An Analysis of State-of-the-Art Models for Situated Interactive MultiModal Conversations (SIMMC)

Satwik Kottur*, **Paul Crook***, **Seungwhan Moon***,
Ahmad Beirami, Eunjoon Cho†, Rajen Subba†, Alborz Geramifard
Facebook Research, Facebook AI
simmc@fb.com

Abstract

There is a growing interest in virtual assistants with multimodal capabilities, *e.g.*, inferring the context of a conversation through scene understanding. The recently released Situated and Interactive Multimodal Conversations (SIMMC) dataset addresses this trend by enabling research to create virtual assistants, which are capable of taking into account the scene that user sees when conversing with the user and also interacting with items in the scene. The SIMMC dataset is novel in that it contains fully annotated user-assistant, task-oriented dialogs where the user and an assistant co-observe the same visual elements and the latter can take actions to update the scene.

The SIMMC challenge, held as part of the Ninth Dialog System Technology Challenge (DSTC9), propelled the development of various models which together set a new state-of-the-art on the SIMMC dataset. In this work, we compare and analyze these models to identify *'what worked?'*, and the remaining gaps; *'what next?'*. Our analysis shows that even though pretrained language models adapted to this setting show great promise, there are indications that multimodal context isn't fully utilised, and there is a need for better and scalable knowledge base integration. We hope this first-of-its-kind analysis for SIMMC models provides useful insights and opportunities for further research in multimodal conversational agents.

1 Introduction

The Situated Interactive MultiModal Conversations (SIMMC) challenge[1] at DSTC9 (Gunasekara et al., 2020) aims to lay the foundations for virtual assistant agents that can engage with the real-world, handle multimodal inputs, and perform multimodal actions. It focuses on task-oriented dialogs that encompass a situated multimodal user context in the form of a co-observed image or virtual reality (VR) environment, which is dynamically updated on each turn based on the user input and the assistant action.

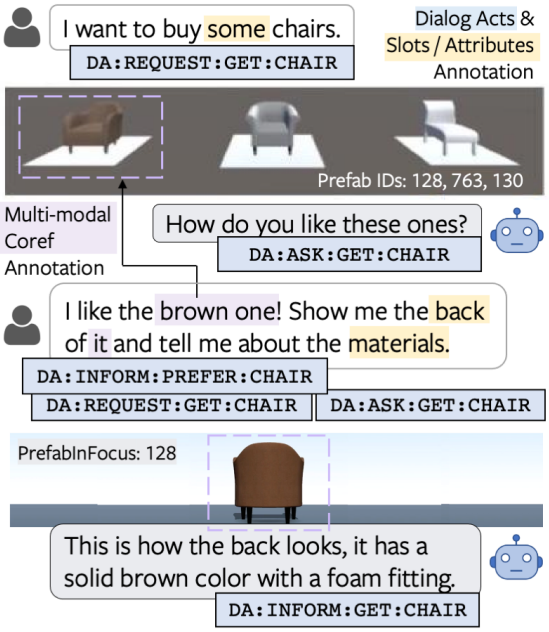


Figure 1: Illustration of a SIMMC dialog: a user and an assistant interact in a co-observed, *evolving* multimodal environment for a shopping scenario. For the sake of brevity, the annotations shown are incomplete. For details of the annotation schema, see Moon et al. (2020). Figure adapted from Moon et al. (2020).

Figure 1 illustrates an exemplary SIMMC dialog, where a user interacts with an assistant with the goal of browsing for furniture. Here, the assistant updates the co-observed environment leading to a new multimodal context based on the dialog, *e.g.*, visually presenting recommended chairs in a VR environment, 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. These actions in turn update the co-observed multimodal context, which grounds

* Joint first authors
† Work done when EC and RS were at Facebook
[1] github.com/facebookresearch/simmc

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 144–153
July 29–31, 2021.

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	N/A	Visual	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 et al., 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**

Table 1: **Comparison with the existing multimodal dialog corpora (Moon et al., 2020). Notation**: (U ↔ A) Utterance to action pair labels. (Task-oriented) Includes API action prediction, Q&A, recommendation, item / image retrieval and interaction. (Semantic) Dialog annotations such as NLU, NLG, DST, and Coref. (Situated) VR environment and/or new highlighted images.

the next turn of the dialog. The example highlights challenges such as multimodal action prediction (*italics* above) and multimodal coreference resolution (underlined elements).

