Abstract
Owing to the success of deep learning techniques for tasks such as Q/A and text-based dialog, there is an increasing demand for AI agents in several domains such as retail, travel, entertainment, etc. that can carry on multimodal conversations with humans employing both text and images within a dialog seamlessly. However, deep learning research in this area has been limited primarily due to the lack of availability of large-scale, open conversation datasets. To overcome this bottleneck, in this paper we introduce the task of multimodal, domain-aware conversations, and propose the MMD benchmark dataset towards this task. This dataset was gathered by working in close coordination with large number of domain experts in the retail domain and consists of over 150K conversation sessions between shoppers and sales agents. With this dataset, we propose 5 new sub-tasks for multimodal conversations along with their evaluation methodology. We also propose two novel multimodal deep learning models in the encode-attend-decode paradigm and demonstrate their performance on two of the sub-tasks, namely text response generation and best image response selection. These experiments serve to establish baseline performance numbers and open new directions of research for each of these sub-tasks.

1 Introduction
In the area of conversation systems, the demand for autonomous AI agents that can assist humans with complex activities requiring a long series of interactions (such as shopping on the web or mobile) has grown by leaps and bounds in several domains such as retail, travel, entertainment, etc. which have witnessed an explosion of multimodal content on the internet. Such AI agents need to demonstrate both of the following capabilities: (i) the ability to converse with humans seamlessly using multimodal content, (ii) ability to exploit domain-specific knowledge.

The recent progress with deep learning techniques for other problems at the intersection of NLP and Computer Vision such as image captioning (Vinyals et al., 2015; Xu et al., 2015), video description (Yu et al., 2016), image question answering (Antol et al., 2015), video question answering (Zeng et al., 2016; Maharaj et al., 2016), is owed in large part to the availability of large-scale open datasets for their respective tasks. However, the primary hindrance for deep learning research towards multimodal conversation systems has been the lack of large-scale open datasets that exemplify both of the challenges mentioned above. Though there has been recent work (Serban et al., 2016b; Yao et al., 2016; Serban et al., 2016a) with different conversation datasets (Lowe et al., 2015; Vinyals and Le, 2015; Ritter et al., 2010), the mode of interaction there is limited to text conversations only, rendering them inadequate for multimodal conversation research. While such multimodal, human-to-human conversation transcripts (for example, between shoppers and retail salespersons) might be available in industry settings, they are both limited in scale and are proprietary, hindering open deep learning research.

In this paper, we attempt to alleviate both of these challenges by developing a large-scale multi-modal conversational dataset in the retail domain that embodies the required generic capabilities for such autonomous agents. Since the actual transcripts are both limited and proprietary, we conducted a series of interviews with a large number of retail salespersons in the fashion domain and devel-
oped the dataset interactively with them in a semi-automatic manner at a large scale as follows. The domain experts described in detail various phases of the sales process which were materialized into over 80 states for a conversational agent. Each such state had a specific intent (such as a shopper wanting to see more items similar to a specific one identified, shopper asking for a style tip or about the latest trends being endorsed by celebrities, etc). Corresponding to each such intent, a collection of representative utterance patterns involving both text and images were designed by us along with the experts. Each such state exhibited multi-modality (i.e. involving text and images) in both the utterance/response as well as the dialog context. The domain experts described various possible sales flows where customer went from one state to another, which were captured by transitions of an expert model automata between these states. The experts then inspected the outputs of controlled runs of the automata, provided feedback, which was thereafter employed to further refine the expert automata. Proceeding in this manually-intensive and iterative manner under the supervision of domain experts, we produced a large-scale dataset of over 150K multimodal dialogs capturing a wide variety of conversational sessions exhibiting the challenges described earlier. Note that such a data collection could not have been achieved with typical crowdsourcing platforms such as Amazon Mechanical Turk using ordinary crowd workers since it was necessary to be performed under the supervision of fashion sales experts exhibiting domain-specific understanding.

An example of a typical multimodal conversation involving both text and images, and exhibiting domain knowledge is shown in Figure 16. Notice that the response generated by the system can either be text (for example, see the 3rd turn inside the highlighted portion in red) or a set of images (as seen in 1st turn there). Similarly, at every point, the context available to the system is multimodal which the system has to process to generate a coherent response. For example, to give a relevant response in Turn 2 inside the box the agent needs to first pay attention to the words “espadrille”, “silver” and the sentiment which are mentioned in the text and then understand the style (“strapped ankle”) from the image to fetch images for “silver espadrille with strapped ankle” in response to the given context.

