Situated Interactive Multimodal Conversations (SIMMC) Track at DSTC9

Paul A. Crook[*], Satwik Kottur[*], Seungwhan Moon[*], Ahmad Beirami, Eunjoon Cho, Rajen Subba, Alborz Geramifard

Facebook Reality Labs & Facebook AI
simmc@fb.com

Abstract

The SIMMC challenge at DSTC9 aims to lay the foundations for multimodal virtual assistant agents that can engage with the real-world, handle multimodal inputs, and perform multimodal actions. We thus focus on task-oriented dialogs that encompass a situated multimodal user context in the form of a co-observed image or virtual reality (VR) environment. The context is dynamically updated on each turn based on user input and assistant actions. The SIMMC track focuses on three main subtasks: (1) structural API call prediction, (2) assistant response generation, and (3) dialog state tracking. We describe these tasks and the respective evaluation metrics for each task. We also present models that have achieved state-of-the-art performance on each task. We conclude with a summary of insights and opportunities for further research arising from the results of the first SIMMC challenge.

Introduction

The Situated Interactive MultiModal Conversations (SIMMC) challenge at DSTC9 is based on the recently proposed SIMMC task and accompanying datasets as described in (Moon et al. 2020). The goal of these datasets is to address a gap in the existing field of multimodal dialog research by focusing on task-oriented dialogs grounded in the visual context that is co-observed by both speakers, and evolves as the dialog progresses.

The SIMMC datasets focus on shopping experiences in the fashion and furniture domains as these experiences provide a rich interactive context in which task-oriented dialogs can be situated. The visual context is either a series of images—for fashion—or a virtual reality (VR) scene—for furniture. In both settings the ground truth of which items are present is known and provided in the dataset, thus allowing modelling efforts to be focused on ingesting and utilizing semantics about the scene without necessarily having to deal with raw pixels.

The SIMMC challenge intends to foster progress in the area of enabling virtual assistants to utilize contexts from multiple sources, and make progress towards deployment of such skills in the real world.

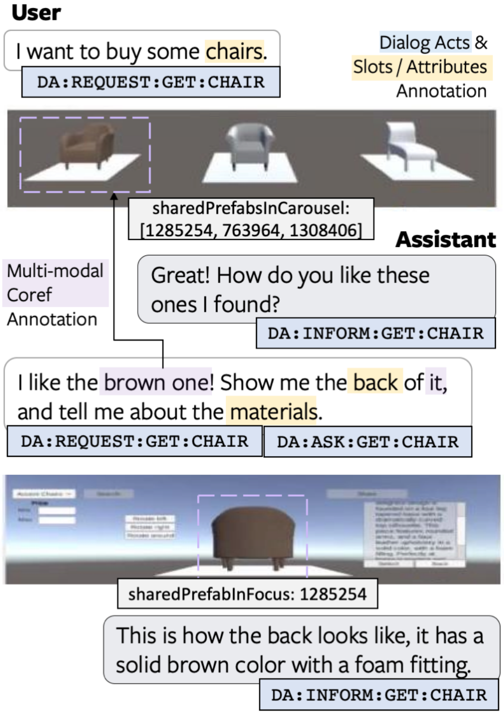

User
I want to buy some chairs.
DA:REQUEST:GET:CHAIR
Dialog Acts & Slots / Attributes Annotation
sharedPrefabsInCarousel: [1285254, 763964, 1308406]
Assistant
Multi-modal Coref Annotation
Great! How do you like these ones I found?
DA:INFORM:GET:CHAIR
I like the brown one! Show me the back of it, and tell me about the materials.
DA:REQUEST:GET:CHAIR
DA:ASK:GET:CHAIR
sharedPrefabInFocus: 1285254
This is how the back looks like, it has a solid brown color with a foam fitting.
DA:INFORM:GET:CHAIR

Figure 1: Illustration of a SIMMC dialog: a user and an assistant interact in a co-observed multimodal environment for a shopping scenario. The dialog is grounded in an *evolving* multimodal context. The ground-truth of which items (*e.g.*, prefabs) appear is known for each view.

An example of the type of dialog found in the SIMMC-Furniture dataset is shown in Fig. 1. In addition to responding to the user, the assistant can update the environment by selecting multimodal actions in this VR setting, e.g.,

> visually presenting recommended chairs ... or responding to the request "I like the brown one. *Show me the back* of it." by executing the actions of *focusing on*, and *rotating* the indicated item (Moon et al. 2020).

In the above example, elements that contribute to the overall novelty of the SIMMC Task, *i.e. action prediction* and multimodal coreference resolution, are indicated using italics and underling, respectively. A further contributor to

[*] Equal contributions.

Table 1: **Comparison with the existing multimodal dialog corpora. Notation**: (U ↔ A) Utterance to action pair labels. (Task-oriented) Includes API action prediction, Q&A, recommendation, item / image retrieval and interaction. (Meta) includes scene descriptions and/or catalog information of items in the scene. (Semantic) Dialog annotations such as NLU, NLG, DST, and Coref. (Situated) VR environment and/or new highlighted images.

