Team 10 proposes a global-local information-aware multimodal model based on the GPT-2 model. Specifically, they provide an efficient GPT-2 implementation which skips the default byte-pair encoding for specially introduced tokens and instead encodes new object tokens with a single identifier, ensuring the grounded representation of those tokens within the model.

Performance Analysis

Starting with a description of the baseline, we will now present the results for the SIMMC 2.0 challenge entries.

Performance Summary

The entries to the challenge set a new state-of-the-art in all four subtasks. The results are summarized in Tab. 5.

The winner of the multimodal disambiguation subtask (subtask 1) was the BART+ResNet model from Team 6. This

Team	1. Disamb.	2. MM-Coref	3. MM-DST		4. Response Retrieval & Generation					
	Acc↑	Coref F1↑	Slot F1↑	Intent F1↑	MRR↑	r@1↑	r@5↑	r@10↑	Mean↓	BLEU↑
GPT-2	73.5	44.1	83.8	94.1	·	·	·	·	·	0.202
MTN	·	·	76.7	92.8	·	·	·	·	·	0.211
Team 1	·	52.1	88.3	96.3	53.5	42.8	65.4	74.9	11.0	0.285
	·	51.9	88.4	96.3	51.7	41.2	62.8	72.5	11.9	0.279
Team 2	·	78.3	·	·	·	·	·	·	·	·
Team 3	89.5	42.2	87.8	96.2	61.2	49.6	74.7	84.5	6.6	0.256
Team 4	93.9	75.8	90.3	95.9	81.5	71.2	95.0	98.2	1.9	0.295
Team 5	93.8	56.4	89.3	96.4	32.0	19.9	41.8	61.2	12.9	0.322
Team 6	94.7	59.5	91.5	96.0	·	·	·	·	·	·
	94.5	·	·	·	·	·	·	·	·	0.309
Team 7	93.1	57.3	·	·	·	·	·	·	·	·
Team 8	93.1	63.4	4.0	41.4	·	·	·	·	·	0.297
	·	63.0	·	·	·	·	·	·	·	·
	·	66.7	·	·	·	·	·	·	·	·
	·	68.2	·	·	·	·	·	·	·	·
Team 9	·	73.3	·	·	·	·	·	·	·	·
Team 10	·	50.6	·	·	·	·	·	·	·	·
	93.6	68.2	87.7	95.8	·	·	·	·	·	0.327

Table 5: Summary of the results on Test-Std split. Best results from each system are shown. **(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, recall@k (k=1,5,10), Mean rank, and mean reciprocal rank (MRR). ↑: higher is better. Baseline performances: Moon et al. (2020) (top), Le et al. (2019) (bottom).

model was the winner for the dialog state tracking subtask (subtask 3) as well. The winner of the multimodal coreference resolution task (subtask 2) and the response retrieval task (subtask 4a) was a BART-based multimodal model from Team 4. The joint winners of the response generation (4b) were Team 5 and 10.

Per-Subtask Analysis

Subtask 1: Fig. 4 shows the distribution of the disambiguation accuracy as turns of the dialog progress.

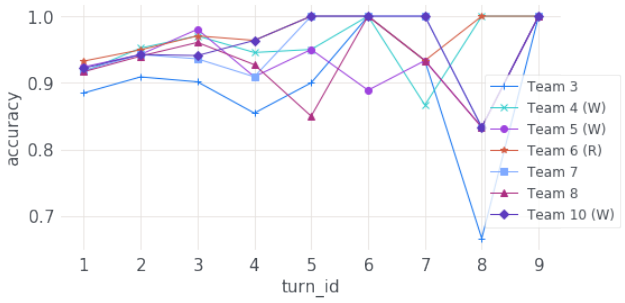


Figure 4: Distribution of disambiguation accuracy as dialog progresses.

Subtask 2 & 3: From Table 5, we find that the universal multimodal transformer approaches that incorporate object token embeddings as part of causal language modeling (encoding information such as object coordinates, visual features, etc.) tend to be the most effective for the MM-Coref task (Team 4, 8 and 9). Interestingly, for the Dialog State Tracking task, we find that the approaches that leverage the stringified tokens as multimodal context (Team 6) perform the best, showing that having more homogeneous token representations is beneficial for the primarily language-focused tasks such as dialog act slot prediction.

Subtask 4: To better understand the performance of the models, we analyze BLEU scores based on the dialog act. An interesting trend is that both the winner entries outperform the rest in `INFORM:COMPARE` act. This is doubly challenging as it involves informing the comparison of two items.

Conclusions

The SIMMC 2.0 challenge saw an increase the diversity of the underlying architectures of the transformers used to fine-tune on the in-domain data. This could be an indication of different backbones offering complementary benefits to fine-tuning on SIMMC 2.0 dataset.