The body of work most relevant to ours is Visual QA(Antol et al., 2015) involving a single question and response, and the work of (de Vries et al., 2016; Mostafazadeh et al., 2017; Das et al., 2016) involving a sequence of QA pairs with an image forming a dialog. There are a few key differences between these datasets and our work as demonstrated by the example above. First, in their case, all questions in a sequence pertain only to a single image. Secondly, their responses are always textual. However, as is the case with natural conversations amongst humans, in our work, (i) there could be multiple images providing context, (ii) these context images could change across turns during the course of the conversation, and (iii) the response at each step can be text or image(s) or both.

Finally, in this paper, we propose a baseline framework to model the agent’s responses in such multimodal conversations. In particular, we propose hierarchical attention-based models for the textual and image response as two separate tasks and empirically estimate the feasibility of these tasks. We also discuss limitations that open new directions for research for these and multiple other tasks enabled by this new dataset. The main contributions of this work can be summarized as follows:

- We introduce the task of Multimodal Conversation which is significantly distinct from the sequential Visual QA driven dialog tasks mentioned before.
- We introduce a large dataset for this task and define several research tasks that can be evaluated using this dataset
- We propose baseline multimodal encoder decoder models for two such tasks and define appropriate metrics for evaluating these tasks

2 The Multimodal Dialogs (MMD) Dataset

As mentioned in the previous section, a key contribution of this paper is a large-scale dataset of 2-party dialogs that seamlessly employ multimodal data in their utterances and context and also demonstrate domain-specific knowledge in their series of interactions. Toward this goal, in this section, we first describe the methodology employed for collecting this dataset and then explain in detail the various sub-tasks exhibited by the dataset that open up new research problems.

2.1 Data Collection Methodology

The data collection process was performed by us in close coordination with a team of 20 fashion experts and primarily consisted of two steps, (i)
curation and representation of a large-scale domain knowledge, and (ii) developing a large collection of multimodal conversations, each consisting of a series of interactions employing this knowledge. We next proceed to describe these 2 steps in detail.

2.1.1 Domain Knowledge Curation

Through our series of interviews with the domain experts, we observed that a lot of the complexity in a natural conversation in this domain comes from the background knowledge that both the expert agent and the shopper employ in their conversation. The expert’s domain knowledge is multitude in nature, varying from knowledge about which attire goes well with which accessory, to which celebrity is presently endorsing which kind of fashion items, or what kind of look is better suited for which occasion. Therefore, the first step in our data collection process was to curate this domain knowledge from unstructured multimodal content on the web at scale and represent them in a machine consumable manner. This process involved a series of steps as enumerated below:

1. Crawling over 1 Million fashion items from the web along with their available semi/un-structured information and associated image(s)

2. Parsing different types of domain knowledge from the free-text information, and curating them in a structured form after a round of manual inspection by domain experts.

2a. Creating a hand-crafted taxonomy of the different types of fashion items. For example, man > apparel > layer-2-lower-body > trouser > formal-trousers, dressed pants, suit pants i.e. formal-trousers is synonymous to dressed pants or suit pants and is a type of trouser which is again a type of layer-2-lower-body apparel. Each entry in the taxonomy has a synonym-set (called “synset”). 282 such fashion “synsets” were collected from domain experts for men and 434 for women.

2b. Identifying the set of fashion attributes relevant (especially for the purpose of shopping) to each of the fashion synsets. Overall 52 such attributes (like color, pattern, style, price, wash-care information) were identified by domain experts, where 45 of them are visual attributes and remaining are metadata attributed about the synset (e.g. wash-care information, price, seller ranking).

2c. Seeding the lexicon for each of these attributes with a set of realistic values provided by the domain experts.

More examples of dialogue sessions are provided in the supplementary material.
Examples of color, pattern, material, brand, style, dial

| Product | Description |
|---------|-------------|
| 4200    | shirt & trouser, tshirts & sneakers, tuxedo, size, wash-care, |
| 1.05M   | 45           |
| 411     | shirt, trouser, tuxedo, style, neck, sleeves, length, sole-type, closing, |
| 8       | price, wash-care, product ranking, brand, size, suitable occasions, |

| Domain Specific Knowledge Base Stats | Examples |
|-------------------------------------|----------|
| Mean Images per item                | 4        |
| Avg. Fashion Synsets                | 4        |
| Avg. Fashion Attributes             | 52       |
| Avg. Fashion Attributes             | 43       |
| Avg. Meta-Info Attributes           | 8        |
| Avg. Price Attributes per item      | 16       |
| Avg. Fringes per attribute          | 2/0      |
| #Fashion Synsets endorsed by Synset | 4/05     |
| #Coarse-Grained StyleTips           | 6/71     |
| #Fine-Grained StyleTips             | 3/01     |
| #Celebrities profiles               | 4/11     |
| Arg. Avg. Fashion Synsets endorsed  | 4        |
| Arg. Avg. Fringes endorsed by Synset| 15       |
| Arg. Avg. Fringes endorsed          | 252      |

Table 1: Domain Specific Knowledge Base Stats

3. Parsing the semi-structured catalog data into a uniform unstructured catalog form of the tuple <fashion synset, {fashion attribute: {attribute values}}> where {} denotes a set.

