Overview of Situated and Interactive Multimodal Conversations (SIMMC) 2.1 Track at DSTC 11

Satwik Kottur[1]*, **Seungwhan Moon** [2]*, **Alborz Geramifard** [1], **Babak Damavandi** [2]
Meta Reality Labs & Meta AI
{skottur,shanemoon,alborzg,babakd}@meta.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.

In order to push research towards this next generation virtual assistants, the SIMMC 2.1 challenge[1] was conducted at the Eleventh Dialog System Technology Challenge (DSTC) which had entries from across the world competing to achieve the state-of-the-art performance in the SIMMC 2.1 task. In this report, we present and compare 13 SIMMC 2.1 model entries from 5 teams across the world to understand the current progress made across the last three years (starting with SIMMC 1.0 and 2.0 challenges) 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.

1 Motivation

The Situated and Interactive Multimodal Conversaitional AI (SIMMC) challenges [15, 12], held as part of DSTC9 [8] and DSTC10, pioneered the work for building the real-world assistant agents that can understand multimodal inputs (vision & conversations) and handle user requests. Throughout the two editions of the challenge, we provided two new benchmark datasets (SIMMC 1.0 and 2.0) for studying multimodal conversations with situated user context in the form of a co-observed image or virtual reality (VR) environment. Specifically, the SIMMC 2.0 dataset provided the assistant↔user task-oriented dialogs grounded on diverse photo-realistic VR renders of (synthetic) commercial stores with various referent objects, serving as a proxy for complex real-world scenarios. The earlier SIMMC challenge at DSTC10 saw a total of 16 model entries from participants around the world, establishing a new set of state-of-the-art baselines for the multimodal task-oriented dialog systems.

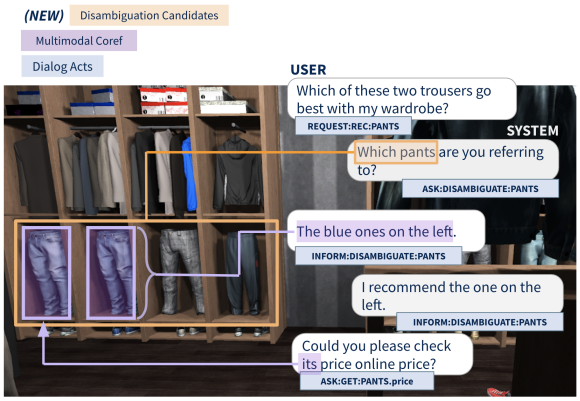

Figure 1: Illustration of a Situated Interactive Multimodal Conversation (SIMMC), which presents a task-oriented user↔assistant dialog grounded in a co-observed photorealistic multimodal context. The new version of the SIMMC dataset includes more fine-grained and precise annotations for referent disambiguation candidates, which poses new challenges for the Multimodal Coreference Resolution task (MM-Coref) and the Ambiguous Candidate Identification task.

While the new SOTA models have drastically improved the performance on the previous benchmark tasks, several challenges remain in building the production-ready agent. One such challenge is the visual disambiguation which is often encountered in real-world multimodal conversations. For instance, the user ambiguously uses *'these two trousers'* in Fig. 1, which cannot be deterministically resolved. Thus, an assistant system needs to reason about this, identify the possible ambiguous candidates, and then ask a follow-up question to disambiguate. This is important to avoid

* Joint first authors

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

Proceedings of The Eleventh Dialog System Technology Challenge, pages 235–241

a wrong resolution of such references leading to subsequent turns with false premises. Another advantage is that the assistant could insert implicit confirmation – "*Which of the two red jackets are you referring to - the one on the left or the one closer to you?*", as opposed to a generic response like: "*Which one are you referring to?*".

