# Visual Dialog

Abhishek Das[1], Satwik Kottur[2], Khushi Gupta[2]*, Avi Singh[3]*, Deshraj Yadav[1], José M.F. Moura[2],
Devi Parikh[4], Dhruv Batra[4]
[1]Virginia Tech, [2]Carnegie Mellon University, [3]UC Berkeley, [4]Georgia Institute of Technology
[1]{abhshkdz, deshraj}@vt.edu [2]{skottur, khushig, moura}@andrew.cmu.edu [3]avisingh@cs.berkeley.edu
[4]{parikh, dbatra}@gatech.edu
visualdialog.org

## Abstract

*We introduce the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a question about the image, the agent has to ground the question in image, infer context from history, and answer the question accurately. Visual Dialog is disentangled enough from a specific downstream task so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress. We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). Data collection is underway and on completion, VisDial will contain 1 dialog with 10 question-answer pairs on all ∼200k images from COCO, with a total of 2M dialog question-answer pairs.*

*We introduce a family of neural encoder-decoder models for Visual Dialog with 3 encoders – Late Fusion, Hierarchical Recurrent Encoder and Memory Network – and 2 decoders (generative and discriminative), which outperform a number of sophisticated baselines. We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a set of candidate answers and evaluated on metrics such as mean-reciprocal-rank of human response. We quantify gap between machine and human performance on the Visual Dialog task via human studies. Our dataset, code, and trained models will be released publicly. Putting it all together, we demonstrate the first 'visual chatbot'!*

## 1. Introduction

We are witnessing unprecedented advances in computer vision (CV) and artificial intelligence (AI) – from 'low-level' AI tasks such as image classification [14], scene recognition [55], object detection [27] – to 'high-level' AI tasks such as learning to play Atari video games [35] and Go [47], answering reading comprehension questions by understanding short stories [15, 57], and even answering questions about images [4, 32, 41] and videos [49, 50]!


Figure 1: We introduce a new AI task – Visual Dialog, where an AI agent must hold a dialog with a human about visual content. We introduce a large-scale dataset (VisDial), an evaluation protocol, and novel encoder-decoder models for this task.

**What lies next for AI?** We believe that the next generation of visual intelligence systems will need to posses the ability to hold a meaningful dialog with humans in natural language about visual content. Applications include:

- Aiding visually impaired users in understanding their surroundings [5] or social media content [58] (AI: *'John just uploaded a picture from his vacation in Hawaii'*, Human: *'Great, is he at the beach?'*, AI: *'No, on a mountain'*).
- Aiding analysts in making decisions based on large quantities of surveillance data (Human: *'Did anyone enter this*

*Work done while KG and AS were interns at Virginia Tech.

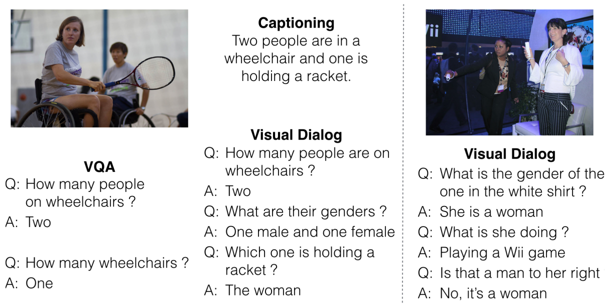



Figure 2: Differences between image captioning, Visual Question Answering (VQA) and Visual Dialog. Two (partial) dialogs are shown from our VisDial dataset, which is curated from a live chat between two Amazon Mechanical Turk workers (Sec. 3).

*room last week?'*, AI: *'Yes, 27 instances logged on camera'*, Human: *'Were any of them carrying a black bag?'*),

- Interacting with an AI assistant (Human: *'Alexa – can you see the baby in the baby monitor?'*, AI: *'Yes, I can'*, Human: *'Is he sleeping or playing?'*).
- Robotics applications (*e.g.* search and rescue missions) where the operator may be 'situationally blind' and operating via language [33] (Human: *'Is there smoke in any room around you?'*, AI: *'Yes, in one room'*, Human: *'Go there and look for people'*).

Despite rapid progress at the intersection of vision and language – in particular, in image captioning and visual question answering (VQA) – it is clear that we are far from this grand goal of an AI agent that can 'see' and 'communicate'. In captioning, the human-machine interaction consists of the machine simply *talking at* the human (*'Two people are in a wheelchair and one is holding a racket'*), with no dialog or input from the human. While VQA takes a significant step towards human-machine interaction, it still represents only *a single round of a dialog* – unlike in human conversations, there is no scope for follow-up questions, no memory in the system of previous questions asked by the user nor consistency with respect to previous answers provided by the system (Q: *'How many people on wheelchairs?'*, A: *'Two'*; Q: *'How many wheelchairs?'*, A: *'One'*).

As a step towards conversational visual AI, we introduce a novel task – **Visual Dialog** – along with a large-scale dataset, an evaluation protocol, and novel deep models.

**Task Definition.** The concrete task in Visual Dialog is the following – given an image $I$, a history of a dialog consisting of a sequence of question-answer pairs (Q1: *'How many people are in wheelchairs?'*, A1: *'Two'*, Q2: *'What are their genders?'*, A2: *'One male and one female'*), and a natural language follow-up question (Q3: *'Which one is holding a racket?'*), the task for the machine is to answer the question in free-form natural language (A3: *'The woman'*). This task is the visual analogue of the Turing Test.

Consider the Visual Dialog examples in Fig. 2. The question *'What is the gender of the one in the white shirt?'* requires the machine to selectively focus and direct attention to a relevant region. *'What is she doing?'* requires co-reference resolution (whom does the pronoun 'she' refer to?), *'Is that a man to her right?'* further requires the machine to have visual memory (which object in the image were we talking about?). Such systems also need to be consistent with their outputs – *'How many people are in wheelchairs?'*, *'Two'*, *'What are their genders?'*, *'One male and one female'* – note that the number of genders being specified should add up to two. Such difficulties make the problem a highly interesting and challenging one.

**Why do we talk to machines?** Prior work in language-only (non-visual) dialog can be arranged on a spectrum with the following two end-points:
goal-driven dialog (*e.g.* booking a flight for a user) $\longleftrightarrow$ goal-free dialog (or casual 'chit-chat' with chatbots).
The two ends have vastly differing purposes and conflicting evaluation criteria. Goal-driven dialog is typically evaluated on task-completion rate (how frequently was the user able to book their flight) or time to task completion [9,36] – clearly, the shorter the dialog the better. In contrast, for chit-chat the longer the user engagement and interaction, the better. For instance, the goal of the 2017 \$2.5 Million Amazon Alexa Prize is to "create a socialbot that converses coherently and engagingly with humans on popular topics for 20 minutes."

We believe our instantiation of Visual Dialog hits a sweet spot on this spectrum. It is *disentangled enough* from a specific downstream task so as to serve as a general test of machine intelligence, while being *grounded enough* in vision to allow objective evaluation of individual responses and benchmark progress. The former discourages task-engineered bots for 'slot filling' [23] and the latter discourages bots that put on a personality to avoid answering questions while keeping the user engaged [56].

**Contributions.** We make the following contributions:

- We propose a new AI task: Visual Dialog, where a machine must hold a dialog with a human about visual content.
- We develop a novel two-person chat data-collection protocol to curate a large-scale Visual Dialog dataset (VisDial). Upon completion[1], VisDial will contain 1 dialog with 10 question-answer pairs on *all* 200k images from COCO [25], for a total of 2M dialog question-answer pairs. When compared to the popular VQA dataset [4], VisDial studies a significantly richer task (dialog), overcomes a 'visual priming bias' in VQA (in VisDial, the person asking the questions does not see the image), contains free-form longer answers (mean length: 3.1 words in VisDial *vs.* 1.1 in VQA), and is *an order of magnitude* larger.

[1]Data collection is underway. As of this submission, VisDial has 68k dialogs on 68k images.

- We introduce a family of neural encoder-decoder models for Visual Dialog with 3 novel encoders
  - Late Fusion: that embeds the image, history, and question into vector spaces separately and performs a 'late fusion' of these into a joint embedding.
  - Hierarchical Recurrent Encoder: that contains a dialog-level Recurrent Neural Network (RNN) sitting on top of a question-answer ($QA$)-level recurrent block. In each $QA$-level recurrent block, we also include an attention-over-history mechanism to choose and attend to the round of the history relevant to the current question.
  - Memory Network: that treats each previous $QA$ pair as a 'fact' in its memory bank and learns to 'poll' the stored facts and the image to develop a context vector.

  We train all these encoders with 2 decoders (generative and discriminative) – all settings outperform a number of sophisticated baselines, including our adaption of state-of-art VQA models to VisDial.
- We propose a retrieval-based evaluation protocol for Visual Dialog where the AI agent is asked to sort a list of candidate answers and evaluated on metrics such as mean-reciprocal-rank of the human response.
- We conduct studies to quantify human performance on this task.
- Putting it all together, we demonstrate a visual chatbot in a video at `https://goo.gl/yjlHxY` that answers a sequence of questions from a user about an image!

## 2. Related Work

**Vision and Language.** A number of problems at the intersection of vision and language have recently gained prominence – image captioning [10, 11, 20, 54], video/movie description [43, 51, 52], text-to-image coreference/grounding [16, 22, 37, 39, 42], Visual Madlibs [60], and of course, visual question answering (VQA) [4, 12, 31, 32, 41]. However, all of these works involve a single-shot natural language interaction with visual content – there is no dialog.

**Visual Turing Test.** Perhaps the most closely related to our work is that of Geman *et al*. [13], who proposed a 'Visual Turing Test' in a fairly restricted setting. First, questions in [13] were automatically generated from 4 kinds of fixed templates (existence, uniqueness, attributes, relationships), while our dataset has *free form, open-ended* natural language questions collected via two workers chatting on Amazon Mechanical Turk (AMT), resulting in a much more realistic and diverse dataset (see Fig. 5). Second, all questions in [13] were binary (yes/no), while our chat interface results in a rich set of free-form responses (see Fig. 5). Third, the dataset in [13] only contains street scenes, while our dataset has considerably more variety since it uses images from the COCO dataset [25]. Moreover, our dataset is planned to be *two orders of magnitude larger* – 2,591 images in [13] vs ∼200k images, 10 question-answer pairs, a total of ∼2M question-answer pairs. Finally, the focus of Geman *et al*. [13] is a statistical templated-question generator and not an actual visual dialog system. We propose several deep models for Visual Dialog, an evaluation protocol, and conduct human accuracy studies for this task.

**Text-based Question Answering.** Our work is related to text-based question answering or 'reading comprehension' tasks studied in the NLP community. Some recent large-scale datasets in this domain include the 30M Factoid Question-Answer corpus [44], 100K SimpleQuestions dataset [6], DeepMind Q&A dataset [15], the 20 artificial tasks in the bAbI dataset [57], and the SQuAD dataset for reading comprehension [38]. VisDial can be viewed as a *fusion* of reading comprehension and VQA. In VisDial, the machine must comprehend the history of the past dialog and then understand the image to answer the question. By design, the answer to any question in VisDial is not present in the past dialog – if it were, the question would not be asked. The history of the dialog *contextualizes* the question – the question *'what else is she holding?'* requires a machine to comprehend the history to realize who the question is talking about and what has been excluded, and then understand the image to answer the question.

**Conversational Modeling and Chatbots.** Visual Dialog is the visual analogue of text-based dialog and conversation modeling. While some of the earliest developed chatbots were rule-based [56], end-to-end learning based approaches are now being actively explored [7,9,19,24,45,46,53]. A recent large-scale conversation dataset is the Ubuntu Dialogue Corpus [28], which contains about 500K dialogs extracted from the Ubuntu channel on Internet Relay Chat (IRC). Liu *et al*. [26] perform a study of problems in existing evaluation protocols for free-form dialog. One important difference between free-form textual dialog and VisDial is that in VisDial, the two participants are not symmetric – one person (the 'questioner') asks questions about an image *that they do not see*; the other person (the 'answerer') sees the image and only answers the questions (in otherwise unconstrained text, but no counter-questions allowed). This role assignment gives a sense of purpose to the interaction (why are we talking? To help the questioner build a mental model of the image), and allows objective evaluation of individual responses. Next, we describe our data collection protocol.