4. Constructing a distribution of attributes and their values for each of the fashion synsets, from the structured catalog data curated in step 3 and filtering them through a close manual inspection by the domain experts.

5. From the unstructured product description in the catalog, spotting and extracting style-tip information (e.g. black trousers go well with white shirt)

6. Creating fashion profiles for celebrities based on the type of clothes and accessories worn or endorsed by them. Since the profile information for real celebrities was proprietary, we generated profiles of imaginary celebrities by simulating a distribution of fashion synsets that each of these celebrities endorse, and a further distribution of fashion attributes preferred by these celebrities for each of these synsets. Note that doing so does not affect the generality of the dataset technically. Statistics about the final domain knowledge curated using this semi-automated methodology are tabulated in Table 1

2.1.2 Gathering multimodal dialogs

During the interviews, the domain experts described in detail various phases of the sales process. For example, a dialog between the sales agent and a shopper who visits an e-commerce website with the objective of either buying or browsing one or more fashion items begins by the shopper providing their shopping requirements to the agent. The agent then browses the corpus and comes back with a multimodal response (i.e. with a set of images that satisfy the shopper’s constraints and/or some associated text). Now, using this response the shopper provides feedback or modifies their requirements. Through this iterative response and feedback loop the shopper continues to explore their items of interest, adding chosen items to their shopping cart. The session continues until they either choose to exit without a purchase or culminates with the shopper buying one or more items. Note that at various steps during such a conversation, the response at the current step of the dialog is based on inference drawn from an aggregate of images and text in the unstructured dialog context as well as a structured background domain knowledge (which is again multimodal in nature).

The domain experts then provided a large number of possible flows for a sales process in terms of customers proceeding from one state to another. These transitions were captured by the automata between these states with expert designed transition probabilities. The domain experts then inspected the outputs of small runs of the automata and provided feedback. This feedback was then incorporated to further refine the expert automata whose runs were again inspected by the experts. This iterative process was manually-intensive, and required closed coordination with the domain experts. Following this process, we produced a large-scale dataset of over 150K multimodal dialogs.

2.1.3 Qualitative Survey

In order to ensure that the dataset is representative and not biased by the specific fashion experts
Table 2: Details of example automata states

| State Type          | State Description                                                                 |
|---------------------|------------------------------------------------------------------------------------|
| greeting            | Shopper greets                                                                     |
| self-date           | Shopper gives information about him/herself                                         |
| give-requirement    | Shopper gives his shopping requirements                                             |
| give-feedback,      | Shopper expresses preference towards one or more results                          |
| show-more           | in the context and possibly modifies his requirements and wants to see more         |
| give-feedback,      | Shopper expresses preference towards one or more results                          |
| show-image, show-   | shows currently or previously, and shows a new image to possibly modify his         |
| more                | requirements and wants to see more                                                 |
| show-orientation    | Shopper wants to see an item from different orientations                           |
| show-similar       | Shopper wants to see similar to a particular item                                   |
| goe-with            | Shopper asks for style-tip                                                         |
| ask-attribute       | Shopper asks about the attributes of the items shown                                |
| filter             | Shopper asks for other types of meta-info about the items shown                     |
| sort-results        | Shopper wants to sort the result by some attribute                                 |
| filter-results      | Shopper wants to filter the results based on some attribute                         |
| celebrity           | Shopper asks questions relating to some celebrities and his fashion interest         |
| switch-synset       | Shopper wants to switch back to the type of fashion synset he had seen previously   |
| buy                 | Shopper wants to buy one or more items                                             |
| exit                | Shopper wants to exit                                                               |