To this end, we conducted the third edition of the SIMMC challenge (SIMMC 2.1) for the community to tackle and continue the effort towards building a successful multimodal assistant agent. In this edition of the challenge, we specifically focused on the key challenge of fine-grained visual disambiguation, which adds an important skill to assistant agents studied in the previous SIMMC challenge. To accommodate for this challenge, we provided the improved version of the of the dataset, SIMMC 2.1, where we augment the SIMMC 2.0 dataset with additional annotations (*i.e.* identification of all possible referent candidates given ambiguous mentions) and corresponding re-paraphrases to support the study and modeling of visual disambiguation (SIMMC 2.1). This new version of the dataset poses several interesting challenges such multimodal dialog state tracking given ambiguity, coreference resolutions ("*directly behind it*", "*grey jacket to the right of the one I mentioned*"), and disambiguation strategies ("*How much is that shirt*" → "*Which shirt are you referring to?*").

2 The Third SIMMC Challenge

We now detail the new version of the multimodal conversational dataset (SIMMC 2.1) (Sec. 2.2) and the four main sub-tasks (Tab. 1, Sec. 2.3).

2.1 Problem Setup

The SIMMC challenge studies the conversational scenarios where the virtual assistant shares a co-observed scene with a user. Specifically, the dataset targets the shopping experience as the domain of study, which often induces rich multimodal interactions around browsing visually grounded items in a physical store (fashion or furniture). The assistant agent is assumed to have access to the ground-truth meta information of every object in the scene, while users observe those objects only through the visual modality to describe and compose a request, as in the real-world applications. Each dialog in the dataset includes multiple viewpoints at different time steps throughout the session, corresponding to the scenarios where users are physically navigating the scene during the conversation. Therefore, the conversational models for the SIMMC challenge need to understand both user requests using both the dialog history and the state of the environment as multimodal context.

Note that the SIMMC problem setup where user and assistant co-observe the same scene allows for more natural multimodal coreferences to be used as part of user-assistant conversations. The previous literature in multimodal dialogs [1, 10, 4, 13, 6, 5] often assumes that dialog participants take the roles of a primary and secondary observer respectively, *i.e. questioner* and *answerer* similar to the Visual Question Answering [2] tasks, which does not address the real-world consumer scenario we are targetting. The SIMMC challenge also extends many of the key dialog tasks studied in the previous literature on conventional task-oriented dialog systems [9, 18, 3, 7] (*e.g.* DST, slot carryovers) to the unique multimodal settings.

2.2 The New SIMMC 2.1 Dataset

The SIMMC 2.1 dataset extends the original SIMMC 2.0 dataset [12] with additional annotations for fine-grained referent candidates and new utterances (details below).

Collection Process: The original dataset used the two-phase pipeline to collect dialogs (multimodal dialog simulation & manual human paraphrase), which can effectively collect natrual dialogs with the minimum annotation overheads. Note that this approach extends the popular machine↔human collaborative dialog collection approaches [18, 19] to the unique multimodal settings. More details on the data collection approach can be found in [12].

Data Statistics: The dataset includes 11,244 dialogs (117,236 utterances), with the fine-grained ground-truth dialog labels (NLU/NLG/Coref) already in place. Table 2 shows the statistics of the dataset.

Additional Annotations in SIMMC 2.1. The key differences are illustrated in Fig. 1. We collect candidates in the scene for turns with ambiguous references using human annotators. Further, we also ask the annotators to paraphrase the turn in case there is not enough ambiguity. This data collection process thus allows for richer coreferences, referential expressions, and disambiguation scenarios.

Task Name	Goal	Evaluation
1. Ambiguous Candidate Identification	Given user utterances with ambiguous object mentions, resolve all referent candidate objects to their canonical ID(s) as defined by the catalog.	Object identification Precision / Recall / F1
2. Multimodal Coreference Resolution	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	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-4 score

Table 1: **Proposed tracks and descriptions.**

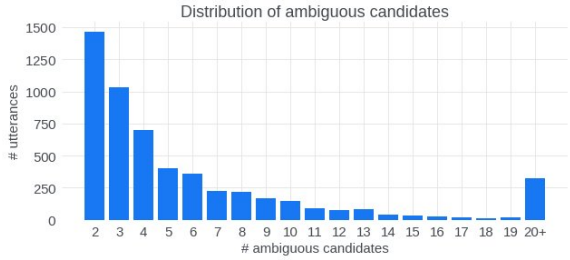

Figure 2: Distribution of ambiguous items.