## 3. The Visual Dialog Dataset (VisDial)

We now describe our VisDial dataset. We begin by describing the chat interface and data-collection process on AMT, analyze the dataset,and then discuss the evaluation protocol.

Consistent with previous data collection efforts, we collect visual dialog data on images from the Common Objects in


(a) What the 'questioner' sees. (b) What the 'answerer' sees. (c) Example dialog from our VisDial dataset.

Figure 3: Collecting visually-grounded dialog data on Amazon Mechanical Turk via a live chat interface where one person is assigned the role of 'questioner' and the second person is the 'answerer'. We show the first two questions being collected via the interface as Turkers interact with each other in Fig. 3a and Fig. 3b. Remaining questions are shown in Fig. 3c.

Context (COCO) [25] dataset, which contains multiple objects in everyday scenes. The visual complexity of these images allows for engaging and diverse conversations to be held about them.

**Live Chat Interface.** Good data for this task should include dialogs that have (1) temporal continuity, (2) grounding in the image, and (3) mimic natural 'conversational' exchanges. To elicit such responses, we paired 2 workers on AMT to chat with each other in real-time (Fig. 3). Each worker was assigned a specific role. One worker (the 'questioner') sees only a single line of text describing an image (caption from COCO); the image remains hidden to the questioner. Their task is to ask questions about this hidden image so as to 'imagine the scene better'. The second worker (the 'answerer') sees the image and the caption. Their task is to answer the questions asked by their chat partner. Unlike VQA [4], answers are not restricted to be short or concise, instead workers will be encouraged to reply as naturally and 'conversationally' as possible. An example dialog is shown in Fig. 3c.

This process is an unconstrained 'live' chat, with the only exception that the questioner must wait to receive an answer before posting the next question. The workers are allowed to end the conversation after 20 messages are exchanged (10 pairs of questions and answers). Further details about our final interface can be found in the supplement.

We also piloted a different setup where the questioner saw a highly blurred version of the image, instead of the caption. The conversations seeded with blurred images resulted in questions that were essentially 'blob recognition' – *'What is the pink patch at the bottom right?'*. For our full-scale data-collection, we decided to seed with just the captions since it resulted in more 'natural' questions and more closely modeled the real-world applications discussed in Section 1 where no visual signal is available to the human.

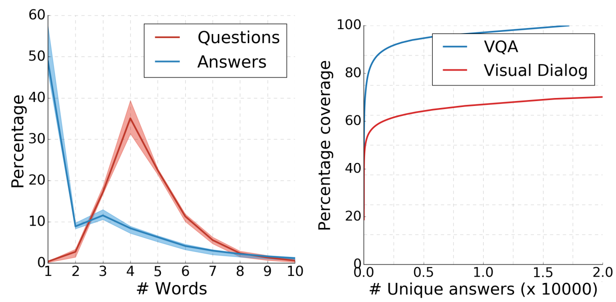
**Building a 2-person chat on AMT.** Despite the popularity of AMT as a data collection platform in computer vision, our setup had to design for and overcome some unique challenges – the key issue being that AMT is simply not designed for multi-user Human Intelligence Tasks (HITs). Hosting a live two-person chat on AMT meant that none of the Amazon tools could be used and we developed our own backend messaging and data-storage infrastructure based on Redis messaging queues and Node.js. To support data quality, we ensured that a worker could not chat with themselves (using say, two different browser tabs) by maintaining a pool of worker IDs paired. To minimize wait time for one worker while the second was being searched for, we ensured that there was always a significant pool of available HITs. If one of the workers abandoned a HIT (or was disconnected) midway, automatic conditions in the code kicked in asking the remaining worker to either continue asking questions or providing facts (captions) about the image (depending on their role) till 10 messages were sent by them. Workers who completed the task in this way were fully compensated, but our backend discarded this data and automatically launched a new HIT on this image so a real two-person conversation could be recorded. Our entire data-collection infrastructure (front-end UI, chat interface, backend storage and messaging system, error handling protocols) will be publicly available to help future efforts.

## 4. VisDial Dataset Analysis

We now analyze the v0.5 subset of our VisDial dataset collected so far – it contains 1 dialog (10 question-answer pairs) on 68k images from COCO (58k `train` and 10k `val`), or a total of 680,000 QA pairs.

### 4.1. Analyzing VisDial Questions

**Visual Priming Bias.** One key difference between VisDial and previous image question-answering datasets (VQA [4], Visual 7W [62], Baidu mQA [12]) is the lack of a 'visual priming bias' in VisDial. Specifically, in all previ-




(a) (b)

Figure 4: Distribution of lengths for questions and answers (left); and percent coverage of unique answers over all answers from the train dataset (right), compared to VQA. For a given coverage, VisDial has more unique answers indicating greater answer diversity.

ous datasets, subjects saw an image while asking questions about it. As described in [61], this leads to a particular bias in the questions – people only ask *'Is there a clocktower in the picture?'* on pictures actually containing clock towers. This allows language-only models to perform remarkably well on VQA and results in an inflated sense of progress [61]. As one particularly perverse example – for questions in the VQA dataset starting with *'Do you see a . . . '*, blindly answering *'yes'* without reading the rest of the question or looking at the associated image results in an average VQA accuracy of 87%! In VisDial, questioners *do not* see the image. As a result, this bias is reduced.

**Distributions.** Fig. 7a shows the distribution of question lengths in VisDial – we see that most questions range from four to ten words. Fig. 5 shows 'sunbursts' visualizing the distribution of questions (based on the first four words) in VisDial *vs*. VQA. While there are a lot of similarities, some differences immediately jump out. There are more binary questions[2] in VisDial as compared to VQA – the most frequent first question-word in VisDial is 'is' *vs*. 'what' in VQA. A detailed comparison of the statistics of VisDial *vs*. other datasets is available in Table 1 in the supplement.

Finally, there is a stylistic difference in the questions that is difficult to capture with the simple statistics above. In VQA, subjects saw the image and were asked to stump a smart robot. Thus, most queries involve specific details, often about the background (Q:'What program is being utilized in the background on the computer?'). In VisDial, questioners did not see the original image and were asking questions to build a mental model of the scene. Thus, the questions tend to be open-ended, and often follow a pattern:

- Generally starting with the **entities in the caption**:

  *'An elephant walking away from a pool in an exhibit'*, *'Is there only 1 elephant?'*,

- digging deeper into their **parts, attributes, or properties**:

  *'Is it full grown?'*, *'Is it facing the camera?'*,

- asking about the **scene category or the picture setting**:

  *'Is this indoors or outdoors?'*, *'Is this a zoo?'*,

- **the weather**: *'Is it snowing?'*, *'Is it sunny?'*,

- simply **exploring the scene**:

  *'Are there people?'*, *'Is there shelter for elephant?'*,

- and asking **follow-up questions** about the new visual entities discovered from these explorations:

  *'There's a blue fence in background, like an enclosure'*, *'Is the enclosure inside or outside?'*.

Such a line of questioning does not exist in the VQA dataset, where the subjects were shown the questions already asked about an image, and explicitly instructed to ask about *different entities* [4].

### 4.2. Analyzing VisDial Answers

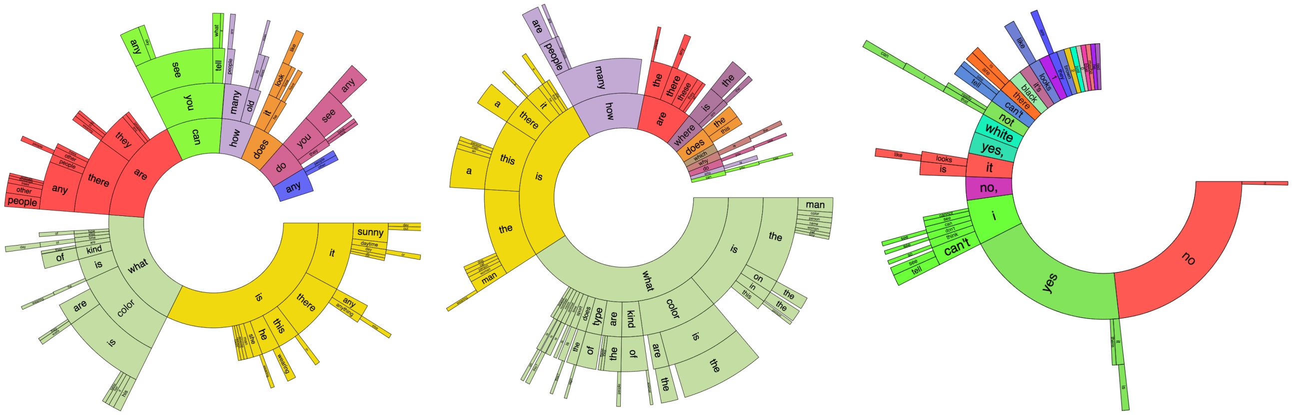
**Answer Lengths.** Fig. 7a shows the distribution of answer lengths. Unlike previous datasets, answers in VisDial are longer and more descriptive – mean-length 3.1 words (VisDial) vs 1.1 (VQA), 2.0 (Visual 7W), 2.8 (Visual Madlibs).

Fig. 7b shows the cumulative coverage of all answers (y-axis) by the most frequent answers (x-axis). The difference between VisDial and VQA is stark – the top-1000 answers in VQA cover ∼83% of all answers, while in VisDial that figure is only ∼58%. There is a significant heavy tail in VisDial – most long strings are unique, and thus the coverage curve in Fig. 7b becomes a straight line with slope 1. In total, there are 171,502 unique answers in VisDial.

**Answer Types.** Since the answers in VisDial are longer strings, we can visualize their distribution based on the starting few words (Fig. 9). An interesting category of answers emerges – *'I think so'*, *'I can't tell'*, or *'I can't see'* – expressing doubt, uncertainty, or lack of information. This is a consequence of the questioner not being able to see the image – they are asking contextually relevant questions, but not all questions may be answerable with certainty from that image. We believe this is rich data for building more human-like AI that refuses to answer questions it doesn't have enough information to answer. See [40] for a related, but complementary effort on question relevance in VQA.

**Binary Questions vs Binary Answers.** In VQA, binary questions are simply those with 'yes', 'no', 'maybe' as answers [4]. In VisDial, we must distinguish between binary questions and binary answers. Binary questions are those starting in 'Do', 'Did', 'Have', 'Has', 'Is', 'Are', 'Was', 'Were', 'Can', 'Could'. Answers to such questions can (1) contain only 'yes' or 'no', (2) begin with 'yes', 'no', and contain additional information or clarification, (3) involve

[2] Questions starting in 'Do', 'Did', 'Have', 'Has', 'Is', 'Are', 'Was', 'Were', 'Can', 'Could'.




(a) VisDial Questions (b) VQA Questions (c) VisDial Answers

Figure 5: Distribution of first n-grams for (left to right) VisDial questions, VQA questions and VisDial answers. The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

ambiguity (*'It's hard to see'*, *'Maybe'*), or (4) answer the question without explicitly saying 'yes' or 'no' (Q: *'Is there any type of design or pattern on the cloth?'*, A: *'There are circles and lines on the cloth'*). We call answers that contain 'yes' or 'no' as binary answers – 149,367 and 76,346 answers in subsets (1) and (2) from above respectively. Binary answers in VQA are biased towards 'yes' [4, 61] – 61.40% of yes/no answers are 'yes'. In VisDial, the trend is reversed. Only 46.96% are 'yes' for all yes/no responses. This is understandable since workers did not see the image, and were more likely to end up with negative responses.

### 4.3. Analyzing VisDial Dialog

In Section 4.1, we discussed a typical flow of dialog in VisDial. We analyze two quantitative statistics here.

**Coreference in dialog.** Since language in VisDial is the result of a sequential conversation, it naturally contains pronouns – 'he', 'she', 'his', 'her', 'it', 'their', 'they', 'this', 'that', 'those', *etc*. In total, 38% of questions, 22% of answers, and *nearly all* (99%) dialogs contain at least one pronoun, thus confirming that a machine will need to overcome coreference ambiguities to be successful on this task. We find that pronoun usage is low in the first round (as expected) and then picks up in frequency. A fine-grained per-round analysis is available in the supplement.