Table 3: Multimodal Dialog Dataset Statistics

| Dataset Statistics | Train | Valid | Test |
|--------------------|-------|-------|------|
| #Utterances per dialog | 10,410 | 12,595 | 15,493 |
| %Punctuation in terms of dialog | 70% | 15% | 15% |
| #Utterances with shopping questions | 201,113 | 246,610 | 295,472 |
| #Utterances ending in agent’s image response | 904,233 | 1,045,245 | 1,103,446 |
| #Utterances ending in agent’s text response | 1,514,183 | 1,711,536 | 1,900,379 |
| #Positive images in agent’s image response | 2 | 2 | 2 |
| #Negative images in agent’s image response | 4 | 4 | 4 |
| #Words in agent’s text response | 12 | 12 | 12 |
| #Words in shopper’s text response | 2 | 2 | 2 |
| #Utterances per dialog | 15 | 15 | 15 |
| Avg. #Utterances having a particular automata state | 20,246 | 24,346 | 29,335 |
| Avg. #Utterances having a particular state-type per dialog | 12 | 12 | 12 |
| Avg. #Utterances having a particular state-type | 59,638 | 12,806 | 12,764 |
| Multi-state states | 87 | 87 | 87 |
| #Automata states types | 2 | 2 | 2 |
| Vocabulary Size (threshold frequency >2) | 7,238 | 7,238 | 7,238 |

Additional standardization, but this will be included in a subsequent release.

2.2 Tasks

The proposed MMD dataset consists of multimodal, domain-aware conversations between 2 agents, and hence can be used for evaluating a wide variety of tasks. We describe each of these tasks and explain the technical challenges involved:

1. **Text Response**: Given a context of $k$ turns the task here is to generate the next text response.
2. **Image Response**: Given a context of $k$ turns the task here is to output the most relevant image(s).
3. **Image Retrieval**: Given a context of $k$ turns a database of images, retrieve and rank $n$ images based on their relevance to the given context.
4. **Image Generation**: Given a context of $k$ turns, generate the most relevant image (typically performed using generative models e.g. contextual GANs(Red et al., 2016; Goodfellow et al., 2014)).

We propose both tasks since the evaluation criteria for each approach is quite different.

**Employing Domain-knowledge**: This is essentially performing tasks (1) and (2) of text and image response generation using both the unstructured dialog context along with the structured domain knowledge. We propose this as a separate task to evaluate the impact of domain-knowledge.

**User Modeling**: Another important aspect of conversations is to study user behavior. Since different users exhibit significantly varying shopping behaviors such as buying preferences, speed of decision making, etc. we propose a task that explicitly models the shopper since it impacts the agent’s most appropriate response at each step.

**Setup**: In this work, we focus on tasks (1) and (2.1). More specifically, the conversation system should be able to perform the following: (i) by utilizing the context, decide the modality of the response, i.e., whether the response should be a text or an image (ii) generate a text response if
required, or (iii) if an image response is required then rank all the images in the catalog based on their contextual relevance as opposed to ranking a given set of \( m \) \( (m < \text{catalog size}) \) images.

In this work, we make two simplifications to this setup: (a) We evaluate the text response and image response task separately, which means that the system does not need to decide the modality of the response, and (b) instead of retrieving and ranking all of the images in the catalog/database, the system needs to rank only a given smaller subset of \( m \) images, which contain the correct image(s) in addition to a few incorrect ones. This simplified evaluation protocol of “selecting/ranking” the best response was proposed for the Ubuntu Dialog Corpus (Lowe et al., 2015) and helps provide more control on the experimental setup and evaluate individual parts of the system in a more thorough manner.

Given this setup, in the next section we propose baseline models for these tasks based on the encode-attend-decode paradigm.

### 3 Models

To empirically estimate the feasibility of the two tasks described earlier we implement one baseline method (and some variations thereof) for each task. These methods are based on the popular hierarchical encode-attend-decode paradigm (Serban et al., 2016a) typically used for (text) conversation systems. We split the description below into two parts

- **(i) Multimodal encoder** which is common for the two tasks
- **(ii) Multimodal decoder** which is different depending on whether we need to generate a text response or predict an image response.

#### 3.1 Multimodal encoder

As mentioned earlier, for both the tasks, the context contains \( k \) utterances. Each utterance in the context could either be (i) a text only utterance or (ii) an image only utterance or (iii) a multimodal utterance containing both text and images (as shown in Figure 2). We use a multimodal hierarchical encoder for computing the encoder representation in each of these cases as described below.

- **(a) Text only utterance:** Every text utterance in the context is encoded using a recurrent neural network with GRU (Chung et al., 2014) cells in a process similar to the utterance level encoder described in (Serban et al., 2016a). This is the level 1 encoder in the hierarchical encoder.

- **(b) Image only utterance:** In the current setup, each image or multimodal utterance can contain \( n_{\text{max}} \) number of images (in the example shown in Figure 2 \( n_{\text{max}} \) is 3). We encode each image using a 4096 dimensional representation obtained from the FC7 layer of a VGGNet-16 (Simonyan and Zisserman, 2014) convolutional neural network. The representation of the utterance is simply the concatenation of the individual images. This is also a part of the first level in the hierarchical encoder.