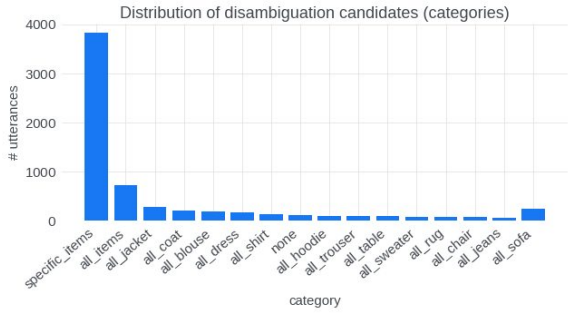

Figure 3: Distribution of ambiguous items per category.

Total # dialogs	11,244
Total # utterances	117,236
Total # scenes	3,133
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
Avg # candidates per ambiguous turn	5.6

Table 2: **SIMMC 2.1 Dataset Statistics**

Annotation Analysis: We have $6.5k$ turns with ambiguous candidates, where the average number of candidates is 5.6. Fig. 2 and Fig. 3 represent the distribution of the candidates in these utterances per the number of items or per category.

Data format: The SIMMC 2.1 data has been provided in the same format as the earlier version of the datasets, making it easier for participants to use the various benchmark models publicly available [15, 12], or augment it with the previous version of the dataset for pretraining, etc. The raw pixel images of each scene as well as the pre-computed visual embeddings have been provided, allowing for easier adaptation for the NLP audience.

2.3 Challenge Tasks

We invited the DSTC community to build multimodal conversational agents for the following four benchmark tasks, addressing the key challenges in the multimodal conversational reasoning (summarized in Tab. 1). All the benchmark tasks require a strong computer vision capability as well as a multimodal conversational reasoning capability, to jointly process both the dialog and the visual contexts.

2.3.1 Ambiguous Candidates Identification

As a main focus of this edition of the challenge, this sub-task will evaluate the models' performance on identifying *all* candidate objects referred by a given user utterance, as their canonical object IDs as defined for each scene (*e.g.* U: "*How much is that blue shirt on the hanger?*" → (object IDs of all blue shirts on hangers in the scene).

The task will provide the ground-truth bounding boxes defining each object ID to make evaluation easier. The performance will mainly be measured for its F1 score.

2.3.2 Multimodal Coreference Resolution (MM-Coref)

The goal of this task is to resolve referential mentions in user utterances to their canonical object IDs as defined for each scene. The resolving contexts can come either through (1) the dialog context (*e.g.* A: "*This shirt comes in XL and is $29.*" → U: "*Please add it to cart.*", (2) the multimodal context (*e.g.* U: "*How much is that red shirt back there?*"), or (3) both (*e.g.* U: "*How much is the one next to the one you mentioned?*").

2.3.3 Multimodal Dialog State Tracking (MM-DST)

Following the earlier challenge, the goal of the MM-DST task is to predict slots and their corresponding values grounded on the co-existing multimodal context. This requires tracking the states of multimodal objects (in addition to textual tokens) as part of dialog states. Note that this task extends the traditional notion of the unimodal dialog state tracking (DST) problem widely studied by the DSTC community.

2.3.4 Assistant Response Generation

The goal of the task is to generate Assistant responses given user utterances, ground-truth APIs and object IDs. While the assistant agent has the ground-truth meta information on each object, the referent objects need to be described *as observed and understood* by the user through the co-observed scene or the dialog context, adding an interesting challenge to the traditional response generation tasks (*e.g.* `INFORM:RECOMMEND (OBJ_ID: 3)` → A: "*I recommend the blue jacket directly behind the one I mentioned*".