**Temporal Continuity in Dialog Topics.** It is natural for conversational dialog data to have continuity in the 'topics' being discussed. We have already discussed qualitative differences in VisDial questions *vs*. VQA. In order to quantify the differences, we performed a human study where we manually annotated question 'topics' for $40$ images (a total of $400$ questions), chosen randomly from the `val` set. The topic annotations were based on human judgement with a consensus of 4 annotators, with topics such as: asking about a particular object (*'What is the man doing?'*) , scene (*'Is it outdoors or indoors?'*), weather (*"Is the weather sunny?'*), the image (*'Is it a color image?'*), and exploration (*'Is there anything else?"*). We performed similar topic annotation for questions from VQA for the same set of $40$ images, and compared topic continuity in questions. Across $10$ rounds, VisDial question have $4.55 \pm 0.17$ topics on average, confirming that these are not independent questions. Recall that VisDial has $10$ questions per image as opposed to 3 for VQA. Therefore, for a fair comparison, we compute average number of topics in VisDial over all subsets of 3 successive questions. For $500$ bootstrap samples of batch size $40$, VisDial has $2.14 \pm 0.05$ topics while VQA has $2.53 \pm 0.09$. Lower mean suggests there is more continuity in VisDial because questions do not change topics as often.

### 4.4. VisDial Evaluation Protocol

One fundamental challenge in dialog systems is evaluation. If the response by a system at a particular round is incorrect, how does the dialog proceed? Moreover, similar to the state of affairs in captioning and machine translation, it is an open problem to automatically evaluate the quality of long free-form answers since existing metrics such as BLEU, METEOR, and ROUGE are known to correlate poorly with human judgement in evaluating dialog responses [26].

Instead of evaluating success on a downstream task [7] or holistically evaluating the entire conversation [3] (as in goal-free chit-chat), we evaluate *individual responses* independently at each round ($t = 1, 2, \ldots, 10$) in a retrieval or multiple-choice setup.

Specifically, at test time, a VisDial system is given an image $I$, the 'ground-truth' dialog history (including the image caption) $C, (Q_1, A_1), \ldots, (Q_{t-1}, A_{t-1})$, the question $Q_t$, *and* a list of $N = 100$ candidate answers, and asked to return a sorting of the candidate answers. The model is evaluated on retrieval metrics – (1) rank of human response (lower is better), (2) recall@$k$, *i.e.* existence of the human response in top-$k$ ranked responses, and (3) mean reciprocal rank (MRR) of the human response (higher is better).

The evaluation protocol is compatible with both discriminative models (that simply score the input candidates, *e.g.* via a softmax over the options, and cannot generate new answers), and generative models (that generate an answer string, *e.g.* via Recurrent Neural Networks) by ranking the candidates by the model's log-likelihood scores.

**Candidate Answers.** We generate a candidate set of correct and incorrect answers from four sets:
**Correct**: The ground-truth human response to the question.
**Plausible**: Answers to 50 most similar questions. Similar questions are those that start with similar tri-grams and mention similar semantic concepts in the rest of the question. To capture this, all questions are embedded into a vector space by concatenating the GloVe embeddings of the first three words with the averaged GloVe embeddings of the remaining words in the questions. Euclidean distances are used to compute neighbors. Since these neighboring questions were asked on different images, their answers serve as 'hard negatives'.
**Popular**: The 30 most popular answers from the dataset – *e.g.* 'yes', 'no', '2', '1', 'white', '3', 'grey', 'gray', '4', 'yes it is'. The inclusion of popular answers forces the machine to pick between likely *a priori* responses and plausible responses for the question, thus increasing the task difficulty.
**Random**: The remaining are answers to random questions in the dataset. To generate 100 candidates, we first find the union of the correct, plausible, and popular answers, and include random answers until a unique set of 100 is found.

## 5. Neural Visual Dialog Models

In this section, we develop a number of novel neural Visual Dialog models. Recall that the model is given as input – an image $I$, the 'ground-truth' dialog history (including the image caption) $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$, the question $Q_t$, and a list of 100 candidate answers $\mathcal{A}_t = \{A_t^{(1)}, \ldots, A_t^{(100)}\}$ – and asked to return a sorting of $\mathcal{A}_t$.

At a high level, all our models follow the encoder-decoder framework, *i.e.* factorize into two parts – (1) an encoder that converts the input $(I, H, Q_t)$ into a vector space, and (2) a decoder that converts the embedded vector into an output. We describe choices for each component next and present experiments with all encoder-decoder combinations.

**Decoders:** We use two types of decoders:

- **Generative** (LSTM) decoder: where the encoded vector is set as the initial state of the Long Short-Term Memory (LSTM) RNN language model. During training, we maximize the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end). To evaluate, we use the model's log-likelihood scores and rank candidate answers.
Note that this decoder does not need to score options during training. As a result, such models do not exploit the biases in option creation and typically underperform models that do [18], but it is debatable whether exploiting such biases is really indicative of progress. Moreover, generative decoders are more practical in that they can actually be deployed in realistic applications.
- **Discriminative** (softmax) decoder: computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options. These dot products are fed into a softmax to compute the posterior probability over the options. During training, we maximize the logliklihood of the correct option. During evaluation, options are simply ranked based on their posterior probabilities.

**Encoders:** We develop 3 different encoders (listed below) that convert inputs $(I, H, Q_t)$ into a joint representation. In all cases, we represent $I$ via the $\ell 2$-normalized activations from the penultimate layer of VGG-16 [48]. For each encoder $E$, we experiment with all possible ablated versions: $E(Q_t), E(Q_t, I), E(Q_t, H), E(Q_t, I, H)$ (for some encoders, not all combinations are 'valid'; details below).

- **Late Fusion (LF) Encoder**: In this encoder, we treat $H$ as a long string with the entire history $(H_0, \ldots, H_{t-1})$ concatenated. $Q_t$ and $H$ are separately encoded with 2 different LSTMs, and individual representations of participating inputs $(I, H, Q_t)$ are concatenated and linearly transformed to a desired size of joint representation.
- **Hierarchical Recurrent Encoder (HRE)**: In this encoder, we capture the intuition that there is a hierarchical nature to our problem – each question $Q_t$ is a sequence of words that need to be embedded, and the dialog as a whole is a sequence of question-answer pairs $(Q_t, A_t)$. Thus, similar to [46], as shown in Fig. 6, we propose an HRE model that contains a dialog-RNN sitting on top of a recurrent block ($R_t$). The recurrent block $R_t$ embeds the question and image jointly via an LSTM (early fusion), embeds each round of the history $H_t$, and passes a concatenation of these to the dialog-RNN above it. The dialog-RNN produces both an encoding for this round ($E_t$ in Fig. 6) and a dialog context to pass onto the next round. We also add an attention-over-history ('Attention' in Fig. 6) mechanism allowing the recurrent block $R_t$ to choose and attend to the round of the history relevant to the current question. This attention mechanism consists of a softmax over previous rounds $(0, 1, \ldots, t-1)$ computed from the history and question+image encoding.
- **Memory Network (MN) Encoder**: We develop a MN encoder that maintains each previous question and answer

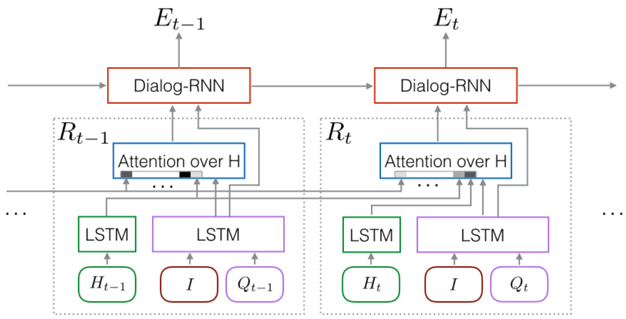



Figure 6: Architecture of HRE encoder with attention. At the current round $R_t$, the model has the capability to choose and attend to relevant history from previous rounds, based on the current question. This attention-over-history feeds into a dialog-RNN along with question to generate joint representation $E_t$ for the decoder.

as a 'fact' in its memory bank and learns to 'poll' the stored facts and the image to answer the question. Specifically, we encode $Q_t$ with an LSTM to get a 512-d vector, encode each previous round of history $(H_0, \ldots, H_{t-1})$ with another LSTM to get a $t \times 512$ matrix. We compute inner product of question vector with each history vector to get scores over previous rounds, which are fed to a softmax to get attention-over-history probabilities. Convex combination of history vectors using these attention probabilities gives us the 'context vector', which is passed through an fc-layer, added to the question vector, and followed by another fc-layer to construct the MN encoding. In the language of Memory Network [7], this is a '1-hop' encoding.

We use a '[encoder]-[input]-[decoder]' convention to refer to model-input combinations. For example, 'LF-QI-D' has a Late Fusion encoder with question+image inputs (no history), and a discriminative decoder. Implementation details about the models can be found in the supplement.

## 6. Experiments

**Splits.** VisDial v0.5 contains 58k dialogs on COCO-`train` and 10k on COCO-`val` images. We split the 58k into 50k for training, 8k for validation, and use the 10k as test.

Data preprocessing, choice of hyperparameters and training details are included in the supplement.

**Baselines** We compare to a number of baselines: **Answer Prior:** Answer options to a test question are encoded with an LSTM and scored by a linear classifier. This captures ranking by frequency of answers in our training set without resolving to exact string matching. **NN-Q:** Given a test question, we find $k$ nearest neighbor questions (in GloVe space) from train, and score answer options by their mean-similarity with these $k$ answers. **NN-QI:** First, we find $K$ nearest neighbor questions for a test question. Then, we find a subset of size $k$ based on image feature similarity. Finally, we rank options by their mean-similarity to answers to these $k$ questions. We use $k = 20, K = 100$.

Finally, we adapt several (near) state-of-art VQA models (SAN [59], HieCoAtt [30]) to Visual Dialog. Since VQA is posed as classification, we 'chop' the final VQA-answer softmax from these models, feed these activations to our discriminative decoder (Section 5), and train end-to-end on VisDial. Note that our LF-QI-D model is similar to that in [29]. Altogether, these form fairly sophisticated baselines.

| | Model | MRR | R@1 | R@5 | R@10 | Mean |
|---|---|---|---|---|---|---|
| Baseline | Answer prior | 0.311 | 19.85 | 39.14 | 44.28 | 31.56 |
| | NN-Q | 0.392 | 30.54 | 46.99 | 49.98 | 30.88 |
| | NN-QI | 0.385 | 29.71 | 46.57 | 49.86 | 30.90 |
| Generative | LF-Q-G | 0.403 | 29.74 | 50.10 | 56.32 | 24.06 |
| | LF-QH-G | 0.425 | 32.49 | 51.56 | 57.80 | 23.11 |
| | LF-QI-G | 0.437 | 34.06 | 52.50 | 58.89 | 22.31 |
| | LF-QIH-G | 0.430 | 33.27 | 51.96 | 58.09 | 23.04 |
| | HRE-QH-G | 0.430 | 32.84 | 52.36 | 58.64 | 22.59 |
| | HRE-QIH-G | 0.442 | 34.37 | 53.40 | 59.74 | 21.75 |
| | HREA-QIH-G | 0.442 | 34.47 | 53.43 | 59.73 | 21.83 |
| | MN-QH-G | 0.434 | 33.12 | 53.14 | 59.61 | 22.14 |
| | MN-QIH-G | **0.443** | **34.62** | **53.74** | **60.18** | **21.69** |
| Discriminative | LF-Q-D | 0.482 | 34.29 | 63.42 | 74.31 | 8.87 |
| | LF-QH-D | 0.505 | 36.21 | 66.56 | 77.31 | 7.89 |
| | LF-QI-D | 0.502 | 35.76 | 66.59 | 77.61 | 7.72 |
| | LF-QIH-D | 0.511 | 36.72 | 67.46 | 78.30 | 7.63 |
| | HRE-QH-D | 0.489 | 34.74 | 64.25 | 75.40 | 8.32 |
| | HRE-QIH-D | 0.502 | 36.26 | 65.67 | 77.05 | 7.79 |
| | HREA-QIH-D | 0.508 | 36.76 | 66.54 | 77.75 | 7.59 |
| | MN-QH-D | 0.524 | 36.84 | 67.78 | 78.92 | 7.25 |
| | MN-QIH-D | **0.529** | **37.33** | **68.47** | **79.54** | **7.03** |
| VQA | SAN1-QI-D | 0.506 | 36.21 | 67.08 | 78.16 | 7.74 |
| | HieCoAtt-QI-D | 0.509 | 35.54 | 66.79 | 77.94 | 7.68 |
| | Human Accuracies | | | | | |
| Human | Human-Q | 0.441 | 25.10 | 67.37 | - | 4.19 |
| | Human-QH | 0.485 | 30.31 | 70.53 | - | 3.91 |
| | Human-QI | 0.619 | 46.12 | 82.54 | - | 2.92 |
| | Human-QIH | 0.635 | 48.03 | 83.76 | - | 2.83 |

Table 1: Performance of methods on our VisDial dataset, measured by mean reciprocal rank (MRR), recall@$k$ for $k = \{1, 5, 10\}$ and mean rank. Note that higher is better for MRR and recall@k, while lower is better for mean rank. Memory Network has the best performance in both discriminative and generative settings.