- **(c) Multimodal utterance:** The text portion of the multimodal utterance is encoded using the same GRU cells as used for encoding the text only utterance. Similarly, the images in the multimodal utterance are encoded using the same VGGNet-16 as used for the image only utterance. The final representation of the multimodal utterance is simply the concatenation of the individual utterances.

Note that we expect a fixed number of images...
for each utterance and if there are fewer images we pad the utterance with empty images (similar to what is typically done for text inputs).

The multimodal utterance representation is then fed to a level two encoder which is again a GRU. This second level encoder (or context-level encoder) essentially encodes the sequence of utterances where the representation of each utterance in the sequence is computed and projected as described above. Figure 2 and Figure 3 show this process of computing the encoder representation for a given multimodal context.

3.2 Decoder for generating text responses

As shown in Fig. 2, we use a standard recurrent neural network based decoder with GRU cells. Such a decoder has been used successfully for various natural language generation tasks including text conversation systems (Serban et al., 2016b). We also implemented a version where we couple the decoder with an attention model. The attention model learns to pay attention to different time-steps of the second level encoder (again this has been tried successfully in the context of text conversation systems (Yao et al., 2016)).

3.3 Layer for ranking image responses:

The task here is to rank a given set of images depending on their relevance to the context. While training we are given a set of \( m \) images for each context of which only \( n_{\text{pos}_{\text{max}}} \) are relevant for the context. The remaining \( m - n_{\text{pos}_{\text{max}}} \) are picked from the corresponding false image responses in the dataset. We train the model using a max margin loss. Specifically, we compute the cosine similarity between the learnt image embedding and the encoded representation of the multimodal context.

The model is then trained to maximize the margin between the cosine similarity for the correct image and the incorrect images. Fig. 3 depicts this for the case when \( m = 2 \) and \( n_{\text{pos}_{\text{max}}} = 1 \).

Note that due to space constraints we only provide pictorial representations of these models.

4 Experiments

In this section we describe the experimental setup used to evaluate the following models on the two tasks:

- **Hierarchical Encoder Decoder (ignoring image context),** whose architecture is similar to that proposed in (Serban et al., 2016a)
- The proposed **Multimodal Hierarchical Encoder Decoder**, where we varied the size of the dialog context.
- The proposed **Multimodal Hierarchical Encoder Decoder with attention** over the multimodal utterance representation at each time step

4.1 Evaluating the Text Response Task

For this task we only considered those dialogs which end with a text response. The sizes of the training, validation and test sets are reported in the 6th row of Table 3. We tuned the following hyperparameters using the validation set: learning rate \( \in \{1e-2, 1e-3, 4e-4\} \), RNN hidden unit size \( \in \{256, 512\} \), text embedding size \( \in \{256, 512\} \), image embedding size \( \in \{256, 512\} \), batch size \( \in \{32, 64\} \). The numbers in brackets indicate the distinct values of each hyperparameter we considered. We used Adam (Kingma and Ba, 2014) as the optimization algorithm. The results in Table 6 summarize the BLEU and NIST scores used as the evaluation metric for this task.
Figure 3: Multimodal Hierarchical Encoder Architecture for Image Response Task. The figure shows a single target positive and negative image, but in general, $m(>1)$ images can be provided as target.

| Table 6: Performance of the different models on the “Text Response Generation” task. |
|--------------------------|------------------|------------------|
| Model                     | BLEU          | NIST           |
| HRED (ignoring image context) | 36.70         | 5.30           |
| Multimodal HRED (dialog context=2) | 33.65         | 4.95           |
| Multimodal HRED (dialog context=5) | 37.14         | 5.43           |
| Multimodal HRED (dialog context=10) | 38.2          | 5.37           |
| Multimodal HRED with Attention (dialog context=5) | 30.48         | 4.43           |

| Model                     | Recall@1    | Recall@2  | Recall@3  |
|--------------------------|-------------|-----------|-----------|
| HRED (ignoring image context) | 42.8%      | 59.8%     | 72.4%     |
| Multimodal HRED (dialog context=2) | 43.4%      | 60.0%     | 72.3%     |
| Multimodal HRED (dialog context=5) | 45.5%      | 61.9%     | 73.1%     |
| Multimodal HRED (dialog context=10) | 36.8%     | 54.8%     | 68.6%     |
| Multimodal HRED with Attention (dialog context=5) | 40.0%      | 51.8%     | 65.8%     |

Table 7: Performance of the different models on the “Select Best Image Response” Task.