Dataset	**Modality**	**Task**	**Provided Context**		**Updated**	**Annotation**
			Q'er	**A'er**	**Context**	**Granularity**
Visual Dialog (Das et al. 2017)	Image	Q&A	N/A	Visual	N/A	N/A
CLEVR-Dialog (Kottur et al. 2019)	Simulated	Q&A	N/A	Visual	N/A	N/A
GuessWhat (de Vries et al. 2017)	Image	Q&A	N/A	Visual	N/A	N/A
Audio Visual Scene-Aware Dialog (Hori et al. 2018)	Video	Q&A	Key Frames	Video	N/A	N/A
TalkTheWalk (de Vries et al. 2018)	Image	Navigation	Visual	Visual + Meta	Location	U ↔ A
Visual-Dialog Navigation (Thomason et al. 2019)	Simulated	Navigation	Visual	Visual + Meta	Location	U ↔ A
Relative Captioning (Guo et al. 2018)	Image	Image Retrieval	Visual	Visual + Meta	New Image	U ↔ A
MMD (Saha, Khapra, and Sankaranarayanan 2018)	Image	Image Retrieval	Visual	Visual + Meta	New Image	U ↔ A
SIMMC (Moon et al. 2020)	**Image/VR**	**Task-oriented**	**Visual**	**Visual + Meta**	**Situated**	**U ↔ A + Semantic**

datasets' novelty is that the assistant's actions, which are presumed to be conditioned on the preceding dialog with the user, may change the co-observed context which in turn influences the subsequent turns of the dialog.

Related Datasets and Challenges

Tab. 1, adapted from Moon et al. (2020), compares SIMMC to existing similar VQA and visual dialog datasets and challenges.

Key differentiators with previous multimodal dialog datasets are: *(a)* SIMMC assumes a mutually shared "co-observed" multimodal context between a user and an assistant. This contrasts with visual question answering (Q&A) tasks (Das et al. 2017; Kottur et al. 2019; de Vries et al. 2017, 2018) where the questioner (Q'er) and answerer (A'er) have differing views. *(b)* SIMMC focuses on task orientated dialogs and to this end includes semantic annotations that extend to a multimodal setting areas which have been a key focus of the dialog literature, such as dialog state tracking (DST) and policy learning (Wu et al. 2019; Gao et al. 2019; Chao and Lane 2019). To this end Moon et al. (2020) proposed an associated SIMMC annotation schema which "allows for a more systematic and structural approach for visual grounding of conversations." *(c)* In contrast to image retrieval (Guo et al. 2018; Saha, Khapra, and Sankaranarayanan 2018) and visual navigation tasks (Thomason et al. 2019; de Vries et al. 2018) where context updates are limited to introduction of new scenes or images, SIMMC-Furniture allows agent actions at both scene level and object level *e.g.*, changing the view of a specific object within a scene, while SIMMC-Fashion introduces the concept of visual memory as part of the preceding context for the dialog. *(d)* The object level manipulations in SIMMC-Furniture additionally introduce multimodal agent actions *e.g.*, *'rotate,'*, *'search,'* and *'add to cart'* not found in tradition task-oriented conversational datasets (Henderson, Thomson, and Williams 2014; Budzianowski et al. 2018; Eric et al. 2019; Rastogi et al. 2019; Chen et al. 2020). *(e)* SIMMC tasks emphasize semantic processing, while work in visual Q&A and visual dialog has heavily focused on language grounding in raw image pixels.

Table 2: **SIMMC Datasets Statistics**. †Additional dialogs in aural medium where annotators exchanged audio messages instead of text.

Statistics	**Furniture (VR)**		**Fashion (Image)**
	Text	Audio†	
Total # dialogs	6.4k	1.3k	6.6k
Total # utterances	97.6k	15.8k	71.2k
Avg # rounds / dialog	7.62	7.16	5.39
Avg # tokens (user)	11.0	N/A	11.10
Avg # tokens (assistant)	12.2	N/A	10.87

SIMMC Datasets

The SIMMC datasets contain about $13k$ human-to-human dialogs (totaling about $169k$ utterances) split across two domains; furniture (VR) and fashion (images). Datasets statistics are shown in Tab. 2; reproduced from (Moon et al. 2020). Both datasets were collected through the SIMMC Platform (Crook et al. 2019), an extension to ParlAI (Miller et al. 2017) that enables a multi-player / Wizard of Oz (Kelley 1984) setting for multimodal conversational data collection, or a single player mode for system evaluation.

The included semantic-level fine-grained annotations ground the visual context, allowing for a more systematic and structural study for visual grounding of conversations. Annotation labeling is centered around atomic *objects* that appear in the text, visual context and associated catalogs of object attributes. This allows for flexible annotation of natural utterances as well as multimodal coreferences that link the annotated language with objects in the context.

Task Definitions

We present three subtasks primarily aimed at replicating human-assistant actions in order to enable rich and interactive shopping scenarios.

Subtask 1: Structural API Call Prediction focuses on predicting the assistant action as an API call given the dialog and the multimodal contexts as inputs. Since accuracy does not account for the existence of multiple valid actions,

Table 3: Summary of each team's results on Test-Std split, average of Furniture and Fashion (*Team 5 submitted results only for Fashion). Best results from each team are shown. **(1) API prediction** via accuracy, perplexity and attribute accuracy, and, **(2) Response prediction** via BLEU, recall@k (k=1,5,10), mean rank, and mean reciprocal rank (MRR). **(3) Dialog State Tracking (DST)**, via slot and intent prediction F1. ↑: higher is better, ↓: lower is better. ‡ Tied result; within one standard error.