2 SIMMC Challenge Details

We briefly review the datasets, task definitions, and evaluation used in the SIMMC challenge. See Moon et al. (2020) for additional details.

Datasets. Two SIMMC datasets in the domain of interactive shopping have been provided: (1) Furniture and (2) Fashion. These datasets collectively contain about $13k$ human-to-human dialogs (totaling about $169k$ utterances). Moon et al. (2020) argue that shopping domains provide a dynamic environment, where rich multimodal interactions happen around visually grounded items.

Annotations. The SIMMC datasets are accompanied with the semantic-level annotation of utterances (dialog acts), multimodal state tracking, multimodal co-reference, actions and also ground truth semantic information about each scene. The latter allows training of virtual assistant models without the necessity of focusing on computer vision.

Tasks and Evaluation. There are three subtasks in the challenge with a priority list of metrics:

(Subtask 1) **Structural API Call Prediction** focuses on predicting the human-assistant action as an API call given the dialog and the multimodal contexts. Metrics for this subtask: action accuracy, action attribute accuracy, and action perplexity.

(Subtask 2) **Assistant Response Prediction** evaluates the relevance of the assistant response in the current turn; *(a)* as a conditional language model generation problem that uses BLEU-4 to score the similarity to the ground-truth response, and, *(b)* as a retrieval problem, where the goal is to retrieve ground-truth responses from a pool of 100 candidates (randomly chosen and unique to each turn). Priority metric list is mean reciprocal rank, recall@k ($k = \{1, 5, 10\}$), and mean rank.

(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 (Moon et al., 2020). The metrics for this subtask are slot and intent prediction F1, in line with prior work in DST.

3 Related Datasets and Challenges

Table 1 presents main distinctions of SIMMC compared to the the existing multimodal dialog datasets/challenges. The SIMMC dataset provides scenarios in which the situated multimodal context is dynamically updated, reflecting the agent actions. In the SIMMC settings, agent actions can be enacted on both the object-level – changing the view of a specific object within a scene, and the scene-level – introducing a new scene or an image. While the dialog-based image retrieval tasks (Guo et al., 2018; Saha et al., 2018) and the visual navigation tasks (Thomason et al., 2019; de Vries et al., 2018) do comprise context updates, they are limited to the introduction of new visual scenes, *e.g.*, new images or locations.

Compared with previous multimodal dialog datasets SIMMC offers four key advantages : (a) SIMMC assumes a co-observed multimodal context between a user and an assistant and records the ground-truth item appearance logs of each item that appears. (b) Compared with the conventional task-oriented conversational datasets, the agent actions in the SIMMC dataset span across a diverse mul-

Systems	Models	Eval.	Joint Train		Ens.	Pretrain Model	MM Rep.	Discrim. Train	Approx. Rank			
			subtasks	x-domain					sub1	sub2a	sub2b	sub3
Kung et al. (2021)	GPT-2 + FullCon.	1, 2a, 3	1, 2a, 3	yes	yes	GPT-2	stringified	·	4	5	·	5
	above + BLEU/METEOR	2b	1, 2a, 3	yes	yes	GPT-2	stringified	no	·	·	6 (7)	·
Kim et al. (2021)	MM Fusion Ens.A	1	1, 2a	no	yes	–	MAG/MMI	·	1	·	·	·
	MM Fusion Ens.B	2a	1, 2a	no	yes	–	MAG/MMI	·	·	7	·	·
	MM Fusion Ens.C	2b	1, 2a	no	yes	GPT-2	MAG/MMI	no	·	·	7 (8)	·
Jeong et al. (2021)	GPT-2 Ens.A	1	1, 2a, 3	no	yes	GPT-2	stringified	·	5	·	·	·
	GPT-2 Ens.B	2a, 3	2a, 3	no	yes	GPT-2	stringified	·	·	3	·	2
	GPT-2 Ens.C	2a, 3	2a, 3	no	yes	GPT-2	stringified	·	·	1	·	1
	GPT-2 Ens.D	2a, 3	2a, 3	no	yes	GPT-2	stringified	·	·	2	·	3
	B,C,D + cosine sim.	2b	2a, 3	no	yes	GPT-2	stringified	no	·	·	3-5 (4-6)	·
Huang et al. (2021)	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
	BART-L Bi-Encoder	2b	2b	no	no	BART	stringified	yes	·	·	1 (1)	·
	BART-L Poly-Encoder	2b	2b	no	no	adapted on 1, 2a, 3	stringified	yes	·	·	2 (2)	·
Senese et al. (2021)	BERT+log-likelihood	2b	2b	no	no	BERT	stringified	no	·	·	- (3)	·