4.2 Evaluating Image Response Task

Here we only consider those dialogs that end in an image response from the system. The sizes of the training, validation and test sets are reported in the 5th row of Table 3. Both during training and testing, the model is provided with $m=5$ target images out of which only $n_{pos_{max}}=1$ is relevant and at test time the model has to rank the images in order of their relevance as a response to the given context. The hyperparameters of the model were tuned in the same way as mentioned above.

We use Recall@top-m as the evaluation metrics where top-m is varied from 1 to 3, and the model prediction is considered to be correct only if the true response is among the top-m entries in the ranked list. The results are summarized in Table 7. In the supplementary material we provide examples of the model’s outputs for the two tasks.

4.3 Discussions

We make a few observations from the results.

- For both tasks, the basic model which does not use any image information performs almost at par with the models which use this information. This suggests that we need better models for capturing the interactions between text and images. In particular, we need some common representation learning for the two modalities.

- While BLEU gives more importance to word ordering in calculating similarity, NIST emphasizes more on informative words. The low NIST scores suggest that the models need some external knowledge to understand the informative words in the domain.

- Adding attention degrades the performance. Though counter intuitive, this suggests we need better models for multimodal attention.

- Overall, we feel there is a lot of scope for improvement and the current models only establish the feasibility of the two tasks.

5 Conclusion

In this paper, we introduced the Multimodal Dialogs (MMD) dataset that is well suited to the task of studying multimodal domain-aware conversations in a wide variety of settings. This dataset was gathered by working closely with a group of 20 experts in the retail domain and consists of over 15 Million conversation sessions between shoppers and sales agents involving both text and images. Driven by the MMD dataset, we proposed 5 new sub-tasks along with their evaluation methodologies. We also proposed 2 multimodal deep learning models using the encode-attend-decode paradigm and demonstrated their performance on both text response generation and best image response selection. The performance numbers obtained demonstrate the feasibility of the involved sub-tasks and highlight the challenges involved.
References
Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pages 2425–2433. https://doi.org/10.1109/ICCV.2015.279.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. CoRR abs/1611.08669. http://arxiv.org/abs/1611.08669.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. CoRR abs/1611.08481. http://arxiv.org/abs/1611.08481.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pages 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980. http://arxiv.org/abs/1412.6980.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic. pages 285–294. http://aclweb.org/anthology/W/W15/W15-4640.pdf.

Tegan Maharaj, Nicolas Ballas, Aaron C. Courville, and Christopher Joseph Pal. 2016. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. CoRR abs/1611.07810. http://arxiv.org/abs/1611.07810.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. CoRR abs/1701.08251. http://arxiv.org/abs/1701.08251.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text-to-image synthesis. In Proceedings of The 33rd International Conference on Machine Learning.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings. June 2-4, 2010, Los Angeles, California, USA. pages 172–180. http://www.aclweb.org/anthology/N10-1020.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, AAAI’16, pages 3776–3783. http://dl.acm.org/citation.cfm?id=3016387.3016435.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Lauren Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. CoRR abs/1605.06069. http://arxiv.org/abs/1605.06069.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. CoRR abs/1506.05869. http://arxiv.org/abs/1506.05869.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pages 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. pages 2048–2057. http://jmlr.org/proceedings/papers/v37/xuc15.html.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. CoRR abs/1606.01292. http://arxiv.org/abs/1606.01292.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas,
Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2016. Leveraging video descriptions to learn video question answering. CoRR abs/1611.04021. http://arxiv.org/abs/1611.04021.
6 Supplementary Materials

6.1 Examples of Text Responses generated by the proposed Multimodal HRED model

This section enumerates some of the examples taken from the test set, of the model generating the agent’s text responses. In each figure, the last row (in red) is the predicted output of the system and directly above it is the true response from the original dialog in the dataset. The multimodal content appearing above is the context fed into the model to generate this response.

**Figure 4: Example text response generated by the proposed multimodal HRED model**

| SHOPPER: And what about celebrity cel_334 for the 5th one? |
|-----------------------------------------------------------|
| AGENT: No.                                                |
|                                                           |
| SHOPPER: I don’t like the 1st image. Show me something in material as in this image |
| AGENT:                                                    |
|                                                           |
| SHOPPER: How about the trackpants in the 4th image? What will go well with it? |
| AGENT: It can go well with casual style, cotton fabrics t-shirt |
| SHOPPER: And what about with the 4th one?                 |
| AGENT: It can go well with lifestyle, silver colouring footwear and with sports style, white coloring, casual-sports type t-shirt |
| **PREDICTED:** it can go well with casual style             |