Teams	**Subtask 1. API Prediction**			**Subtask 2. Response Prediction**						**Subtask 3. DST**	
	Acc↑	A.Acc↑	Perp↓	BLEU↑	MRR↑	r@1↑	r@5↑	r@10↑	Mean↓	Slot F1↑	Intent F1↑
Baseline	79.3	63.7	1.9	0.061	0.145	7.2	19.8	27.3	39.2	62.4	62.1
Team 1	80.2	74.6	2.0	0.105	0.326	21.1	43.6	56.8	18.8	77.8	76.7
Team 2	**82.5‡**	69.8	1.8	0.082	0.074	2.5	8.3	13.6	47.7	-	-
Team 3	79.4	73.2	-	**0.128**	0.381	26.3	50.3	61.8	15.5	**79.1**	78.1
Team 4	**81.3‡**	**73.9**	3.5	0.108	**0.673**	52.6	87.4	95.1	3.2	78.6	77.7
Team 5*	-	-	-	-	0.390	26.7	52.1	66.0	14.8	-	-

we use perplexity (defined as the exponential of the mean log-likelihood) alongside accuracy. To also measure the correctness of the predicted action (API) arguments, we use attribute accuracy compared to the collected datasets.

Subtask 2: Assistant Response Generation examines the relevance of the assistant response in the current turn. We propose two variants of this subtask; *(a)* a generative approach (conditional language modeling problem), which uses BLEU-4 as a proxy for measuring the closeness between the generated response and the ground-truth response given in the corpus, and, *(b)* as a retrieval/ranking problem, where the model retrieves the ground-truth response from a pool of 100 candidates (randomly chosen and unique to each turn). We use recall@k ($k = \{1, 5, 10\}$), mean rank, and mean reciprocal rank as the corresponding metrics.

Subtask 3: Dialog State Tracking (DST) aims to systematically track the dialog acts and the associated slot pairs across multiple turns, as represented in the flexible ontology developed to represent the SIMMC multimodal context. We use the intent and slot prediction metrics (F1), inline with prior work in DST.

Evaluation. For each subtask in this challenge, we enforce the following priority over the metrics discussed above, designed to highlight the desired model behavior. Refer (Moon et al. 2020) for an elaborate discussion on evaluation.

- **Subtask 1:** Action accuracy, action attribute accuracy, action perplexity
- **Subtask 2:** *(Generation)* BLEU-4, *(Retrieval)* mean reciprocal rank, recall@k ($k = \{1, 5, 10\}$), mean rank
- **Subtask 3:** Slot F1 score, Intent F1 score

The entry with the most favorable (higher or lower) performance on the first metric is labelled as a *winner candidate*. Further, all other entries within one standard error of the *winner candidate's* performance are also considered as candidates. If there are more than one candidate according to the metric, we move to the next metric in the priority list and repeat this process until we have a single winner candidate, which is declared as the **subtask winner**. The declaration of winners and runners up is made on a per team basis, considering only the best performing model submitted by that team for that subtask.

Baselines and Submitted Systems

The challenge saw a total of 13 model entries from 5 teams across the world. We describe the major modeling decisions of these teams in this section and provide a comparative summary in Tab. 4. For more details please refer to the referenced teams' papers.

Baselines In our analysis, we use the baselines from (Moon et al. 2020) to compare the performance of the submitted challenge entries. At a high level, Moon et al. (2020) develop a joint conversational model for subtask 1 and subtask 2, and model subtask 3 separately. For the former, they follow a pipeline architecture to *(a)* encode the user utterance and dialog history, *(b)* fuse it with the multimodal context, *(c)* predict the API call for the turn, and finally, *(d)* produce the assistant response using a conditional language model. For the latter, an end-to-end task-oriented dialog model (Hosseini-Asl et al. 2020) is adopted and extended to ingest the multimodal context, and predict the intent and slot values as a classification problem. Please see (Moon et al. 2020) for more details.

Team 1 (Kung et al. 2020) submitted an ensemble of GPT-2 models trained jointly on all three subtasks and across both domains. Specifically, they added a discriminative classifier consisted of multiple fully connected layers for subtask 1 (API Prediction), while keeping subtasks 2a (Response Generation) and 3 (DST) as generative tasks, following the baseline provided by (Moon et al. 2020). For the response retrieval subtask 2b, they ranked the retrieval candidates based on their BLEU and METEOR similarity scores with the generated responses from subtask 2a. In addition, auxiliary features as input such as segment embeddings were used to better leverage the visual information.