**Results.** Tab. 1 shows the results for our proposed models and baselines. A few key takeaways – First, as expected, all learning based models significantly outperform non-learning baselines. Second, all discriminative models significantly outperform generative models, which as we discussed is expected since discriminative models can tune to the biases in the answer options. This improvement comes with the significant limitation of not being able to actually generate responses, and we recommend the two decoders be viewed as separate use cases. Third, our best generative and discriminative models are MN-QIH-G with 0.44 MRR, and MN-QIH-D with 0.53 MRR that outperform a suite of models and sophisticated baselines. Fourth, we observe that models with $H$ perform better than $Q$-only models, highlighting the importance of history in VisDial. Additionally, modeling history through HRE improves performance of generative decoders, while discriminative decoders do not seem to benefit much. Fifth, models looking at $I$ significantly outperform both the blind models ($Q$, $QH$) by at least $2\%$ on recall@1 in both decoders. This improvement confirms that VisDial is indeed grounded in the image. Finally, models that use both $H$ and $I$ have best performance. The gap between $QIH$ and $Q$-only blind models is only likely to increase as VisDial becomes larger.

**Human Studies.** We conduct studies on AMT to quantitatively evaluate human performance on this task for all combinations of {with image, without image}$\times${with history, without history}. Specifically, we show humans top-9 predicted respones from our 'LF-QIH-D' model and ground truth answer in jumbled order, and ask them to rank responses. Results of our study are included in Tab. 1. Note that these numbers are not directly comparable to machine performance as models are tasked with ranking 100 responses, as opposed to 10 for humans. The former would be too cumbersome for humans. We find that without image, humans perform better when they have access to dialog history. As expected, this gap narrows down when they have access to the image. Explicit human-machine comparison can be found in supplement.

## 7. Conclusions

To summarize, we introduce a new AI task – Visual Dialog, where an AI agent must hold a dialog with a human about visual content. We develop a novel two-person chat data-collection protocol to curate a large-scale dataset (VisDial), propose retrieval-based evaluation protocol, and develop a family of encoder decoder models for Visual Dialog with novel encoders, which model the particular features of this task and consequently outperform sophisticated baselines. We conduct human studies to quantify human performance on this task. Putting it all together, in the supplement we demonstrate the first visual chatbot. Our results and analysis indicates that there is significant scope for improvement, and we believe this task can serve as a testbed for measuring progress towards visual intelligence.

**Acknowledgements.** We thank Harsh Agrawal and Jiasen Lu for help on the AMT data collection interface. We also thank Xiao Lin, Ramprasaath Selvaraju and Latha Pemula for model discussions. Finally, we are grateful to the developers of Torch [2] for building an excellent framework. This work was funded in part by NSF CAREER awards to DP and DB, an ONR YIP award to DP, ONR Grant N00014-14-1-0679 to DB, a Sloan Fellowship to DP, ARO YIP awards to DB and DP, a Allen Distinguished Investigator award to DP from the Paul G. Allen Family Foundation, ICTAS Junior Faculty awards to DB and DP, Google Faculty Research Awards to DP and DB, Amazon Academic Research Awards to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donations to DB. SK was supported by ONR Grant N00014-12-1-0903. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

## Appendix Overview

This supplementary document is organized as follows:

- Sec. A studies how and why VisDial is more than just a collection of independent Q&As.
- Sec. B shows qualitative examples from our dataset.
- Sec. C presents detailed human studies along with comparisons to machine accuracy. The interface for human studies is demonstrated in a video[3].
- Sec. D shows snapshots of our two-person chat data-collection interface on Amazon Mechanical Turk. The interface is also demonstrated in the video[3].
- Sec. E presents further analysis of VisDial, such as question types, question and answer lengths per question type. A video with an interactive sunburst visualization of the dataset is included[3].
- Sec. F presents implementation-level training details including data preprocessing, and model architectures.
- Putting it all together, we compile a video demonstrating our visual chatbot[3] that answers a sequence of questions from a user about an image. This demo uses one of our best generative models from the main paper, MN-QIH-G, and uses sampling (without any beam-search) for inference in the LSTM decoder. Note that these videos demonstrate an 'unscripted' dialog – in the sense that the particular QA sequence is not present in VisDial and the model is not provided with any list of answer options.

# A. In what ways are dialogs in VisDial more than just 10 visual Q&As?

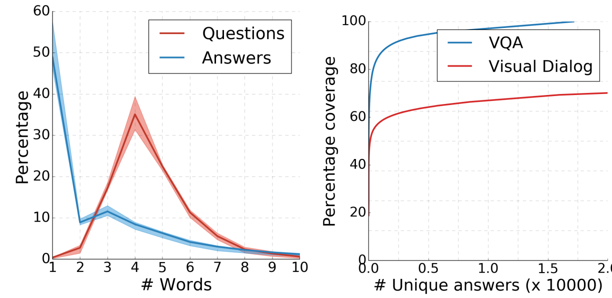
In this section, we lay out an exhaustive list of differences between VisDial and existing image question-answering datasets, with the VQA dataset [4] serving as the representative.

In essence, we characterize what makes an instance in VisDial more than a collection of 10 independent question-answer pairs about an image – *what makes it a dialog*.

In order to be self-contained and an exhaustive list, some parts of this section repeat content from the main document.

### A.1. VisDial has longer free-form answers

Fig. 7a shows the distribution of answer lengths in VisDial. and Tab. 2 compares statistics of VisDial with existing image question answering datasets. Unlike previous datasets, answers in VisDial are longer, conversational, and more descriptive – mean-length 3.1 words (VisDial) vs 1.1 (VQA), 2.0 (Visual 7W), 2.8 (Visual Madlibs). Moreover, 42.2% of answers in VisDial are longer than 2 words while the VQA dataset has only 3.8% answers longer than 2 words.



(a) (b)

Figure 7: Distribution of lengths for questions and answers (left); and percent coverage of unique answers over all answers from the train dataset (right), compared to VQA. For a given coverage, VisDial has more unique answers indicating greater answer diversity.

Fig. 7b shows the cumulative coverage of all answers (y-axis) by the most frequent answers (x-axis). The difference between VisDial and VQA is stark – the top-1000 answers in VQA cover ~83% of all answers, while in VisDial that figure is only ~58%. There is a significant heavy tail of answers in VisDial – most long strings are unique, and thus the coverage curve in Fig. 7b becomes a straight line with slope 1. In total, there are 171,502 unique answers in VisDial (out of the 680,000 answers currently in the dataset).

### A.2. VisDial has co-references in dialogs

People conversing with each other tend to use pronouns to refer to already mentioned entities. Since language in VisDial is the result of a sequential conversation, it naturally contains pronouns – 'he', 'she', 'his', 'her', 'it', 'their', 'they', 'this', 'that', 'those', *etc.* In total, 38% of questions, 22% of answers, and *nearly all* (99%) dialogs contain at least one pronoun, thus confirming that a machine will need to overcome coreference ambiguities to be successful on this task. As a comparison, only 9% of questions and 0.25% of answers in VQA contain at least one pronoun.

In Fig. 8, we see that pronoun usage is lower in the first round compared to other rounds, which is expected since there are fewer entities to refer to in the earlier rounds. The pronoun usage is also generally lower in answers than questions, which is also understandable since the answers are generally shorter than questions and thus less likely to contain pronouns. In general, the pronoun usage is fairly consistent across rounds (starting from round 2) for both questions and answers.

[3] https://goo.gl/yjlHxY

| | # QA | # Images | Q Length | A Length | A Length > 2 | Top-1000 A | Human Accuracy |
|---|---|---|---|---|---|---|---|
| DAQUAR [31] | 12,468 | 1,447 | $11.5 \pm 2.4$ | $1.2 \pm 0.5$ | 3.4% | 96.4% | - |
| Visual Madlibs [60] | 56,468 | 9,688 | $4.9 \pm 2.4$ | $2.8 \pm 2.0$ | 47.4% | 57.9% | - |
| COCO-QA [41] | 117,684 | 69,172 | $8.7 \pm 2.7$ | $1.0 \pm 0$ | 0.0% | 100% | - |
| Baidu [12] | 316,193 | 316,193 | - | - | - | - | - |
| VQA [4] | 614,163 | 204,721 | $6.2 \pm 2.0$ | $1.1 \pm 0.4$ | 3.8% | 82.7% | ✓ |
| Visual7W [62] | 327,939 | 47,300 | $6.9 \pm 2.4$ | $2.0 \pm 1.4$ | 27.6% | 63.5% | ✓ |
| VisDial (Ours) | 680,000 | 68,000 | $4.64 \pm 0$ | $3.08 \pm 0$ | 42.2% | 58.1% | ✓ |

Table 2: Comparison of existing image question answering datasets with VisDial

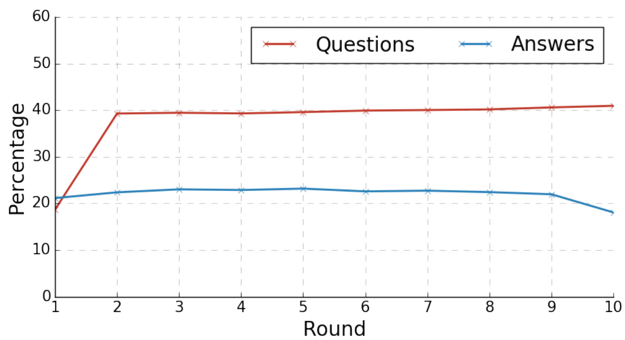



Figure 8: Percentage of QAs with pronouns for different rounds. In round 1, pronoun usage in questions is low (in fact, slightly lower than usage in answers). From rounds 2 through 10, pronoun usage is higher in questions and fairly consistent across rounds.

### A.3. VisDial has smoothness/continuity in 'topics'

**Qualitative Example of Topics.** There is a stylistic difference in the questions asked in VisDial (compared to the questions in VQA) due to the nature of the task assigned to the subjects asking the questions. In VQA, subjects saw the image and were asked to "stump a smart robot". Thus, most queries involve specific details, often about the background (Q: *'What program is being utilized in the background on the computer?'*). In VisDial, questioners did not see the original image and were asking questions to build a mental model of the scene. Thus, the questions tend to be open-ended, and often follow a pattern:

- Generally starting with the **entities in the caption**:
  *'An elephant walking away from a pool in an exhibit'*, *'Is there only 1 elephant?'*,
- digging deeper into their **parts, attributes, or properties**:
  *'Is it full grown?'*, *'Is it facing the camera?'*,
- asking about the **scene category or the picture setting**:
  *'Is this indoors or outdoors?'*, *'Is this a zoo?'*,
- **the weather**: *'Is it snowing?'*, *'Is it sunny?'*,
- simply **exploring the scene**:
  *'Are there people?'*, *'Is there shelter for elephant?'*,
- and asking **follow-up questions** about the new visual entities discovered from these explorations:
  *'There's a blue fence in background, like an enclosure'*, *'Is the enclosure inside or outside?'*.