**Figure 5: Example text response generated by the proposed multimodal HRED model**

| AGENT: Hello, is there anything I can help you with? |
|-----------------------------------------------------|
| SHOPPER: I am looking for my uncle. His age is 40 years. |
| AGENT: Thanks for the information                    |
| SHOPPER: I am very much interested in going through your collection of training shoes within the range USD between 120.0 and 620.0 that i can gift him. |
| AGENT: Good/I’ll just take a quick scan through my catalog |
| AGENT:                                               |
| SHOPPER: Does celebrity cel_230 wear the kind of training shoes in the 4th image? |
| **PREDICTED:** Yes.                                  |

6.2 Examples of Target Images Ranked by the proposed Multimodal HRED model

Here, we similarly enumerate some of the example outputs for the Image Response Selection task taken from the test set. The model is given a multimodal context of an ongoing dialog and a set of target images, which it has to rank in order of their relevance as a response, given the context. The target images are shown in the bottom row for every example, and the true positive response and the true negative response (according to the dataset) is demarcated by a green or a red box respectively. The model internally scores each of these images based on the likelihood of the image being a response, given the context. In the
SHOPPER: I like the 4th image. Show me something like it but in brand as in this image. Can you please show within my budget?

AGENT:

SHOPPER: Are the products in the 3rd and 5th images suited for casual style?

AGENT: Yes

SHOPPER: And what about the ones in the 2nd, 1st and 4th images?

AGENT: Yes

PREDICTED: Yes

---

SHOPPER: I like the 2nd sandals. Show me something similar to it but in suede soled type

AGENT:

SHOPPER: I don't like the 4th image. Show me something in material as in this image

AGENT:

SHOPPER: What is the 2nd images material?

AGENT: The material in 2nd image is mesh

PREDICTED: The material in 2nd image is synthetic

---

Figure 6: Example text response generated by the proposed multimodal HRED model

Figure 7: Example text response generated by the proposed multimodal HRED model

figures, the target images have been sorted by the cosine similarity score given out by the model, where a higher value of similarity implies a higher confidence of the model in that image being a relevant response.
6.3 Examples of dialog sessions from MMD dataset

In this section we show some of the sample dialog sessions between a shopper and an agent, where each dialog culminates in either a successful purchase or with the shopper exiting without a purchase.
**SHOPER:** What is the color and print in the 3rd result?

**AGENT:** the coat in the 3rd image has black color lace print

**SHOPER:** And fit and color in the 1st one?

**AGENT:** the coat in the 1st image has blue color slim fit

**SHOPER:** I like the 2nd image. Show me something like it but in material as in this image

![Image of a jacket]

**AGENT:**

| Target Images | Cosine Similarity |
|---------------|------------------|
|               | 0.87             |
|               | 0.52             |
|               | 0.41             |
|               | -0.61            |
|               | -0.78            |
|               | -0.84            |

Figure 10: Example target images ranked by the proposed multimodal HRED model

**SHOPER:** Hello

**AGENT:** Hi, How can i help you today?

**SHOPER:** I am buying for my buddy. she is 33-yr old.

**AGENT:** Great!

**SHOPER:** I am interested in going through hand bag within INR 1400.0 that i can buy for her.

**AGENT:** Good!Got it. Just showing you in a minute

**AGENT:**

| Target Images | Cosine Similarity |
|---------------|------------------|
|               | 0.92             |
|               | 0.63             |
|               | 0.75             |
|               | 0.74             |
|               | 0.63             |
|               | 0.6              |
|               | 0.58             |
|               | -0.56            |

Figure 11: Example target images ranked by the proposed multimodal HRED model
Figure 12: Example dialog session between a shopper and an agent

Figure 13: Example dialog session between a shopper and an agent
Figure 14: Example dialog session between a shopper and an agent