Team 2 (Kim et al. 2020) submitted an ensemble of models based on the baselines (Moon et al. 2020) released as part of the competition. While the baselines model subtask 1 and 2 jointly and subtask 3 separately, team 2 used the predicted dialog state outputs from subtask 3 baseline as inputs for subtasks 1 and 2. Additionally, they used two sophisti-

Table 4: **Summary of submitted models. Notation**: *Sub.* is the subtasks for which results were submitted. *MM Rep.* is the method used for ingesting MultiModal context. *Descrim. Train* indicates if discriminative model training on positive and negative candidates was used. *Model Rank* is the raw ranking using the top metric for that subtask without consideration of standard error and thus should be considered as only providing an indicative ordering. *Team 5 submitted results only for Fashion. Rank in parenthesis is for Fashion only. *MM Fusion Ensembles A, B, C* are each different combinations of baseline and multimodal fusion models. *MAG / MMI* specialized multimodal fusion gates; MAG (Rahman et al. 2020) and MMI (Yu et al. 2020). *S*, *M*, and *L* indicate GPT-2 small, medium and large models, respectively. *L(t)* is GPT-2 Large trained on train set only, others are trained on train and dev sets. *FC* indicates additional full-connected layers. *BLEU/METEOR* and *cosine sim.* indicate generative models adapted to the retrieval task (2b) by using BLEU/METEOR or cosine similarity metrics to measure the distance between retrieval candidates and the model's predicted response.

(a) Summary of models submitted for Subtasks 1, 2a and 3; API Prediction, Response Generation and DST.

Teams	**Models**	**Sub.**	**Joint Train**		**Ensemble**	**Pretrain**	**MM Rep.**	**Model Rank**		
			subtasks	x-domain				sub1	sub2a	sub3
Team 1	GPT-2 + FCs	1, 2a, (2b), 3	1, 2a, 3	yes	yes	GPT-2	stringified	4	5	5
Team 2	MM Fusion Ensemble A	1	1, 2a	no	yes	–	MAG / MMI	1	·	·
	MM Fusion Ensemble B	2a	1, 2a	no	yes	–	MAG / MMI	·	7	·
Team 3	GPT-2 (M + S)	1	1, 2a, 3	no	yes	GPT-2	stringified	5	·	·
	GPT-2 (L + L(t))	2a, (2b), 3	2a, 3	no	yes	GPT-2	stringified	·	3	2
	GPT-2 (L + S)	2a, (2b), 3	2a, 3	no	yes	GPT-2	stringified	·	1	1
	GPT-2 (L + L(t) + S)	2a, (2b), 3	2a, 3	no	yes	GPT-2	stringified	·	2	3
Team 4	BART-Base	1, 2a, 3	1, 2a, 3	no	no	BART	stringified	3	6	6
	BART-Large	1, 2a, 3	1, 2a, 3	no	no	BART	stringified	2	4	4

(b) Summary of models submitted for Subtask 2b; Response Retrieval.

Teams	**Models**	**Sub.**	**Joint Train**		**Ensemble**	**Pretrain**	**MM Rep.**	**Descrim. Train**	**Model Rank**
			subtasks	x-domain					
Team 1	GPT-2 + FCs + BLUE/METEOR	2b	1, 2a, 3	yes	yes	GPT-2	stringified	no	6 (7)
Team 2	MM Fusion Ensemble C	2b	1, 2a	no	yes	GPT-2	MAG / MMI	no	7 (8)
Team 3	GPT-2 (L+L(t)) + cosine sim.	2b	2a, 3	no	yes	GPT-2	stringified	no	5 (5)
	GPT-2 (L+S) + cosine sim.	2b	2a, 3	no	yes	GPT-2	stringified	no	4 (6)
	GPT-2 (L+L(t)+S) + cosine sim.	2b	2a, 3	no	yes	GPT-2	stringified	no	3 (4)
Team 4	BART-Large Bi-Encoder	2b	2b	no	no	BART adapted on 1, 2a, 3	stringified	yes	1 (1)
	BART-Large Poly-Encoder	2b	2b	no	no		stringified	yes	2 (2)
Team 5*	BERT + log-likelihood	2b	2b	no	no	BERT	stringified	no	- (3)

cated multimodal fusion models designed for transformer architectures—MAG (Rahman et al. 2020) and MMI (Yu et al. 2020) in their implementation—to fuse the predicted dialog state with the utterance encoding at the current turn. The final predictions from the ensemble was obtained by averaging the individual model scores for subtask 1 and 2. Though this augmentation hurt their performance for subtask 2, their model achieved a gain of about 3 points on action accuracy and 6 points on action attribute accuracy for API call prediction (subtask 2).

Team 3 (Jeong et al. 2020) submitted a varied set of ensembles of GPT-2 (Radford et al. 2019) models that were of differing sizes (large, medium and small) and trained on differing partitions of the training data; train only, or train plus dev. For the ensemble submitted for subtask 1, each GPT-2 model was independently trained on three joint tasks—subtask 1, subtask 2a and subtask 3—using a simple language model loss that optimized over the concatenated string containing the dialog history, multimodal context, user utterance, dialog state, system response, and API call. This model can predict all three subtasks on which it was trained but it's results were only submitted to subtask 1. In the ensemble submitted for subtasks 2a and 3 each GPT-2 model was again independently trained with a simple language model loss but only on the joint tasks of subtask 2a and subtask 3, *i.e.*, the above concatenated string excluding API call. For subtask 2b the generated response of the model trained on subtask 2a and 3 was compare to each candidate response using word tokenization and cosine similarity to select the response. For all models the dialog state representation was preprocessed to remove camel-case and non-natural punctuation before training. An ensemble beam search over each model's prediction was used to generate the final prediction.