Such a line of questioning does not exist in the VQA dataset, where the subjects were shown the questions already asked about an image, and explicitly instructed to ask about *different entities* [4].

**Counting the Number of Topics.** In order to quantify these qualitative differences, we performed a human study where we manually annotated question 'topics' for 40 images (a total of 400 questions), chosen randomly from the `val` set. The topic annotations were based on human judgement with a consensus of 4 annotators, with topics such as: asking about a particular object (*'What is the man doing?'*), the scene (*'Is it outdoors or indoors?'*), the weather (*"Is the weather sunny?'*), the image (*'Is it a color image?'*), and exploration (*'Is there anything else?"*). We performed similar topic annotation for questions from VQA for the same set of 40 images, and compared topic continuity in questions.

Across 10 rounds, VisDial questions have $4.55 \pm 0.17$ topics on average, confirming that these are not 10 independent questions. Recall that VisDial has 10 questions per image as opposed to 3 for VQA. Therefore, for a fair comparison, we compute average number of topics in VisDial over all 'sliding windows' of 3 successive questions. For 500 bootstrap samples of batch size 40, VisDial has $2.14 \pm 0.05$ topics while VQA has $2.53 \pm 0.09$. Lower mean number of topics suggests there is more continuity in VisDial because questions do not change topics as often.

**Transition Probabilities over Topics.** We can take this analysis a step further by computing topic transition probabilities over topics as follows. For a given sequential dialog exchange, we now count the number of topic transitions between consecutive QA pairs, normalized by the total number of possible transitions between rounds (9 for VisDial and 2 for VQA). We compute this 'topic transition probability' (how likely are two successive QA pairs to be about two different topics) for VisDial and VQA in two different settings – (1) in-order and (2) with a permuted sequence of QAs. Note that if VisDial were simply a collection of

10 independent QAs as opposed to a dialog, we would expect the topic transition probabilities to be similar for in-order and permuted variants. However, we find that for 1000 permutations of 40 topic-annotated image-dialogs, in-order-VisDial has an average topic transition probability of $0.61$, while permuted-VisDial has $0.76 \pm 0.02$. In contrast, VQA has a topic transition probability of $0.80$ for in-order *vs*. $0.83 \pm 0.02$ for permuted QAs.

There are two key observations: (1) In-order transition probability is lower for VisDial than VQA (*i.e*. topic transition is less likely in VisDial), and (2) Permuting the order of questions results in a larger increase for VisDial, around $0.15$, compared to a mere $0.03$ in case of VQA (*i.e*. in-order-VQA and permuted-VQA behave significantly more similarly than in-order-VisDial and permuted-VisDial).

Both these observations establish that there is smoothness in the temporal order of topics in VisDial, which is indicative of the narrative structure of a dialog, rather than independent question-answers.

### A.4. VisDial has the statistics of an NLP dialog dataset

In this analysis, our goal is to measure whether VisDial *behaves like a dialog dataset*.

In particular, we compare VisDial, VQA, and Cornell Movie-Dialogs Corpus [8]. The Cornell Movie-Dialogs corpus is a text-only dataset extracted from pairwise interactions between characters from approximately 617 movies, and is widely used as a standard dialog corpus in the natural language processing (NLP) and dialog communities.

One popular evaluation criteria used in the dialog-systems research community is the *perplexity* of language models trained on dialog datasets – the lower the perplexity of a model, the better it has learned the structure in the dialog dataset.

For the purpose of our analysis, we pick the popular sequence-to-sequence (Seq2Seq) language model [17] and use the perplexity of this model trained on different datasets as a measure of temporal structure in a dataset.

As is standard in the dialog literature, we train the Seq2Seq model to predict the probability of utterance $U_t$ given the previous utterance $U_{t-1}$, *i.e.* $\mathbf{P}(U_t \mid U_{t-1})$ on the Cornell corpus. For VisDial and VQA, we train the Seq2Seq model to predict the probability of a question $Q_t$ given the previous question-answer pair, *i.e.* $\mathbf{P}(Q_t \mid (Q_{t-1}, A_{t-1}))$.

For each dataset, we used its `train` and `val` splits for training and hyperparameter tuning respectively, and report results on `test`. At test time, we only use conversations of length 10 from Cornell corpus for a fair comparison to VisDial (which has 10 rounds of QA).

For all three datasets, we created 100 permuted versions of `test`, where either QA pairs or utterances are randomly shuffled to disturb their natural order. This allows us to compare datasets in their natural ordering w.r.t. permuted orderings. Our hypothesis is that since dialog datasets have linguistic structure in the sequence of QAs or utterances they contain, this structure will be significantly affected by permuting the sequence. In contrast, a collection of independent question-answers (as in VQA) will not be significantly affected by a permutation.

| **Dataset** | **Perplexity Per Token** | | **Classification** |
|---|---|---|---|
| | Orig | Shuffled | |
| VQA | 7.83 | 8.16 ± 0.02 | 52.8 ± 0.9 |
| Cornell (10) | 82.31 | 85.31 ± 1.51 | 61.0 ± 0.6 |
| VisDial (Ours) | 6.61 | 7.28 ± 0.01 | 73.3 ± 0.4 |

Table 3: Comparison of sequences in VisDial, VQA, and Cornell Movie-Dialogs corpus in their original ordering *vs*. permuted 'shuffled' ordering. Lower is better for perplexity while higher is better for classification accuracy. Left: the absolute increase in perplexity from natural to permuted ordering is highest in the Cornell corpus (3.0) followed by VisDial with 0.7, and VQA at $0.35$, which is indicative of the degree of linguistic structure in the sequences in these datasets. Right: The accuracy of a simple threshold-based classifier trained to differentiate between the original sequences and their permuted or shuffled versions. A higher classification rate indicates the existence of a strong temporal continuity in the conversation, thus making the ordering important. We can see that the classifier on VisDial achieves the highest accuracy (73.3%), followed by Cornell (61.0%). Note that this is a binary classification task with the prior probability of each class by design being equal, thus chance performance is 50%. The classifier on VQA performs close to chance.

Tab. 3 compares the original, unshuffled `test` with the shuffled testsets on two metrics:

**Perplexity:** We compute the standard metric of *perplexity per token*, *i.e*. exponent of the normalized negative-log-probability of a sequence (where normalized is by the length of the sequence). Tab. 3 shows these perplexities for the original unshuffled `test` and permuted `test` sequences.

We notice a few trends.

First, we note that the absolute perplexity values are higher for the Cornell corpus than QA datasets. We hypothesize that this is due to the broad, unrestrictive dialog generation task in Cornell corpus, which is a more difficult task than question prediction about images, which is in comparison a more restricted task.

Second, in all three datasets, the shuffled `test` has statistically significant higher perplexity than the original `test`, which indicates that shuffling does indeed break the linguistic structure in the sequences.

Third, the absolute increase in perplexity from natural to permuted ordering is highest in the Cornell corpus (3.0) followed by our VisDial with 0.7, and VQA at 0.35, which is indicative of the degree of linguistic structure in the sequences in these datasets. Finally, the relative increases in perplexity are 3.64% in Cornell, 10.13% in VisDial, and 4.21% in VQA – VisDial suffers the highest relative increase in perplexity due to shuffling, indicating the existence of temporal continuity that gets disrupted due to shuffling.

**Classification:** As our second metric to compare datasets in their natural *vs*. permuted order, we test whether we can reliably classify a given sequence as natural or permuted.

Our classifier is a simple threshold on perplexity of a sequence. Specifically, given a pair of sequences, we compute the perplexity of both from our Seq2Seq model, and predict that the one with higher perplexity is the sequence in permuted ordering, and the sequence with lower perplexity is the one in natural ordering. The accuracy of this simple classifier indicates how easy or difficult it is to tell the difference between natural and permuted sequences. A higher classification rate indicates the existence of a strong temporal continuity in the conversation, thus making the ordering important.

Tab. 3 shows the classification accuracies achieved on all datasets. We can see that the classifier on VisDial achieves the highest accuracy (73.3%), followed by Cornell (61.0%). Note that this is a binary classification task with the prior probability of each class by design being equal, thus chance performance is 50%. The classifiers on VisDial and Cornell both significantly outperforming chance. On the other hand, the classifier on VQA is near chance (52.8%), indicating a lack of general temporal continuity.

To summarize this analysis, our experiments show that VisDial is significantly more dialog-like than VQA, and *behaves* more like a standard dialog dataset, the Cornell Movie-Dialogs corpus.

### A.5. VisDial eliminates visual priming bias in VQA

One key difference between VisDial and previous image question answering datasets (VQA [4], Visual 7W [62], Baidu mQA [12]) is the lack of a 'visual priming bias' in VisDial. Specifically, in all previous datasets, subjects saw an image while asking questions about it. As described in [61], this leads to a particular bias in the questions – people only ask *'Is there a clocktower in the picture?'* on pictures actually containing clock towers. This allows language-only models to perform remarkably well on VQA and results in an inflated sense of progress [61]. As one particularly perverse example – for questions in the VQA dataset starting with *'Do you see a . . . '*, blindly answering *'yes'* without reading the rest of the question or looking at the associated image results in an average VQA accuracy of 87%! In VisDial, questioners *do not* see the image. As a result, this bias is reduced.

This lack of visual priming bias (*i.e*. not being able to see the image while asking questions) and holding a dialog with another person while asking questions results in the following two unique features in VisDial.

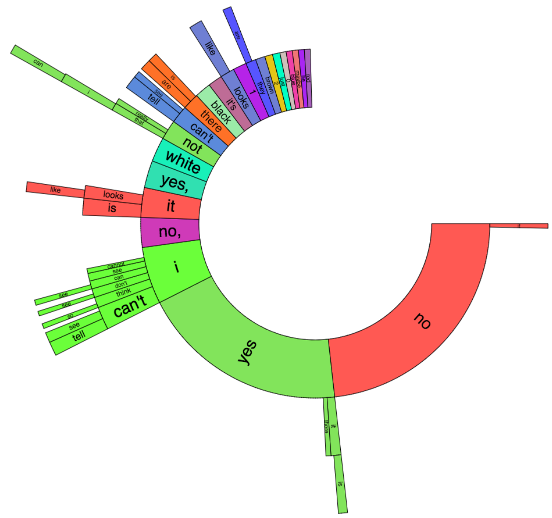



Figure 9: Distribution of answers in VisDial by their first four words. The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

**Uncertainty in Answers in VisDial.** Since the answers in VisDial are longer strings, we can visualize their distribution based on the starting few words (Fig. 9). An interesting category of answers emerges – *'I think so'*, *'I can't tell'*, or *'I can't see'* – expressing doubt, uncertainty, or lack of information. This is a consequence of the questioner not being able to see the image – they are asking contextually relevant questions, but not all questions may be answerable with certainty from that image. We believe this is rich data for building more human-like AI that refuses to answer questions it doesn't have enough information to answer. See [40] for a related, but complementary effort on question relevance in VQA.

**Binary Questions $\neq$ Binary Answers in VisDial.** In VQA, binary questions are simply those with 'yes', 'no', 'maybe' as answers [4]. In VisDial, we must distinguish between binary questions and binary answers. Binary questions are those starting in 'Do', 'Did', 'Have', 'Has', 'Is', 'Are', 'Was', 'Were', 'Can', 'Could'. Answers to such questions can (1) contain only 'yes' or 'no', (2) begin with 'yes', 'no', and contain additional information or clarification (Q: *'Are there any animals in the image?'*, A: *'yes, 2 cats and a dog'*), (3) involve ambiguity (*'It's hard to see'*, *'Maybe'*), or (4) answer the question without explicitly saying 'yes' or 'no' (Q: *'Is there any type of design or pattern on the cloth?'*, A: *'There are circles and lines on the cloth'*). We call answers that contain 'yes' or 'no' as binary answers – 149,367 and 76,346 answers in subsets (1) and (2) from above respectively. Binary answers in VQA are biased towards 'yes' [4,61] – 61.40% of yes/no answers are 'yes'. In VisDial, the trend is reversed. Only 46.96% are 'yes' for all yes/no responses. This is understandable since workers did not see the image, and were more likely to end up with negative responses.