Table 2: **Summary of the developed models**. Rank in parenthesis is for SIMMC-Fashion only.

```
System : This is our Hedon Kitchen Island with Stainless Steel Top. It
features a natural wood countertop. User : and what are the dimensions?
<SOM> OBJECT_0 : pos left color ['White'] class_name Kitchen Islands decor_style
['Rustic', 'Sophisticated'] OBJECT_1 : pos center color ['White'] class_name
Kitchen Islands decor_style ['Traditional', 'Modern'] <EOM> System : The width
is 52 inches, depth 18 inches, and height is 36 inches. User : and how much is it
```

Table 3: **Example of "stringified" multimodal context concatenated with user and system utterances.**

timodal action space (*e.g.*, *'rotate,' 'search,'* and *'add to cart'*). (c) Agent actions can be enacted on both the object level (*e.g.*, changing the view of a specific object within a scene) and the scene level (*e.g.*, introducing a new scene or an image). (d) SIMMC tasks emphasize semantic processing, while work in this area has traditionally focused heavily on raw image processing. The SIMMC annotation schema allows for a more systematic and structural approach for "visual" grounding of conversations, which is essential for solving challenging problems in real-world scenarios.

4 Survey of the Developed Systems

Table 2 provides a comparative summary of the 13 models that were developed by 5 different groups. As an example of how to read this table; Jeong et al. (2021) proposed four different ensembles (*Ens.*) of GPT-2 (Radford et al., 2019) models (*A, B, C, D*). Ens.A was evaluated (*Eval.*) only for subtask 1 but was jointly trained on three subtasks. Multimodal context was ingested by the model as a string of "word" tokens (*stringified*), *i.e.* formal descriptions of the scenes were flattened into a sequence of tokens and concatenated along with assistant and user utterances as shown in Table 3. Other ingestion approaches used specialized multimodal fusion (MM Fusion) gates; MAG (Rahman et al., 2020) and MMI (Yu et al., 2020). Ens.B, C and D were trained and evaluated on 2 substasks and adapted to the response retrieval task (2b) using cosine similarly over word vectors between the predicted response (2a) and candidate responses. Discriminative training (*Discrim. Train*) on subtask 2b was used only by Huang et al. (2021). *Approx. Rank* is the model rank using the top metric for each subtask without std. err considerations and is thus only indicative. We provide the detailed descriptions of each entry below.

Kung et al. (2021) proposed an ensemble of GPT-2 (Radford et al., 2019) models trained jointly on all three subtasks and across both domains. Specifically, they added a discriminative classifier consisting 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 such as segment embeddings were used as input to better leverage the visual information.

Kim et al. (2021) proposed an ensemble of models based on the baselines by Moon et al. (2020). While the baselines model subtask 1 and 2 jointly and subtask 3 separately, Kim et al. (2021) used

the predicted dialog state outputs from subtask 3 baseline as inputs for subtasks 1 and 2. Additionally, they used two sophisticated 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 1).

Jeong et al. (2021) proposed a varied set of ensembles of GPT-2 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 evaluated 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 its results were only evaluated for subtask 1. In the ensemble developed 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 compared 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.

With reference to Table 2; (a) *Ens.A* by Jeong et al. (2021) consists of a medium and small GPT-2 model, both trained on the train and dev sets, (b) *Ens.B* is two large GPT-2 models, one trained on just the training set and other trained on both train and dev sets, (c) *Ens.C* is a large and small GPT-2 model, both trained on the train and dev sets, and, (d) *Ens.D* is two large and one small GPT-2 model, where all but one large model were trained on train and dev sets, while the large model was trained on just the training set.

Huang et al. (2021) proposed two BART (Lewis et al., 2020) models (BART-Large and BART-Base) for subtasks 1, 2a, and 3. Both were trained to jointly predict 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 proposed 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. These model weights are then further adapted. Four model combinations exist for this subtask (2b), *i.e.*, BART-Large or BART-Base with Bi-encoder or Poly-encoder, but Table 2 only includes results for BART-Large Bi/Poly-encoders.

Senese et al. (2021) proposed a BERT-based model addressing the Assistant response retrieval task (subtask 2b), trained using the cross-entropy loss. Specifically, the proposed model includes a self-attention module, an encoder-decoder attention module, and an item-attention module. The item-attention module (part of the decoder) computes attention over the states of a transformer which encodes the attributes of the reference item, *e.g.* the shared item in the scene. 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. The latter score rewards responses that mention item attributes that appear in the reference item. Candidate responses with the highest sum of these two scores were used as final predictions.