Team 4 (Huang et al. 2020) submitted two BART (Lewis et al. 2020) models (BART-Large and BART-Base) for subtasks 1, 2a, and 3. Both were trained to jointly predicted the dialog state (subtask 3), API call (subtask 1) and response (subtask 2a) as a single string target when given the dialog history, multimodal context and user utterance. For response retrieval they submitted two BART-encoder based models; Bi-encoder and Poly-encoder (Humeau et al. 2020; Mazaré et al. 2018; Dinan et al. 2019). In both of these models the encoder weights were initialized from the jointly trained BART models trained on subtasks 1, 2a, and 3. This gave a combination of four models on this subtask, *i.e.*, BART-Large or BART-Base with Bi-encoder or Poly-encoder. We, however, only include results for BART-Large Bi/Poly-encoders. Model weights were then adapted to the retrieval task.

Team 5 (Senese et al. 2020) submitted a BERT-based model addressing the Assistant response retrieval task (subtask 2b), trained using the cross-entropy loss. Specifically, the submitted model includes a self-attention module, an encoder-decoder attention module, and an item-attention module. At inference time, the log-likelihood of each candidate response (given the input utterances and multimodal context) is calculated for each token. To rank the candidate responses, two scoring modules were used: (1) normalized sum of log-likelihood scores for each token (to avoid a scoring bias towards short responses), and (2) token match rate of the annotated item attributes in each candidate response. Candidate responses with the highest sum of these two scores were used as final predictions.

Challenge Results

The submitted entries set new state-of-the-art in all three subtasks as summarized in (Tab. 3).

The winner of the structural API call prediction subtask (subtask 1) was the BART-Large model from Team 4. This model was also one of two runners up on subtask 2a, and the runner up on subtask 3.

The winner of the response retrieval subtask (subtask 2b) was the BART-Large Bi-encoder from Team 4. This model achieved a mean reciprocal rank (MRR) of 0.67, beating their BART-Large Poly-Encoder model by 0.02 points, and with a substantial lead of 0.29 points compared to the runner up team on this subtask.

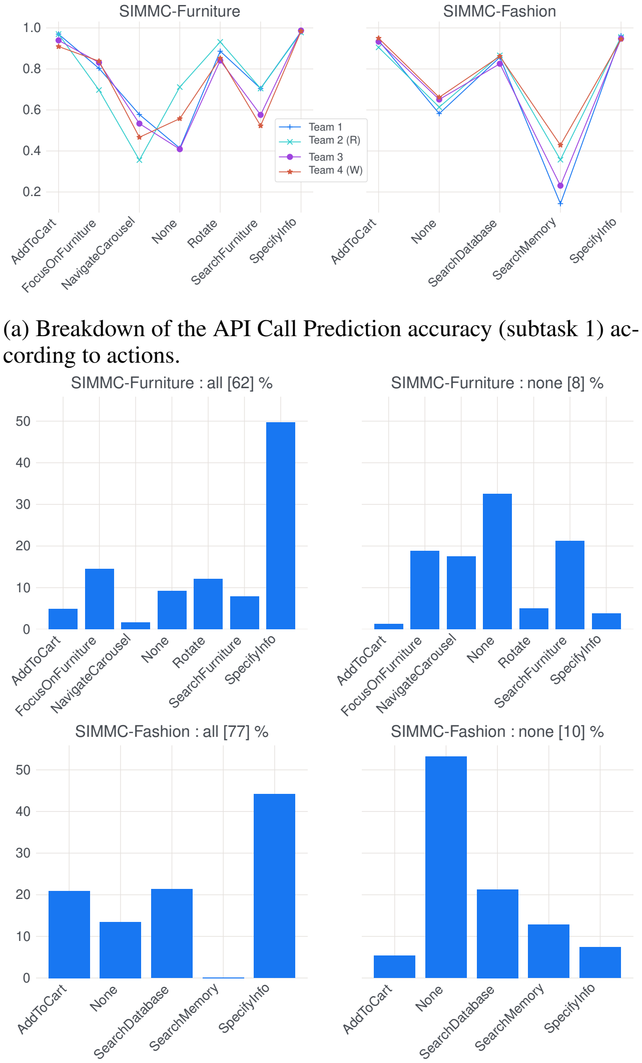
The winner of the response generation and DST subtasks (subtask 2a and subtask 3) was an ensemble of GPT-2 models from Team 3.

Performance Analysis

In this section, we analyze the performance of model entries by breaking it down across several aspects for each subtask.

Subtask 1: Structural API Call Prediction. Fig. 2a shows the breakdown of API prediction accuracy for each team by action type for both SIMMC-Furniture and SIMMC-Fashion. The key observations are:

- All teams successfully predict the AddToCart and SpecifyInfo actions with 90% and 95% accuracy respectively, for both the domains. This is intuitive as the models seem to pick up on important cues informing the user intents for these particular API calls. For example, *"Can you please add this to my cart?"* clearly indicates the intention to add the discussed product to the cart. Similarly, *"What is its price and customer rating?"* denotes a request to provide additional information about the product under discussion.