## B. Qualitative Examples from VisDial

Fig. 10 shows random samples of dialogs from the VisDial dataset.

## C. Human-Machine Comparison

| | Model | MRR | R@1 | R@5 | Mean |
|---|---|---|---|---|---|
| Human | Human-Q | 0.441 | 25.10 | 67.37 | 4.19 |
| | Human-QH | 0.485 | 30.31 | 70.53 | 3.91 |
| | Human-QI | 0.619 | 46.12 | 82.54 | 2.92 |
| | Human-QIH | 0.635 | 48.03 | 83.76 | 2.83 |
| Machine | HREA-QIH-G | 0.477 | 31.64 | 61.61 | 4.42 |
| | MN-QIH-G | 0.481 | 32.16 | 61.94 | 4.47 |
| | MN-QIH-D | 0.553 | 36.86 | 69.39 | 3.48 |

Table 4: Human-machine performance comparison on our VisDial dataset, measured by mean reciprocal rank (MRR), recall@$k$ for $k = \{1, 5\}$ and mean rank. Note that higher is better for MRR and recall@k, while lower is better for mean rank.

We conducted studies on AMT to quantitatively evaluate human performance on this task for all combinations of {with image, without image}×{with history, without history} on 100 random images at each of the 10 rounds. Specifically, in each setting, we show human subjects a jumbled list of 10 candidate answers for a question – top-9 predicted responses from our 'LF-QIH-D' model and the 1 ground truth answer – and ask them to rank the responses. Each task was done by 3 human subjects.

Results of this study are shown in the top-half of Tab. 4. We find that without access to the image, humans perform better when they have access to dialog history – compare the Human-QH row to Human-Q (R@1 of 30.31 *vs*. 25.10). As perhaps expected, this gap narrows down when humans have access to the image – compare Human-QIH to Human-QI (R@1 of 48.03 *vs*. 46.12).

Note that these numbers are not directly comparable to machine performance reported in the main paper because models are tasked with ranking 100 responses, while humans are asked to rank 10 candidates. This is because the task of ranking 100 candidate responses would be too cumbersome for humans.

To compute comparable human and machine performance, we evaluate our best discriminative (MN-QIH-D) and generative (HREA-QIH-G, MN-QIH-G)[4] models on the same 10 options that were presented to humans. Note that in this setting, both humans and machines have R@10 = 1.0, since there are only 10 options.

Tab. 4 bottom-half shows the results of this comparison. We can see that, as expected, humans with full information (*i.e*. Human-QIH) perform the best with a large gap in human and machine performance (compare R@5: Human-QIH 83.76% *vs*. MN-QIH-D 69.39%). This gap is even larger when compared to generative models, which unlike the discriminative models are not actively trying to exploit the biases in the answer candidates (compare R@5: Human-QIH 83.76% *vs*. HREA-QIH-G 61.61%).

Furthermore, we see that humans outperform the best machine *even when not looking at the image*, simply on the basis of the context provided by the history (compare R@5: Human-QH 70.53% *vs*. MN-QIH-D 69.39%).

Perhaps as expected, with access to the image but not the history, humans are significantly better than the best machines (compare R@5: Human-QI 82.54% *vs*. MN-QIH-D 69.39%). With access to history humans perform even better.

From in-house human studies and worker feedback on AMT, we find that the dialog history plays the following roles for humans: (1) provides a context for the question and paints a picture of the scene, which helps eliminate certain answer choices (especially when the image is not available), (2) gives cues about the answerer's response style, which helps identify the right answer among similar answer choices, and (3) disambiguates amongst likely interpretations of the image (*i.e*., when objects are small or occluded), again, helping identify the right answer among multiple plausible options.

## D. Interface

In this section, we show our interface to connect two Amazon Mechanical Turk workers live, which we used to collect our data.

**Instructions.** To ensure quality of data, we provide detailed instructions on our interface as shown in Fig. 11a. Since the workers do not know their roles before starting the study, we provide instructions for both questioner and answerer roles.

[4] We use both HREA-QIH-G, MN-QIH-G since they have similar accuracies.

**Caption**: The skiers stood on top of the mountain.
**Person A (1)**: how many skiers are there
**Person B (1)**: hundreds
**Person A (2)**: are they getting ready to go downhill
**Person B (2)**: i think so my view is at end of line
**Person A (3)**: is it snowing
**Person B (3)**: no, there is lot of snow though
**Person A (4)**: can you see anybody going downhill
**Person B (4)**: no my view shows people going up small hill on skis i can't see what's going on from there
**Person A (5)**: do you see lift
**Person B (5)**: no
**Person A (6)**: can you tell if they are male or female
**Person B (6)**: skiers closest to me are male
**Person A (7)**: are there any children
**Person B (7)**: i don't see any but there could be it's huge crowd
**Person A (8)**: does anybody have hat on
**Person B (8)**: they all have winter hat of some sort on
**Person A (9)**: is sun shining
**Person B (9)**: yes, all blue sky
**Person A (10)**: do you see any clouds
**Person B (10)**: no clouds

(a)

**Caption**: an image of a man in a boat with a dog
**Person A (1)**: how old does man look
**Person B (1)**: he has white hair, so probably 60's
**Person A (2)**: can you tell what breed of dog
**Person B (2)**: it is dalmatian
**Person A (3)**: is this photo in color
**Person B (3)**: yes it is
**Person A (4)**: does this photo look like it is older photo or from more recent times
**Person B (4)**: it looks fairly recent
**Person A (5)**: is boat in water
**Person B (5)**: yes it is
**Person A (6)**: do you see any fishing equipment
**Person B (6)**: i do not
**Person A (7)**: is boat focal point of photo
**Person B (7)**: yes it is
**Person A (8)**: does water look smooth
**Person B (8)**: it is quite smooth
**Person A (9)**: is it sunny
**Person B (9)**: yes it is
**Person A (10)**: does this look like professional photo
**Person B (10)**: this is not professional photo

(b)

**Caption**: A man and woman on bicycles are looking at a map.
**Person A (1)**: where are they located
**Person B (1)**: in city
**Person A (2)**: are they on road
**Person B (2)**: sidewalk next to 1
**Person A (3)**: any vehicles
**Person B (3)**: 1 in background
**Person A (4)**: any other people
**Person B (4)**: no
**Person A (5)**: what color bikes
**Person B (5)**: 1 silver and 1 yellow
**Person A (6)**: do they look old or new
**Person B (6)**: new bikes
**Person A (7)**: any buildings
**Person B (7)**: yes
**Person A (8)**: what color
**Person B (8)**: brick
**Person A (9)**: are they tall or short
**Person B (9)**: i can't see enough of them to tell
**Person A (10)**: do they look like couple
**Person B (10)**: they are

(c)

**Caption**: a horse is pulling a white cart with two people a car and buildings
**Person A (1)**: is it day time
**Person B (1)**: i think it's day
**Person A (2)**: can you see other people
**Person B (2)**: no
**Person A (3)**: can you see trees
**Person B (3)**: yes
**Person A (4)**: can you see lot of buildings
**Person B (4)**: only 2 maybe 3
**Person A (5)**: does it look warm
**Person B (5)**: no, there are no leaves on trees
**Person A (6)**: do they nave jackets on
**Person B (6)**: no, long sleeve shirts though
**Person A (7)**: are they driving cart
**Person B (7)**: yes
**Person A (8)**: what color is car
**Person B (8)**: i can't tell photo is in black and white
**Person A (9)**: is building brick
**Person B (9)**: no, looks like wood
**Person A (10)**: do trees look old
**Person B (10)**: no they are still small

(d)


**Caption**: A statue depicting a bear breaking into a car.
**Person A (1)**: how big is statue
**Person B (1)**: about size of real full grown bear
**Person A (2)**: so is car full size then as well
**Person B (2)**: yes replica of car
**Person A (3)**: is statue all 1 color
**Person B (3)**: no brown and black
**Person A (4)**: what color is car
**Person B (4)**: dark red
**Person A (5)**: where is this, do you think
**Person B (5)**: in wooded area someplace
**Person A (6)**: do you see any people in image
**Person B (6)**: yes 1 man
**Person A (7)**: how old is man
**Person B (7)**: 35-40
**Person A (8)**: what is man doing
**Person B (8)**: sitting in car behind replica
**Person A (9)**: do you see any signs
**Person B (9)**: yes, on car door warning sign
**Person A (10)**: what else can you tell me about this image
**Person B (10)**: there are many trees in background

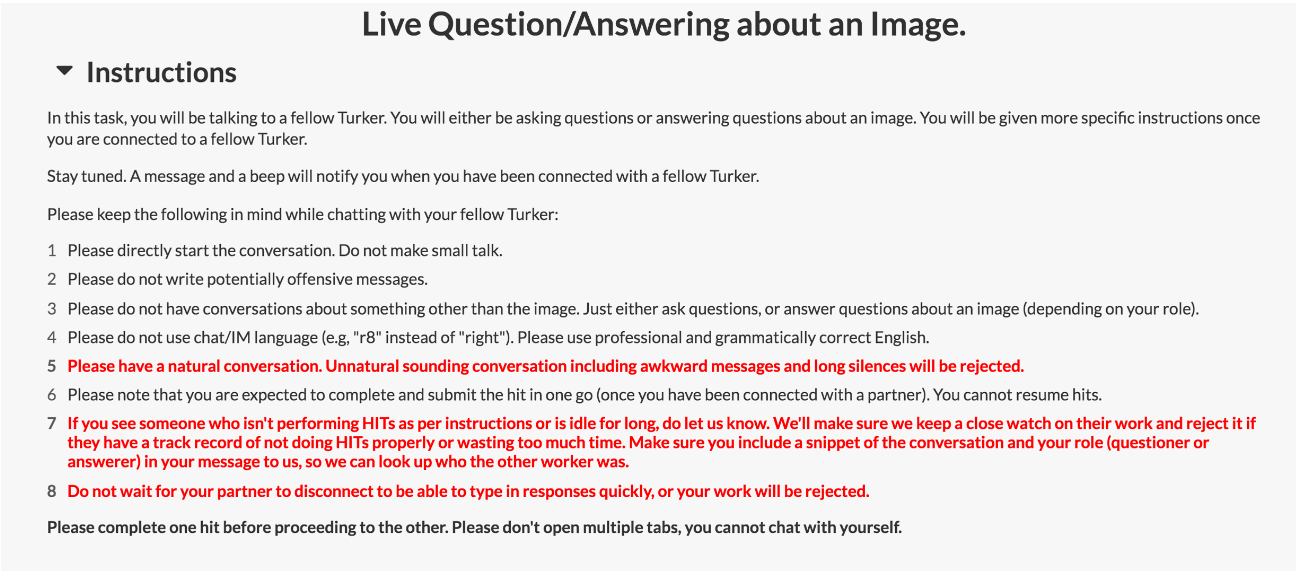
(e)

**Caption**: A dog with goggles is in a motorcycle side car.
**Person A (1)**: can you tell what kind of dog this is
**Person B (1)**: he looks like beautiful pit bull mix
**Person A (2)**: can you tell if motorcycle is moving or still
**Person B (2)**: it's parked
**Person A (3)**: is dog's tongue lolling out
**Person B (3)**: not really
**Person A (4)**: is picture in color
**Person B (4)**: yes it is
**Person A (5)**: what color is dog
**Person B (5)**: light tan with white patch that runs up to bottom of his chin and he has white paws on 2 front feet
**Person A (6)**: can you see motorcycle
**Person B (6)**: from side, yes
**Person A (7)**: what color is motorcycle
**Person B (7)**: black with white or silver accents, sun is glaring so it's hard to tell
**Person A (8)**: is there anybody sitting on motorcycle
**Person B (8)**: no
**Person A (9)**: is there anybody in picture
**Person B (9)**: in cars on street behind motorcycle
**Person A (10)**: does dog look like he's having fun
**Person B (10)**: yes