In addition, with the new annotations and utterances added for SIMMC 2.1, we expect that the entries that can leverage the identified ambiguous candidate list as part of the response will achieve the best performance (*e.g.* A: "*Which of the two red jackets are you referring to - the one on the left or the one closer to you?*", as opposed to a generic response like: "*Which one are you referring to?*") – which is the main focus of this edition of the challenge.

3 Baselines

There are two baselines, adopted from [12]:

(a) **MM-DST model** by [15], where we train a multi-task GPT-2 [17] 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 [16, 11]. 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)** [14] 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 [14].

Specifically for the task of ambiguous candidate identification (subtask 1), we extract the feature representation of the last token as the textual feature and compute similarity with the visual features extracted for each item in a given scene. Visual features for the items are extracted similar to the above MTN network.

4 Submitted Systems

4.1 Models

We now provide brief description of the entries submitted to the SIMMC 2.1 challenge.

Team 1 uses a large transformer-encoder based discriminative model (Longformer). To predict ambiguous candidates, coreference resolution, and belief state, they encode dialogue history and at-

Team	Model	1. ACI.	2. MM-Coref	3. MM-DST		4. Gen.
		ID F1↑	Coref F1↑	Slot F1↑	Intent F1↑	BLEU↑
Baseline	GPT-2 [15]	42.6	30.5	71.8	92.6	0.199
	MTN [14]	·	·	66.5	91.5	0.210
Team 1	Longformer	67.3	94.3	94.2	96.0	·
	FairSeq-Generative	·	·	·	·	0.409
Team 2	CoCondenser+ViT	65.2	·	·	·	·
Team 3	ALBEF	63.8	75.6	·	·	·
	BART	·	·	90.5	96.9	0.303
Team 4	Combiner	·	·	·	·	0.252
Team 5	BART-DSTCLA-Ens.	69.9	80.1	92.6	97.3	·
	BART-DSTCLA	70.5	80.3	92.7	97.3	·
	BART-Ens.	70.3	79.9	91.6	97.8	·
	BART-Sys.	·	·	·	·	0.351
	BART-Sys-Alt.	·	·	·	·	0.353
	BART-Sys-Ens.	·	·	·	·	0.365
	BART-Sys-Attr.	·	·	·	·	0.350

Table 3: Summary of the results on Test-Std split. Best results from each system are shown. **(1) Ambiguous Candidate Identification (ACI.)**, via identification F1, **(2) Multimodal Coreference Resolution (MM-Coref)**, via coref prediction F1, **(3) Dialog State Tracking (DST)**, via slot and intent F1, **(4) Response Generation (Gen.)** via BLEU. ↑: higher is better. Baseline performances: [15] (top), [14] (bottom).

tach task-specific heads to the output of encoder. Additionally, they line up the item vectors (bbox position embedding) with their respective attribute tokens in the inputted token embedding. Auxiliary heads for predicting attributes are added as additional supervision signals. For task 4, a generative multimodal model is used which takes dialogue history and non-visual attributes as textual input, and corresponding scene images as visual input. The model then generates system response autoregressively.

Team 2 proposes a model based on CoCondenser and Vision Transformer (ViT) specifically trained for Subtask 1, leveraging their powerful encoding ability. Image models were fine-tuned for the in-domain image sets (for fashion and furniture domains) to improve the visual encoding ability of the models.

Team 3 proposes separate models tuned for each subtask. For subtask 1, they train a vision language transformer model using ALBEF as backbone. Similarly, they use the same architecture for subtask 2, trained separately using only the co-reference labels as the target task. For subtask 3, they use additional pre-training on the BART model with a tailored masked language modeling task to adapt to the dialogue domain, later fine-tuned on the Dialogue State Tracking task. They use a similar model fur sub-task 4 as well, framed as a generative model. Object images are pre-processed using a ResNext model.

Team 4 did not release their implementation, thus details about their approach are not public.