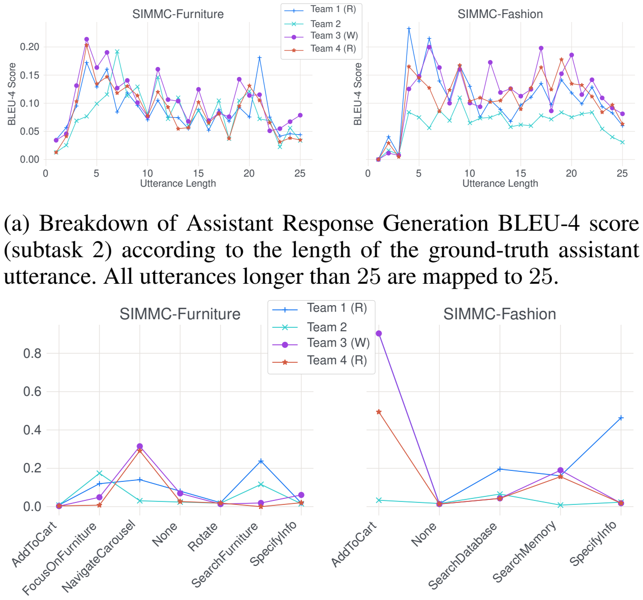


(a) Breakdown of the API Call Prediction accuracy (subtask 1) according to actions.

(b) Breakdown of instances categorized based on whether **all** or **none** of the participating model entries accurately predicted the ground-truth API call.

Figure 2: Analysis of the competition entries for API Call prediction (Subtask 1). See text for more details.

- On the other hand, all models perform poorly on NavigateCarousel and None actions for SIMMC-Furniture, and SearchMemory for fashion. The accuracy for these actions are in the 20%–40% range for most models. A possible explanation is due to the equally valid choice betwen either showing items from the catalog with existing filters (mapped to SearchFurniture or SearchDatabase) or requesting more information to refine the search (mapped to None).
- Notice that the models from both team 4 (winner) and team 2 (runner-up) perform closely with respect to the API call prediction task, as seen from an overall accuracy of 81.3% and 82.5% (Tab. 3). The winner has been declared based on the API action attribute accuracy.



(a) Breakdown of Assistant Response Generation BLEU-4 score (subtask 2) according to the length of the ground-truth assistant utterance. All utterances longer than 25 are mapped to 25.

(b) Breakdown of Assistant Response Generation BLEU-4 score (subtask 2) according to actions.

Figure 3: Analysis of the competition entries for Assistant Response Generation (Subtask 2). See text for more details.

Further, we identify instances on which **all** and **none** of the competition entries were able to accurately predict the corresponding ground-truth API call. We breakdown each of these instance categories further into the ground-truth actions in Fig. 2b. For SIMMC-Furniture, the **all** and **none** categories compose 62% and 8% of all the test instances, respectively. The corresponding numbers for SIMMC-Fashion are 77% and 10%. Using these categories as weak indicators of *easy* and *hard* instances for subtask 1, one could conclude that SIMMC-Furniture contains a smaller percent of both *easy* and *difficult* instances when compared to SIMMC-Fashion. Finally, Fig. 2b also provides additional evidence for the trends observed earlier.

Subtask 2: Assistant Response Generation. We analyze the performance of models by comparing BLEU-4 scores (generation category) based on: *(a)* length of ground-truth assistant utterance in Fig. 3a, and *(b)* corresponding ground-truth API call in Fig. 3b. Following are the takeaways:

- As expected, BLEU-4 score decreases with the length of the utterances for both the domains on an average.
- Though the smoothing for BLEU-4 contributes partially to the low values for utterance lengths of 1–3, a good proportion of these utterances contained information about the catalog item, e.g., price and dimension. On further investigation, we found that most of the models were unable to correctly respond with these attributes. This highlights the need for a better catalog integration with the response generation model.
- Comparing BLEU-4 scores for the action `AddToCart`, models perform better on SIMMC-Fashion on an average compared to SIMMC-Furniture. This could be due to a larger percent of `AddToCart` in the former (18%) when compared to the latter (3%), leading to this discrepancy.
- The BLEU-4 score for `SpecifyInfo` is lower than the overall score for all the models. Once again, this points to the need for a better modeling of the catalog information. Examples of model responses for `SpecifyInfo` by the winner and runner-up team are given in Tab. 5.

Another important distinction between the winner (Team 4) and the rest of the entries (and baseline) is the use of *discriminative* training for the assistant response generation. Specifically, their loss function trains to not only increase the likelihood of ground-truth response (similar to a language model) but also to decrease the likelihood of other assistant response targets in the batch, which act as negative examples. This results in a superior performance in the subtask 2 (retrieval category), where the former outperforms the rest by at least 26 points on the recall@1 metric (Tab. 3).