Shopping Assistant ChatLog

SHOPPER: Hello, please tell me, anything I can help you with today?
AGENT: I am keen on looking at some of your best handbags that would suit me.
SHOPPER: It would help if you could know your gender.
AGENT: Great. Knowing you, just a galleria.
SHOPPER: Ignore me the image of the hot shoes from all different directions.
AGENT: Image from the profile, right, back and left views respectively.
SHOPPER: And how about the side?
AGENT: Image from the profile orientation.
SHOPPER: Which of the celebrities wore similar looking hand bag as in the 3rd image?
AGENT: Celebrity: ol, not countries like that hand bag.
SHOPPER: Any to the 3rd one?
AGENT: Celebrity: ol, 2016, and ol, 2016, have similar this kind of hand bag.
SHOPPER: Can I have the most images from all possible orientations?
AGENT: Image from the profile, right, back and left views respectively.
SHOPPER: Great.
AGENT: Good morning.
SHOPPER: I would like to buy some v-neck shaped casual jacket in a regular fit that can be buy for my family.
AGENT: May I please ask the gender?
SHOPPER: female.
SHOPPER: She is in 45 years of age.
AGENT: Okay.
SHOPPER: Can I just take few seconds to quickly browse through your catalog?
AGENT: sure, I can't find anything in v-neck that I can show you some in different necks.
SHOPPER: Can you show me something in tops that looks long to type. I just like the 5th casual jacket.
AGENT: Yes.
SHOPPER: And how about the 3rd model's complement? Will it go well with a fawn colored pant?
AGENT: Yes, and sandals that looks brown colored.
SHOPPER: Yes.
SHOPPER: Which of the celebrities usually wear casual jacket similar to the one in the 4th image?
AGENT: Celebrities: ol, and ol, 2016, and ol, 2016, wear the kind of casual jacket.
SHOPPER: I am here to look for brown leather type sole material contained in casual that can gift her.
AGENT: You.
SHOPPER: I like the kind of purse usually worn by celebrity ol, 2016. Can you show more like these?
AGENT: Yes.
SHOPPER: The kind of purse is as shown.
AGENT: Yes.
SHOPPER: Dear celebrity ol, 2016 tastes the kind of sandals in the 4th image?
AGENT: Yes.
SHOPPER: I have 2nd result images from some possible orientation?
AGENT: Image from the front, right, back and left views respectively.
SHOPPER: No I think still the sandals I want to go back and look at the casual jacket you had shown before.
SHOPPER: Do you think it is style military in the 2nd image?
AGENT: No. The style in the 2nd image is casual.
SHOPPER: And for this pair product as well?
AGENT: Yes.
SHOPPER: I like this kind of bag usually worn by celebrity ol, 2016. Can you show more like these?
AGENT: Yes.
SHOPPER: How about a few more option in the material as in the 3rd image?
AGENT: Yes.
SHOPPER: I would like to look at your collection of winter type chains does containing some made out of metals of about P1000 to P5000 the appeal.
AGENT: Sorry, I can't find anything in such but would you like something in love?
SHOPPER: Like the sole material in the 4th product but not the 1st. Please show me some more?
AGENT: Absolutely, this is a great Sandoles.
SHOPPER: Will the product in the 3rd match justify the skirt constituted of fancy material in ths image?
AGENT: Yes.
SHOPPER: And how about the picture type sole material contained in connection to this image?
AGENT: No.
SHOPPER: And, how about the style in the 4th image?
AGENT: The casual jacket in the 4th image has elegant and casual style.
SHOPPER: What about the color?
AGENT: The casual jacket in the 4th image has blue color.
SHOPPER: I would like to see wrap around with the soles that she would like.
AGENT: Yes.
SHOPPER: Dear celebrity ol, 2017 usually wear the kind of skirt in the 1st image?
AGENT: No.
SHOPPER: Can you show me something in having material like in cortes.
AGENT: I can't like her furtet.
SHOPPER: I want to buy the 1st one.
AGENT: Sure, this is a great Sandoles.

Figure 15: Example dialog session between a shopper and an agent
Figure 16: Example dialog session between a shopper and an agent

SHOPPER: Hi there.
AGENT: Hello, what can I help you with?
SHOPPER: Hi. Can you show me some pretty styled shawls that would suit me?
AGENT: May I please ask your gender?
SHOPPER: Female.
AGENT: I see. How old are you?
SHOPPER: I am a woman, 25 years of age.
AGENT: Ok great! This helpful!
SHOPPER: Let me just quickly scan through my catalogue.
AGENT:
SHOPPER: Can I have 3rd results images from all possible orientations?
AGENT: Images from the profile, back and left orientations respectively.
SHOPPER: And for the result in the 2nd image?
AGENT: Images from the profile, right back and left orientations respectively.
SHOPPER: Will scarf in the 3rd image complement outdoor style?
AGENT: Yes.
SHOPPER: What style would be suited for scarf in 5th, 6th and 2nd images?
AGENT: The scarf in the 5th and 6th images will suit well for chic and elegant and vibrant style.
SHOPPER: I quite like the style in the 3rd one. Can you show me some more?
AGENT:
SHOPPER: I want to see something similar to the 5th image but in a different material?
AGENT: The similar looking ones are...
SHOPPER: Can you show these sorted by best seller ranking?
AGENT:
SHOPPER: Will the product in the 3rd result agree with the black colored thermal in this image?
AGENT:
SHOPPER: Am interested in buying the 5th one?
AGENT: Great choice! Hope you had a great shopping experience.