Team 5 performs ablation over multiple families of models – a vanilla Encoder-Decoder architecture (BART, T5, UL-2, and BlenderBot), an Multi-Modal encoder architecture (Flava), and a Multi-Modal Encoder-Decoder, all jointly trained for a multi-task (MT) objective. With various models, they perform ablations over using the joint input of textual context, system dialog acts, tokenized objects and scene as multi-modal input.

4.2 Evaluation

The entries to the challenge set a new state-of-the-art in all four subtasks. The results are summarized in Tab. 3.

The winner of the ambiguous candidate identification subtask (subtask 1) was the multi-task BART-based model from Team 1. The winner of the multimodal coreference resolution task (subtask 2), the dialog state tracking task (subtask 3), and the response generation task (subtask 4) was from the Longformer model from Team 5.

For Subtask 1, the BART-based multi-task train-

ing from Team 1 outperforms other competing models by more than 2 point absolute gain. This result implies that the ambiguous candidate identification task can benefit from the learning signals from multiple sub-tasks.

For Subtask 2 through 4, the Longformer-based model from Team 5 shows a huge margin in performance over other models, indicating the effectiveness of the implementation in encoding the visual information (*e.g.* bbox position embeddings, understanding visual attributes of items).

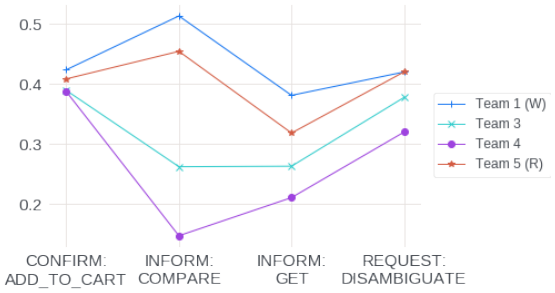

Figure 4: Distribution of BLEU score for subtask 4 over natural language generation act.

Fig. 4 shows the distribution of BLEU scores over natural language generation act, with the following key observations:

- All teams seem to perform comparable on `CONFIRM:ADD_TO_CART` intent, which is to be expected as it usually is less diverse linguistically.
- `INFORM:COMPARE` act clearly distinguishes both the winner and runner-up team from the rest, as this requires the system to understand the objects of comparison and list their attributes for the user.

5 Conclusion

The goal of the SIMMC 2.1 challenge in DSTC11 is to motivate and inspire the research community to work on the problem of creating dialog agents that can handle multiple modes of communication and handle tasks in a specific context. These agents have many practical uses and present a range of challenges related to multimodal communication.

The SIMMC 2.1 challenge specifically saw a number of submissions exploring different transformer architectures, highlighting the trade-offs in performance. We hope that the insights gained from this challenge will shed light on the challenges of multimodal dialogs and inspire further research in this area.

What next? We identify possible future directions for the SIMMC challenge to continue advancing in this area of research.

- **Incorporating additional modalities.** The current approach for the SIMMC 2.1 challenge utilizes screenshots from a shopping website as the multimodal context. While this presents its own set of challenges, it does not take into account additional forms of input such as eye gaze, head position, and gestures that would be present in a real-world setting. These additional cues are often used by human users to refer to objects, for example, "*How much is that shirt* (pointing a finger)". To include these modalities, SIMMC would need to be grounded in a virtual environment with a stream of inputs to capture users' eye gaze, head position, location in the store, *etc*.
- **Incorporating temporal reasoning.** Many of the referring expressions used in real-world scenarios are *temporally* grounded, which is not studied in the current challenge in detail. For instance, to understand a query "*I prefer the one I saw when I entered the store*.", an assistant agent would need to have a way to track the multimodal states (including human motions and movements) over time.

References

[1] H. Al Amri, V. Cartillier, R. G. Lopes, A. Das, J. Wang, I. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, and C. Hori. Audio visual scene-aware dialog (avsd) challenge at dstc7. 2018.

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

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

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

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

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

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

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

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

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

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

[12] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 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*, pages 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

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

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

[15] 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. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*, 2020.

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

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

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

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