Subtask 3: Dialog State Tracking (DST). Fig. 4a shows a breakdown of the DST results based on slot types, for the entries summarized in Tab. 3. Specifically, we report F1 scores for *attribute* slot types that describe objects (*e.g.*, "How many [O.color green] ones do you have?") or intents (*e.g.*, "I am looking for [.intendedRoom bedroom] lamps"), and for *object* slots, which represent object indices that correspond to their parent intents (*e.g.* "[DA:REQUEST:GET:TABLE Please add [TABLE_1 it] to the cart.]") The object slot prediction task thus can also be framed as multimodal coreference resolution problem. It can be seen that the F1 scores for attribute slots have higher variances across different entries compared to those for object slots. This result shows that the different approaches proposed by each team had relatively small influences on the multimodal coreference resolution performance.

Fig. 4b and Fig. 4c show the object slot F1 tracking snapshots at varying turn indices as cohorts, averaged over the dialogs, for SIMMC-Furniture and SIMMC-Fashion respectively. For both domains, we observe that the object slot F1 performances decrease in general as more objects are mentioned and introduced in the multimodal context. Note that none of the proposed models showed significant improvement over other baselines in suppressing the degradation in the object slot prediction performances over time.

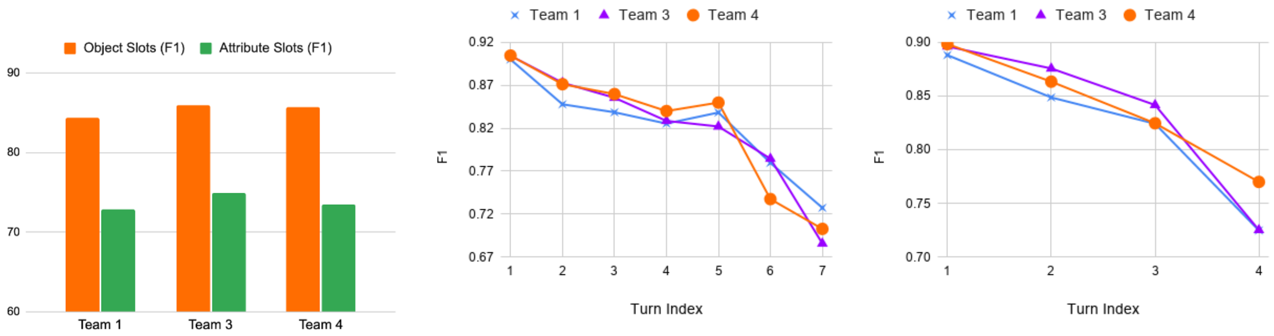
What We Learnt From the Challenge

Pretrained language models show promise in multimodal settings. The strong performance of pretrained language models such as GPT-2 and BART when adapted to these task indicate that such models are flexible enough to be able to ingest relatively simple multimodal context and thus can be used in a multimodal setting with a high degree of success.

Multimodal context helps but gaps remain. Moon et al. (2020) showed that the availability of multimodal context does boost performance on subtask 3 (DST). Although not directly addressed by these sets of results, we believe that modeling multimodal context boosts performance across

Table 5: Examples of assistant responses generated by the winner (Team 3) and runner-up (Team 4) for subtask 2a from SIMMC-Furniture, where the ground-truth API call is `SpecifyInfo`, *i.e.*, seeking information about catalog items. The category of examples compared to the ground-truth assistant response is mentioned in the last column.

Ground Truth	**Winner (Team 3)**	**Runner-up (Team 4)**	**Category**
It is reasonable priced at $69.99.	The price is $69.99.	It is priced at $69.99	Both models predict the right attributes
It’s $48.00.	It is $135.99.	It is $82.81.	Both models are incorrect
The dimensions for that one is 53.5” ×30” ×36” $(L\times W\times H)$	The dimensions are Width 60 Depth 44.92 Height 44.33	The dimensions are 60”W × 44.92”D × 44.33”H.	Both models are incorrect but consistent with each other
The width is 18, depth 15.06 and height is 16.5.	It is 18 inches wide, 15.06 inches deep and 16.5 inches high.	The dimensions are 18.5”W × 15.5”D × 35.5”H.	Winner model is correct while the runner-up is incorrect
Sure. Dimensions are 86 × 37 × 32 inches and cost is $829.	This sofa is made by Jarrard. If you would like to know details such as price or see a different angle, tell me and I can show you.	This is the Jarrard Sofa. It has tapered wood legs and a plush back.	Both models respond with mismatched attributes about the catalog items



(a) F1 per slot types (Average)

(b) Object F1 for SIMMC-Furniture

(c) Object F1 for SIMMC-Fashion

Figure 4: Analysis of the competition entries for Dialog State Tracking (Subtask 3). *(a)* Breakdown of Slot F1 results according to slot types (object & attribute slots). *(b)* Average object slot tracking results at varying turn indices for SIMMC-Furniture and *(c)* for SIMMC-Fashion. See text for more details.

all of the subtasks. However: (1) DST metrics show that accuracy falls off with increasing dialog length. Given that, especially in SIMMC-Furniture, the multimodal context provides a grounded set of items which are likely to be the most salient in each turn, this results suggest that the models are not necessarily taking full advantage of context provided. (2) Relatively low BLEU scores for response generation (subtask 2a) indicate there remains a significant opportunity for improving assistant response prediction. This is further reinforced by the indication that discriminatively trained retrieval models on subtask 2b demonstrate much better performance than the generative models that match candidates based on similarity to a generated response or model loglikehood score.