5 Performance Analysis

5.1 Summary

The developed models set a new state-of-the-art in all three subtasks. Table 4 summarizes their performance. For the structural API call prediction subtask (subtask 1), the BART-Large model by Huang et al. (2021) achieved the best overall performance (taking into account both API and attribute accuracy). This model also achieved the second-best performance on subtask 2a, and on subtask

Systems	Subtask 1. API Prediction			Subtask 2. Response Generation						Subtask 3. DST	
	Acc↑	A.Acc↑	Perp↓	BLEU↑	MRR↑	r@1↑	r@5↑	r@10↑	Mean↓	Slot F1↑	Intent F1↑
Baseline (Moon et al., 2020)	79.3	63.7	1.9	0.061	0.145	7.2	19.8	27.3	39.2	62.4	62.1
Kung et al. (2021)	80.2	74.6	2.0	0.105	0.326	21.1	43.6	56.8	18.8	77.8	76.7
Kim et al. (2021)	**82.5**	69.8	1.8	0.082	0.074	2.5	8.3	13.6	47.7	-	-
Jeong et al. (2021)	79.4	73.2	-	**0.128**	0.381	26.3	50.3	61.8	15.5	**79.1**	78.1
Huang et al. (2021)	**81.3**	**73.9**	3.5	0.108	**0.673**	52.6	87.4	95.1	3.2	78.6	77.7
Senese et al. (2021)*	-	-	-	-	0.390	26.7	52.1	66.0	14.8	-	-

Table 4: Summary of the results on Test-Std split, average of Furniture and Fashion (*Senese et al. (2021) submitted results only for Fashion). Best results from each system are shown. **(1) API prediction** via Accuracy, Perplexity and Attribute Accuracy, and, **(2) Response Generation** 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.

3. For the response retrieval subtask (subtask 2b), the BART-Large Bi-encoder model by Huang et al. (2021) achieved the best performance. For the response generation (subtask 2a) and DST subtasks (subtask 3), the GPT-2 model ensemble by Jeong et al. (2021) achieved the best performance.

5.2 Subtask 1: Structural API Call Prediction

Figure 2 shows the breakdown of action accuracy by type for both datasets. The key observations are:

- All systems successfully predict `AddToCart` and `SpecifyInfo` with 90% and 95% accuracy respectively, for both the domains. Intuitively, 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?"* 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 product information.
- 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 of 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`).
- Note that Huang et al. (2021) (winner) and Kim et al. (2021) (runner-up) perform similarly on the API call prediction task with an overall accuracy of 81.3% and 82.5% respectively (Table 4). The winner was declared based on the action attribute accuracy.

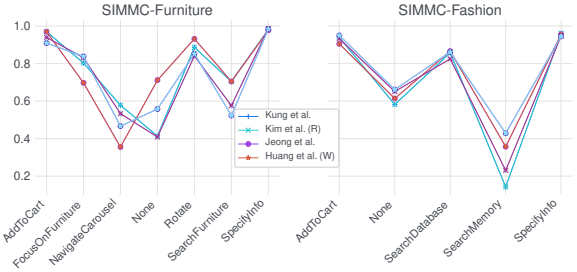


Figure 2: Breakdown of the API Call Prediction accuracy (subtask 1) according to actions.

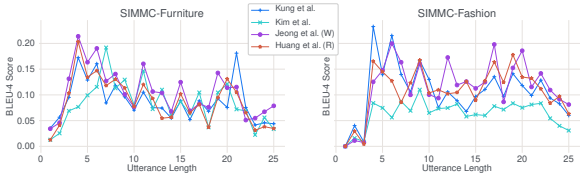
5.3 Subtask 2: Assistant Response Generation

We compare BLEU-4 scores (generation category) based on: (a) length of ground-truth assistant utterance in Figure 3a, and (b) corresponding ground-truth API call in Figure 3b. Following are the takeaways:

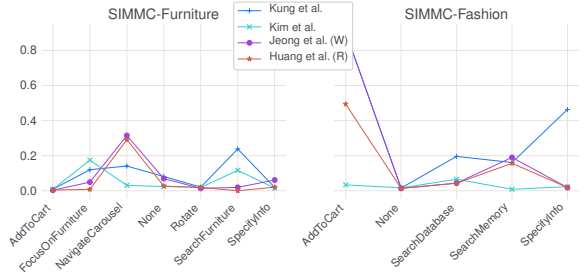
- As expected, BLEU-4 score decreases (on average) with the length of the utterances.
- Though the smoothing for BLEU-4 contributes partially to the low values for utterance lengths of 1–3, a good proportion of these utterances contained information about the catalog item, e.g., price and dimension. 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 `AddToCart`, models perform better on SIMMC-Fashion on 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.
- BLEU-4 for `SpecifyInfo` is lower than the overall score for all models, points to the need

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

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.