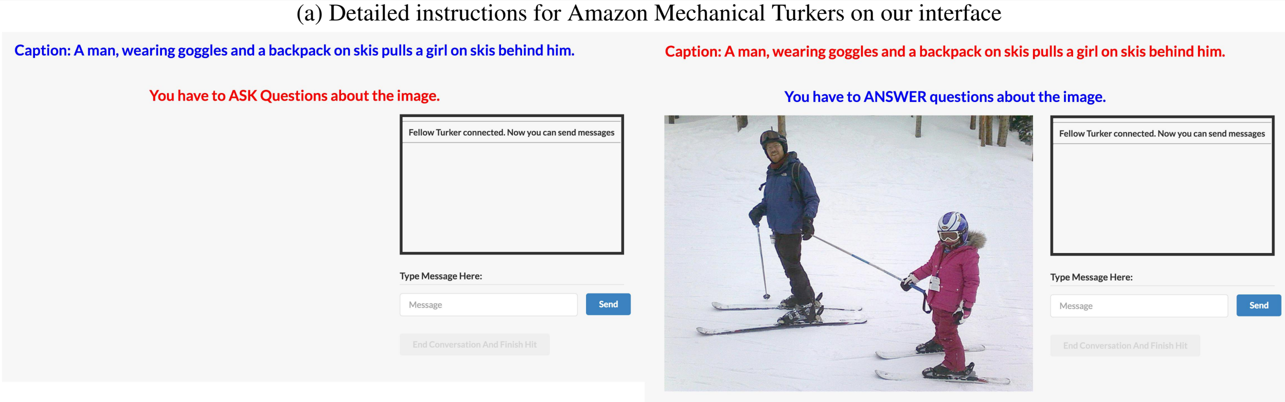
(f)

Figure 10: Examples from VisDial

**After pairing:** Immediately after pairing two workers, we assign them roles of a questioner and a answerer and display role-specific instructions as shown in Fig. 11b. Observe that the questioner does not see the image while the answerer




(a) Detailed instructions for Amazon Mechanical Turkers on our interface




(b) Left: What questioner sees; Right: What answerer sees.

Figure 11

does have access to it. Both questioner and answerer see the caption for the image.

# E. Additional Analysis of VisDial

In this section, we present additional analyses characterizing our VisDial dataset.

## E.1. Question and Answer Lengths

Fig. 12 shows question lengths by type and round. Average length of question by type is consistent across rounds. Questions starting with 'any' (*'any people?'*, *'any other fruits?'*, *etc.*) tend to be the shortest. Fig. 13 shows answer lengths by type of question they were said in response to and round. In contrast to questions, there is significant variance in answer lengths. Answers to binary questions (*'Any people?'*, *'Can you see the dog?'*, *etc.*) tend to be short while answers to 'how' and 'what' questions tend to be more explanatory and long. Across question types, answers tend to be the longest in the middle of conversations.

## E.2. Question Types

Fig. 14 shows round-wise coverage by question type. We see that as conversations progress, 'how' and 'what' questions reduce while 'can', 'does', 'do', 'any' questions occur more often. Questions starting with 'Is' are the most popular in the dataset.

# F. Experimental Details

In this section, we describe details about our models, data preprocessing, training procedure and hyperparameter selection.

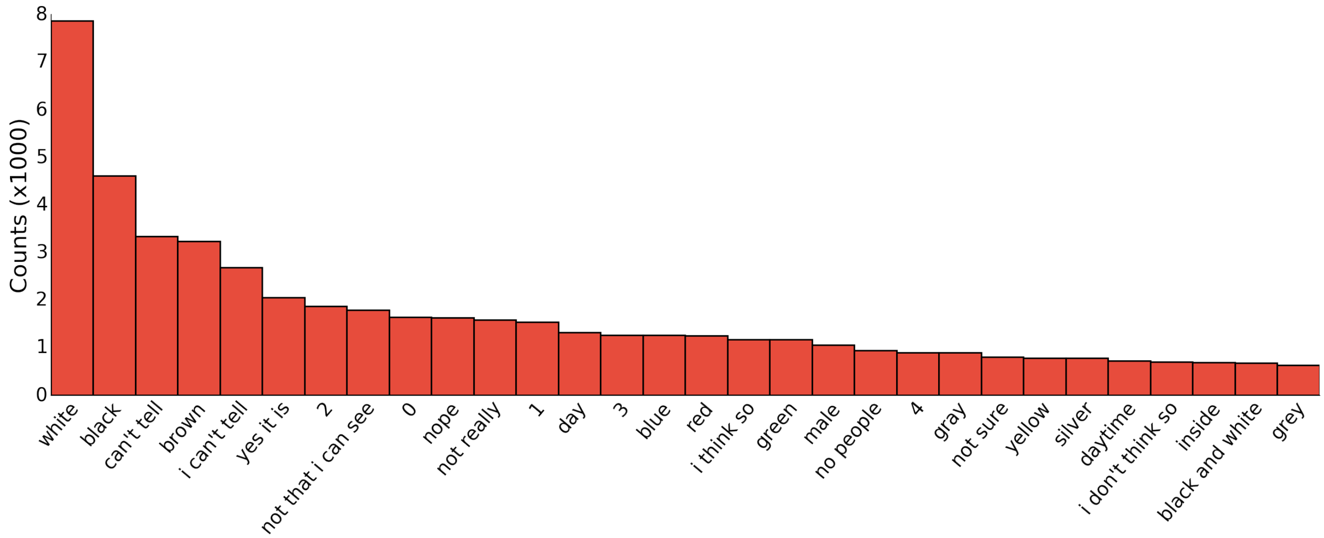



Figure 15: Most frequent answer responses except for 'yes'/'no'

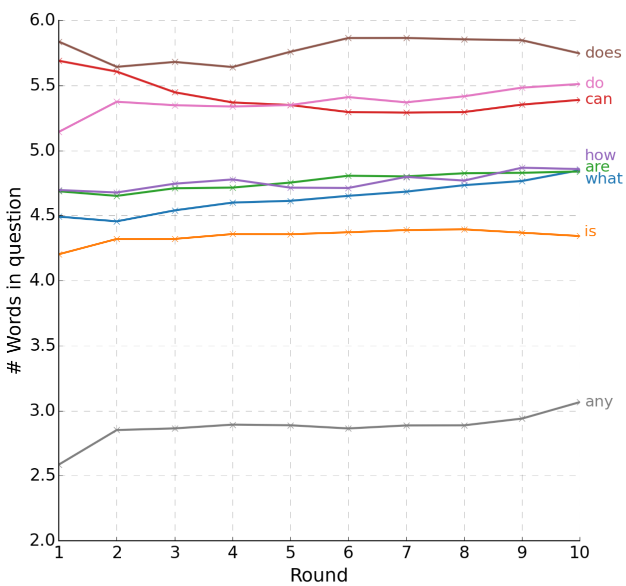



Figure 12: Question lengths by type and round. Average length of question by type is fairly consistent across rounds. Questions starting with 'any' ('any people?', 'any other fruits?', etc.) tend to be the shortest.

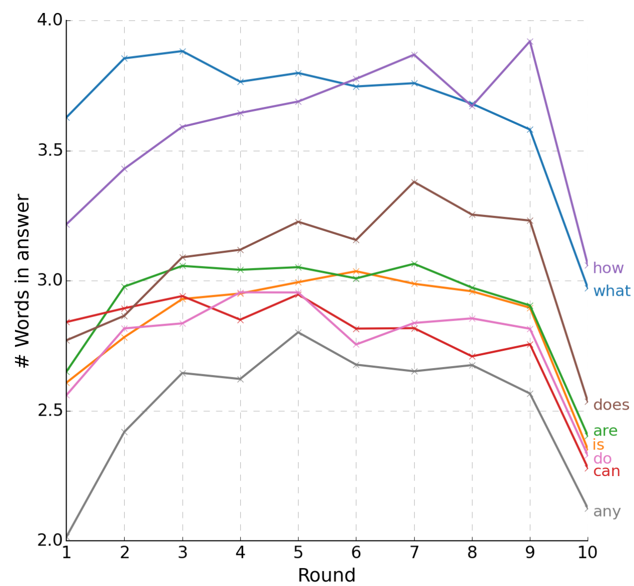



Figure 13: Answer lengths by question type and round. Across question types, average response length tends to be longest in the middle of the conversation.

### F.1. Models

**Late Fusion (LF) Encoder.** We encode the image with a VGG-16 CNN, question and concatenated history with separate LSTMs and concatenate the three representations. This is followed by a fully-connected layer and tanh non-linearity to a 512-d vector, which is used to decode the response. Fig. 16a shows the model architecture for our LF encoder.

**Hierarchical Recurrent Encoder (HRE).** In this encoder, the image representation from VGG-16 CNN is early fused with the question. Specifically, the image representation is concatenated with every question word as it is fed to an LSTM. Each QA-pair in dialog history is independently encoded by another LSTM with shared weights. The image-question representation, computed for every round from 1 through $t$, is concatenated with history representation from the previous round and constitutes a sequence of question-history vectors. These vectors are fed as input to a

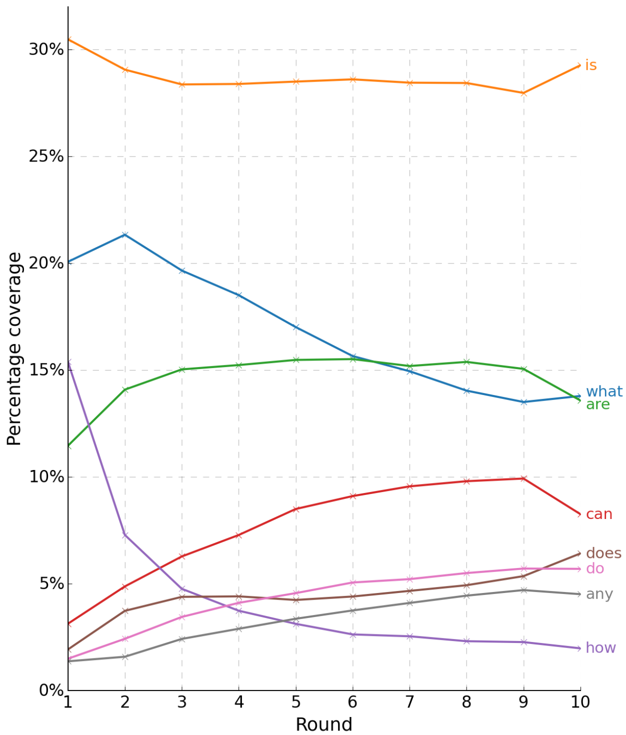



Figure 14: Percentage coverage of question types per round. As conversations progress, 'How' and 'What' questions reduce while 'Can', 'Does', 'Do', 'Any' questions occur more often. Questions starting with 'Is' are the most popular in the dataset.

dialog-level GRU, whose output state at $t$ is used to decode the response to $Q_t$. Fig. 16b shows the model architecture for our HRE.