Need for a better and scalable catalog integration. Eyeballing the generated responses, given in Tab. 5, indicates that these models are powerful enough to avoid returning bland and safe responses (often observed in generative models (Li et al. 2015)) but fail to reliably integrate catalog information. This maybe indicative of a failure of model architectures to utilise the additional context available from the catalog or a more general problem with utilisation of multimodal context in response generation. Better and scalable multimodal integration for catalog information is crucial in task-oriented settings where systems are expected to relay accurate information to users.

Scaling up multimodal complexity. An additional area for future investigation is to examine the related question of how well does the simple ‘stringified’ approach to ingesting multimodal context handle increasing complex scenarios as the number of items in the scene increases and thus format become increasing long and potentially increasing nested.

Conclusion

Through the organization of the SIMMC Challenge in DSTC9, we aim to motivate the research community to consider the important problem of situated and interactive multimodal task-oriented conversations, which paves the ways towards virtual assistants that can handle many everyday, real-world applications. We hope that the insights gained throws light on the challenges of such multimodal dialogs and inspires multiple follow-up lines of research.

References

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*.

Chao, G.-L.; and Lane, I. 2019. BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer. In *INTERSPEECH*.

Chen, M.; Liu, R.; Shen, L.; Yuan, S.; Zhou, J.; Wu, Y.; He, X.; and Zhou, B. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 459–466. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.58.

Crook, P. A.; Poddar, S.; De, A.; Shafi, S.; Whitney, D.; Geramifard, A.; and Subba, R. 2019. SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. *ASRU* .

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *CVPR*.

de Vries, H.; Shuster, K.; Batra, D.; Parikh, D.; Weston, J.; and Kiela, D. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367* .

de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Eric, M.; Goel, R.; Paul, S.; Kumar, A.; Sethi, A.; Ku, P.; Goyal, A. K.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines. *arXiv preprint arXiv:1907.01669* .

Gao, S.; Abhishek Seth and, S. A.; Chun, T.; and Hakkani-Ture, D. 2019. Dialog State Tracking: A Neural Reading Comprehension Approach. In *SIGDIAL*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. 2018. Dialog-based interactive image retrieval. In *NeurIPS*.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The Second Dialog State Tracking Challenge. In *SIGDIAL*.

Hori, C.; Cherian, A.; Marks, T. K.; and Metze, F. 2018. Audio Visual Scene-Aware Dialog Track in DSTC8. *DSTC Track Proposal* .

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. 2020. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations. *9th Dialog System Technology Challenge (DSTC-9) Workshop at AAAI* .

Humeau, S.; Shuster, K.; Lachaux, M.; and Weston, J. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL https://openreview.net/forum?id=SkxgnnNFvH.

Jeong, Y.; Lee, S.; Koa, Y.; and Seo, J. 2020. TOM : End-to-End Task-Oriented Multimodal Dialogue System with GPT-2. *9th Dialog System Technology Challenge (DSTC-9) Workshop at AAAI* .

Kelley, J. 1984. An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.* 2, 1: 26–41. ISSN 1046-8188. doi: 10.1145/357417.357420. URL http://doi.acm.org/10.1145/357417.357420.

Kim, B.; Lee, I.; Jeong, Y.; Ko, Y.; Koo, M.-W.; and Seo, J. 2020. Improving Multimodal API Prediction via Adding Dialog State and Various Multimodal Gates. *9th Dialog System Technology Challenge (DSTC-9) Workshop at AAAI* .

Kottur, S.; Moura, J. M.; Parikh, D.; Batra, D.; and Rohrbach, M. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166* .

Kung, P.-N.; Yang, T.-H.; Chang, C.-C.; Hsu, H.-K.; Liou, Y.-J.; and Chen, Y.-N. 2020. Multi-Task Learning for Situated Multi-Domain End-to-End Dialogue Systems. *9th Dialog System Technology Challenge (DSTC-9) Workshop at AAAI* .

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* .

Mazaré, P.-E.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2775–2779. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1298. URL https://www.aclweb.org/anthology/D18-1298.

Miller, A. H.; Feng, W.; Fisch, A.; Lu, J.; Batra, D.; Bordes, A.; Parikh, D.; and Weston, J. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476* .

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. *The 28th International Conference on Computational Linguistics (COLING)* .

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.

Rahman, W.; Hasan, M. K.; Lee, S.; Bagher Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://www.aclweb.org/anthology/2020.acl-main.214.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2019. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *AAAI*.

Saha, A.; Khapra, M. M.; and Sankaranarayanan, K. 2018. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*.

Senese, M. A.; Benincasa, A.; Caputo, B.; and Rizzo, G. 2020. A Response Retrieval Approach for Dialogue Using a Multi-Attentive Transformer. *9th Dialog System Technology Challenge (DSTC-9) Workshop at AAAI* .

Thomason, J.; Murray, M.; Cakmak, M.; and Zettlemoyer, L. 2019. Vision-and-dialog navigation. *arXiv preprint arXiv:1907.04957* .

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *ACL*.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3342–3352. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.306. URL https://www.aclweb.org/anthology/2020.acl-main.306.