(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 entries for Assistant Response Generation (Subtask 2). See text for more details.

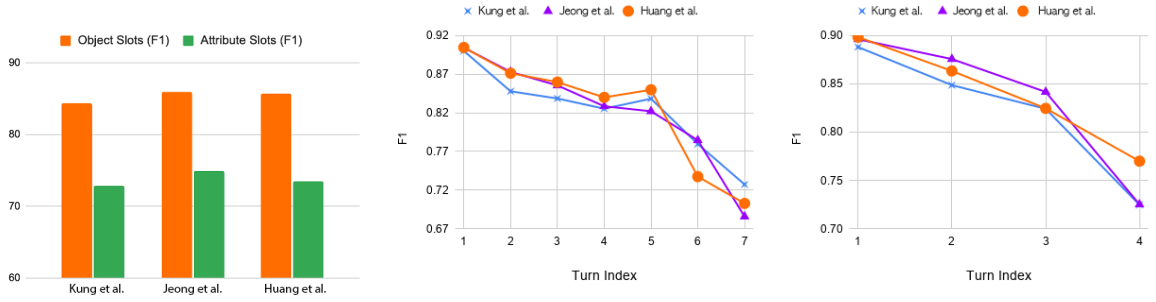
for a better catalog modeling again.

Interestingly, Huang et al. (2021) (the best model for subtask 2) used *discriminative* training for this subtask to achieve superior performance (26 points lead on the r@1). Specifically, they train not only to increase the likelihood of ground-truth response (similar to a language model) but also to decrease that of other response targets in the batch that act as negative examples. This enables the model to discriminatively pick the ground truth over the other distractor candidates. Das et al. (2017) also observe a similar phenomenon.

5.4 Subtask 3: Dialog State Tracking (DST)

Figure 4a shows a breakdown of the DST results based on slot types. 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. F1 scores for attribute slots have higher variances across different entries compared to those for object slots. This shows that the different approaches proposed by each system had relatively small influences on the multimodal coreference resolution performance.

Figure 4b and Figure 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 base-



(a) F1 per slot types (Average) (b) Object F1 for SIMMC-Furniture (c) Object F1 for SIMMC-Fashion

Figure 4: Analysis for Dialog State Tracking (Subtask 3). (a) Breakdown of Slot F1 results by slot types (object & attribute slots). (b, c) Average object slot tracking results at varying turn indices. See text for more details.

Model	Subtask 1. API Prediction			Subtask 2. Response Generation	Subtask 3. DST	
	Acc↑	A.Acc↑	Perp↓	BLEU↑	Slot F1↑	Intent F1↑
Original (Huang et al., 2021)	79.6	**79.5**	5.9	0.099	**61.3**	62.6
multimodal-context-ablated	79.2	78.3	5.9	0.098	55.7	63.2

Table 6: Summary of multimodal-context-ablation results on Dev-Std split, average of Furniture and Fashion. **(1) API prediction** via accuracy, perplexity and attribute accuracy, and, **(2) Response Generation** via BLEU, **(3) Dialog State Tracking (DST)**, via slot and intent prediction F1. ↑: higher is better, ↓: lower is better.

lines in suppressing the degradation in the object slot predictions over time.

5.5 Breakdown based on "all" and "none"

We identify instances on which **all** and **none** of the developed models were able to accurately predict the ground-truth API call. We breakdown each of these instance categories further into the ground-truth actions in Figure 5. 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.