Our aim in organizing the SIMMC 2.0 challenge in DSTC10 is to encourage and motive the research community towards the problem of situated and interactive multimodal task-oriented dialog agents. Such agents have immense practical applications and with them bring a plethora of multimodal challenges. We hope that the insights gained throws light on the challenges of such multimodal dialogs

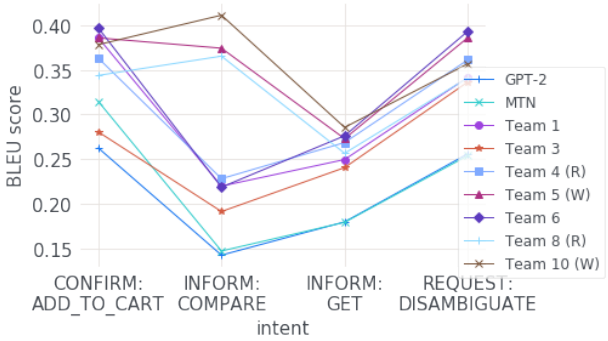


Figure 5: Distribution of BLEU score based on the natural language generation act.

and inspires multiple follow-up lines of research.

What next? We identify potential next steps for the SIMMC 2.0 challenge to further this direction of research (decreasing magnitude of change compared to SIMMC 2.0):

- **Multimodal input streams.** The current setup leverages screenshots from a shopping store as multimodal context. While this is a hard setup in itself with several challenges, it does not capture additional inputs like eye gaze, head position, gestures *etc*., expected in a real setting. Human users often uses these additional cues for referring to objects (*e.g*., *"How much is that shirt (pointing a finger)"*). In order to incorporate such modalities, SIMMC would need to ground conversations on a virtual environment with a stream of inputs to capture users' eye gaze, head position, location in the store, *etc*.
- **Robustness to size, multiple domain, and catalog updates.** Shopping applications often require the assistant to be robust towards the size of the catalog, spanning multiple domains, and updating catalogs. Re-training conversational models for every change in the catalog is not feasible and therefore stricter evaluation must be imposed to discourage catalog memorization.
- **Challenging disambiguation task.** Though multimodal disambiguation occurs less frequently (1 out of 14 turns in SIMMC 2.0), it is an important skill in a conversational agent. In light of a nearly perfect accuracy (95%), a potential future direction is to ensure harder disambiguation turns in SIMMC 2.0 either through synthetic data augmentation or additional data collection.

References

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Gunasekara, C.; Kim, S.; D'Haro, L. F.; Rastogi, A.; Chen, Y.-N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Hosseini-Asl, E.; McCann, B.; Wu, C.-S.; Yavuz, S.; and Socher, R. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Huang, X.; Tan, C. S.; Ng, Y. B.; Shi, W.; Yeo, K. H.; Jiang, R.; and Kim, J. J. 2021. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations. *AAAI 2021 DSTC9 Workshop*.

Jeong, Y.; Lee, S. J.; Ko, Y.; and Seo, J. 2021. TOM : End-to-End Task-Oriented Multimodal Dialog System with GPT-2. *AAAI 2021 DSTC9 Workshop*.

Kim, B.; Lee, I.; Jeong, Y.; Youngjoong, K.; Koo, M.-W.; and Seo, J. 2021. Improving Multimodal API Prediction via Adding Dialog State and Various Multimodal Gates. *AAAI 2021 DSTC9 Workshop*.

Kottur, S.; Crook, P.; Moon, S.; Beirami, A.; Cho, E.; Subba, R.; and Geramifard, A. 2021a. An Analysis of State-of-the-Art Models for Situated Interactive MultiModal Conversations (SIMMC). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 144–153. Singapore and Online: Association for Computational Linguistics.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021b. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4903–4912. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kung, P.-N.; Yang, T.-H.; Chang, C.-C.; Hsu, H.-K.; Liou, Y.-J.; and Chen, Y.-N. 2021. Multi-Task Learning for Situated Multi-Domain End-to-End Dialogue Systems. *AAAI 2021 DSTC9 Workshop*.

Le, H.; Sahoo, D.; Chen, N.; and Hoi, S. 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*, 5612–5623.

Moon, S.; Kottur, S.; Crook, P. A.; De, A.; Poddar, S.; Levin, T.; Whitney, D.; Difranco, D.; Beirami, A.; Cho, E.; Subba, R.; and Geramifard, A. 2020. Situated and Interactive Multimodal Conversations. *arXiv preprint arXiv:2006.01460*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2020. SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model. *arXiv preprint arXiv:2005.05298*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Senese, M. A.; Rizzo, G.; Benincasa, A.; and Caputo, B. 2021. A Response Retrieval Approach for Dialogue Using a Multi-Attentive Transformer. *AAAI 2021 DSTC9 Workshop*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2012. Training data-efficient image transformers & distillation through attention (2021).