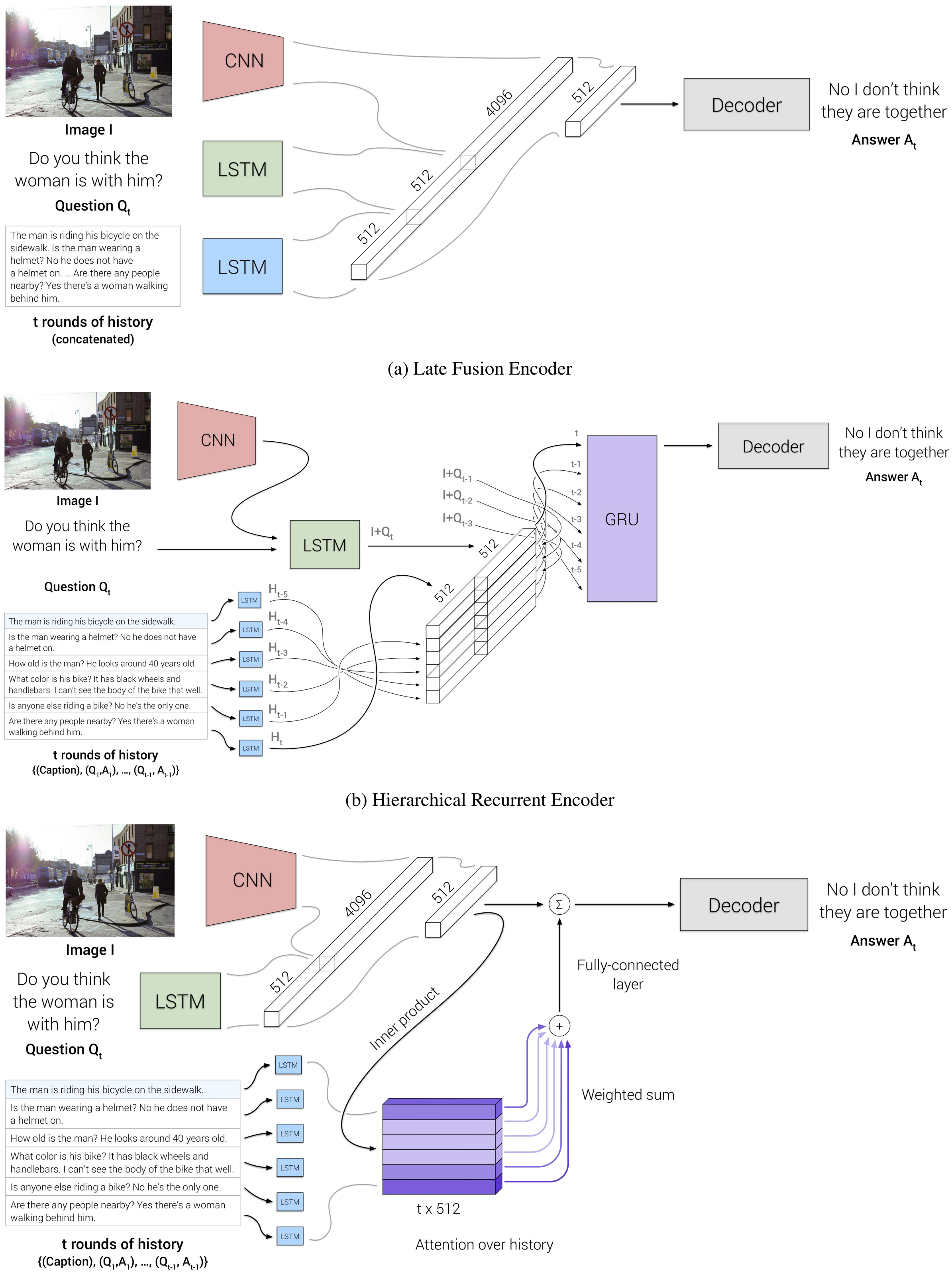
**Memory Network.** The image is encoded with a VGG-16 CNN and question with an LSTM. We concatenate the representations and follow it by a fully-connected layer and tanh non-linearity to get a 'query vector'. Each caption/QA-pair (or 'fact') in dialog history is encoded independently by an LSTM with shared weights. The query vector is then used to compute attention over the $t$ facts by inner product. Convex combination of attended history vectors is passed through a fully-connected layer and tanh non-linearity, and added back to the query vector. This combined representation is then passed through another fully-connected layer and tanh non-linearity and then used to decode the response. The model architecture is shown in Fig. 16c. Fig. 17 shows some examples of attention over history facts from our MN encoder. We see that the model learns to attend to facts relevant to the question being asked. For example, when asked 'What color are kites?', the model attends to 'A lot of people stand around flying kites in a park.' For 'Is anyone on bus?', it attends to 'A large yellow bus parked in some grass.' Note that these are selected examples, and not always are these attention weights interpretable.

### F.2. Training

**Splits.** Recall that VisDial v0.5 contained 58k dialogs on COCO-`train` and 10k on COCO-`val` images, at the time of submission. We split the 58k into 50k for training, 8k for validation, and use the 10k as test.

**Preprocessing.** We spell-correct VisDial data using the Bing API [34]. Following VQA, we lowercase all questions and answers, convert digits to words, and remove contractions, before tokenizing using the Python NLTK [1]. We then construct a dictionary of words that appear at least five times in the train set, giving us a vocabulary of around 7.5k.

**Hyperparameters.** All our models are implemented in Torch [2]. Model hyperparameters are chosen by early stopping on `val` based on the Mean Reciprocal Rank (MRR) metric. All LSTMs/GRUs are 2-layered with 512-dim hidden states. We learn 300-dim embeddings for words and images. These word embeddings are shared across question, history, and decoder LSTMs. For models with generative decoders, we use Adam [21] with a learning rate of $10^{-3}$. For discriminative decoders, we use SGD+Nesterov momentum with a learning rate of $0.1$. Gradients at each iterations are clamped to $[-5, 5]$ to avoid explosion. Our code, architectures, and trained models will be publicly available.




(a) Late Fusion Encoder



(b) Hierarchical Recurrent Encoder



(c) Memory Network Encoder

Figure 16

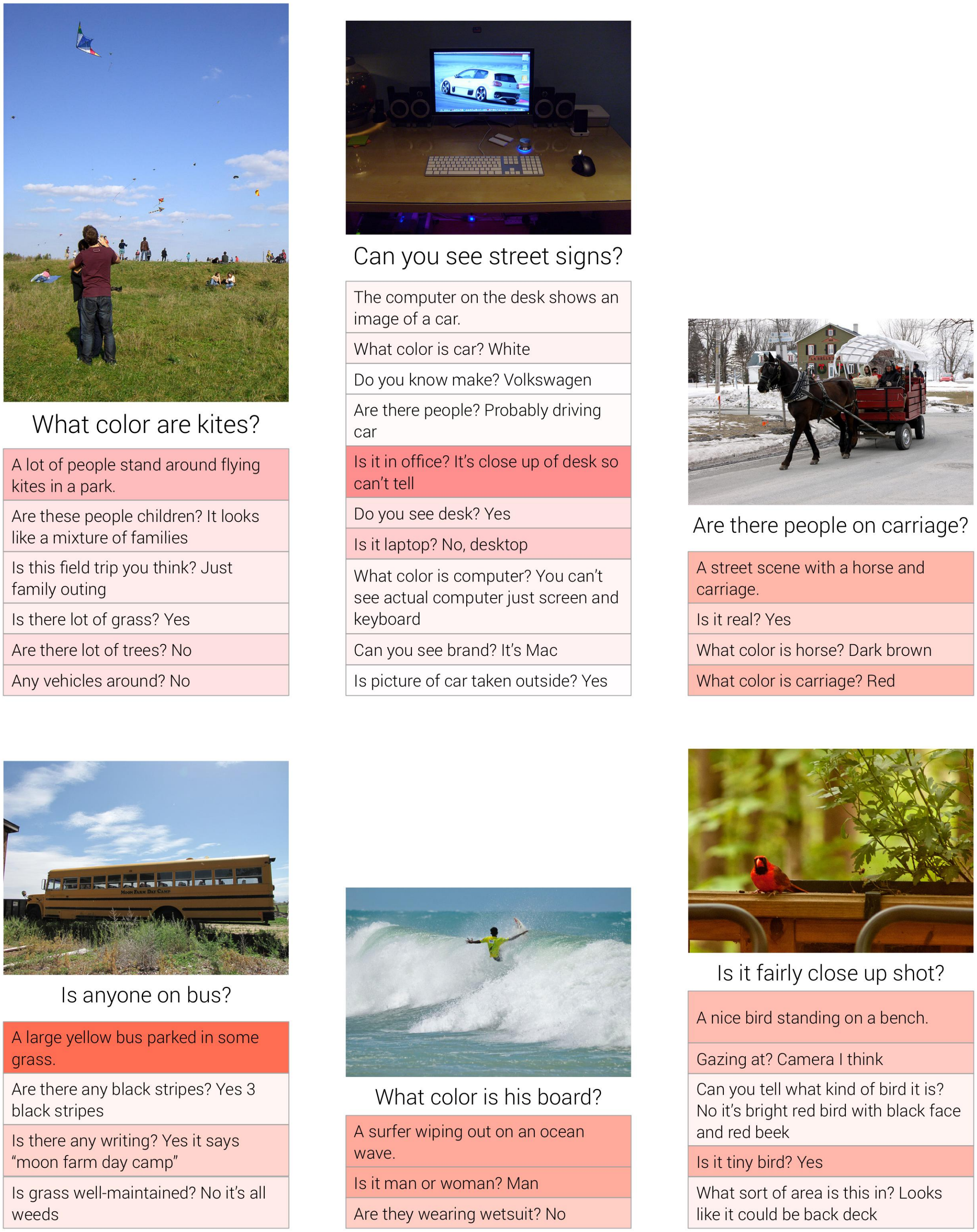



Figure 17: Selected examples of attention over history facts from our Memory Network encoder. The intensity of color in each row indicates the strength of attention placed on that round by the model.

## References


[1] NLTK. http://www.nltk.org/. 18

[2] Torch. http://torch.ch/. 9, 18

[3] Amazon. Alexa. http://alexa.amazon.com/. 6

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2, 3, 4, 5, 6, 10, 11, 13, 14

[5] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *UIST*, 2010. 1

[6] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale Simple Question Answering with Memory Networks. *arXiv preprint arXiv:1506.02075*, 2015. 3

[7] A. Bordes and J. Weston. Learning End-to-End Goal-Oriented Dialog. *arXiv preprint arXiv:1605.07683*, 2016. 3, 6, 8

[8] C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011. 12

[9] J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *ICLR*, 2016. 2, 3

[10] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *CVPR*, 2015. 3

[11] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 3

[12] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *NIPS*, 2015. 3, 4, 11, 13

[13] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In *PNAS*, 2014. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 1

[15] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015. 1, 3

[16] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 3

[17] Q. V. L. Ilya Sutskever, Oriol Vinyals. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014. 12

[18] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016. 7

[19] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, et al. Smart Reply: Automated Response Suggestion for Email. In *KDD*, 2016. 3

[20] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3

[21] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 18

[22] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 3

[23] O. Lemon, K. Georgila, J. Henderson, and M. Stuttle. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In *EACL*, 2006. 2

[24] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*, 2016. 3

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollÃ¡r, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 3, 4

[26] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*, 2016. 3, 6

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016. 1

[28] R. Lowe, N. Pow, I. Serban, and J. Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL*, 2015. 3

[29] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper LSTM and Normalized CNN Visual Question Answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN, 2015. 8

[30] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*, 2016. 8

[31] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014. 3, 11

[32] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 1, 3

[33] H. Mei, M. Bansal, and M. R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI*, 2016. 2

[34] Microsoft. Bing Spell Check API. https://www.microsoft.com/cognitive-services/en-us/bing-spell-check-api/documentation. 18

[35] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015. 1

[36] T. Paek. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, 2001. 2

[37] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3

[38] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD:

100,000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2016. 3
[39] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people with "their" names using coreference resolution. In *ECCV*, 2014. 3
[40] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions. In *EMNLP*, 2016. 5, 13
[41] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015. 1, 3, 11
[42] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 3
[43] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *CVPR*, 2015. 3
[44] I. V. Serban, A. García-Durán, Ç. Gülçehre, S. Ahn, S. Chandar, A. C. Courville, and Y. Bengio. Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. In *ACL*, 2016. 3
[45] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 2016. 3
[46] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *arXiv preprint arXiv:1605.06069*, 2016. 3, 7
[47] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 1
[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 7
[49] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*, 2016. 1
[50] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu. Joint Video and Text Parsing for Understanding Events and Answering Queries. *IEEE MultiMedia*, 2014. 1
[51] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to Sequence - Video to Text. In *ICCV*, 2015. 3
[52] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL HLT*, 2015. 3
[53] O. Vinyals and Q. Le. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*, 2015. 3
[54] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 3
[55] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao. Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNs. *arXiv preprint arXiv:1610.01119*, 2016. 1
[56] J. Weizenbaum. ELIZA. `http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm`. 2, 3
[57] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *ICLR*, 2016. 1, 3
[58] S. Wu, H. Pique, and J. Wieland. Using Artificial Intelligence to Help Blind People 'See' Facebook. http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/, 2016. 1
[59] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked Attention Networks for Image Question Answering. In *CVPR*, 2016. 8
[60] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. In *ICCV*, 2015. 3, 11
[61] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *CVPR*, 2016. 5, 6, 13, 14
[62] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 4, 11, 13