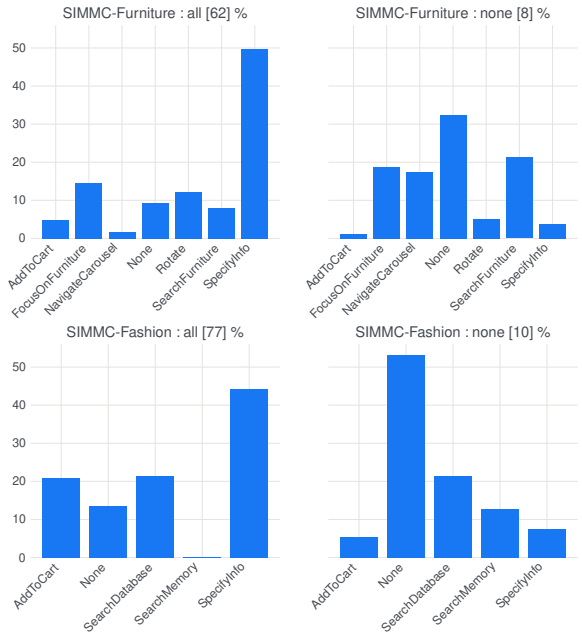


Figure 5: Breakdown of instances categorized based on whether **all** or **none** of the model entries predicted accurately.

6 Ablation Study

To further test the extent to which the available multimodal context is improving model results on the subtask metrics, we conduct an ablation experiment where we prepare a version of the datasets with the multimodal context removed. We then train and test the BART-Large model (Huang et al., 2021) on the original and ablated versions of the datasets.

6.1 Methodology

For model training, we conduct a parameter search over batch size and learning rate, and train three models for each combination of parameters. We select models that achieved the lowest dev set loss during training. We repeat this process for the four combinations of SIMMC-Furniture or SIMMC-Fashion with original or multimodal-context-ablation versions of the dataset. The aim is to ensure that the models trained on the ablated datasets are trained and selected under the same conditions as the models that have the multimodal context available. Note that this process does not guarantee to reproduce the reported results for this model.

6.2 Results

Results are presented in Table 6. Multimodal context does boost performance on slot F1 metric in subtask 3 (DST) in line with findings by Moon et al. (2020). It also provides a marginal improvement in attribute accuracy in subtask 1 (API calls). Other metrics like BLEU are largely unmoved. Given that the multimodal context should inform the assistant's responses, this is somewhat surprising.

7 Findings & Conclusions

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 their flexibility to ingest relatively simple multimodal context and thus be used in a multimodal setting with a high degree of success.

Multimodal context helps but gaps remain. To examine how effectively models use the multimodal context we conduct an ablation experiment where we train the BART-Large-based model (Huang et al., 2021) on two versions of the datasets; including and excluding multimodal context. The results (Table 6) indicate that multimodal context does boost performance on slot F1 metric in subtask 3 (DST) and provides a marginal improvement in attribute accuracy in subtask 1 (API calls). However BLEU scores for response generation (subtask 2a) are relatively unaffected. In SIMMC-Furniture, the multimodal context provides, for each turn, a grounded set of items which are likely to be the most salient. Given this, the ablation results when considered alongside both the overall relatively low BLEU scores, and the accuracy falloff in DST metrics with increasing dialog length, suggests that the multimodal context isn't currently utilized to the fullest extent and indicates that there remains a significant opportunity for improving assistant response prediction.

Need for a better and scalable catalog integration. Generated responses (see Table 5) indicate 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 knowledge in the catalog or a more general problem with utilisation of multimodal context in response generation.

Approaches that may address this issue include: encoding additional information from the catalog, such as price and description, for each item in the scene; integrating explicit database API calls to the catalog and database responses as part of prediction task and model input respectively (*c.f.* Peng et al. (2020); Hosseini-Asl et al. (2020)); discourage memorization of the catalog by randomly varying attributes, such as price, (while maintaining consistency in the data between model input and target); extending the test set with examples drawn from a held out catalog to penalize memorization.

Better and scalable multimodal integration for knowledge bases, *e.g.* catalogs, 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 increasingly complex scenarios. As the number of items in the scene increases, so does the string representation making it harder for the model to capture scene related information due to increased nesting.

References

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chulaka Gunasekara, Abhinav Rastogi, Yun-Nung Chen, Luis Fernando D'Haro, Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkanitur, Baolin Peng, Jianfeng Gao, Jinchao Li, Lars Liden, Minlie Huang, Qi Zhu, Runze Liang, Ryuichi Takanobu, Shahin Shayandeh, Swadheen Shukla, Zheng Zhang, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon (EJ) Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the Ninth Dialog System Technology Challenge: DSTC9. *arXiv preprint arXiv:2011.06486*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *NeurIPS*.

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*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 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.

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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*.

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*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. *arXiv preprint arXiv:1907.04957*.

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*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 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*, pages 3342–3352, Online. Association for Computational Linguistics.