Visual Dialog:
Supplementary Document

Abstract

This supplementary document is organized as follows:

- *Sec. 1 studies how and why VisDial is more than just a collection of independent Q&As.*
- *Sec. 2 shows qualitative examples from our dataset.*
- *Sec. 3 presents detailed human studies along with comparisons to machine accuracy. The interface is also demonstrated in a video*[1].
- *Sec. 4 shows snapshots of our two-person chat data-collection interface on Amazon Mechanical Turk. The interface is also demonstrated in a video*[1].
- *Sec. 5 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*[1].
- *Sec. 6 presents performance of our models on VisDial v0.5* `test`.
- *Sec. 7 presents implementation-level training details including data preprocessing, and model architectures.*
- *Putting it all together, we compile a video demonstrating our visual chatbot*[1] *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.*

[1] https://vimeo.com/193092429

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

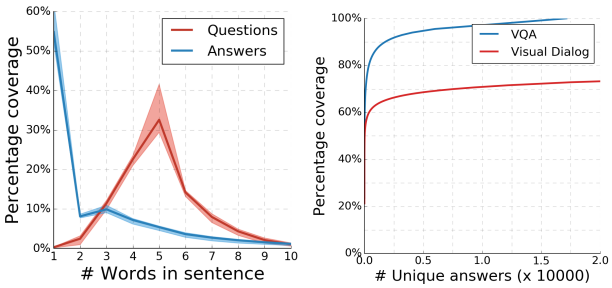
In this section, we lay out an exhaustive list of differences between VisDial and existing image question-answering datasets, with the VQA dataset [3] serving as the representative.

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.

1.1. VisDial has longer free-form answers

Fig. 1a shows the distribution of answer lengths in VisDial. and Tab. 1 compares statistics of VisDial with existing image question answering datasets. Unlike previous datasets, answers in VisDial are longer, conversational, and more descriptive – mean-length 2.9 words (VisDial) vs 1.1 (VQA), 2.0 (Visual 7W), 2.8 (Visual Madlibs). Moreover, 37.1% 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 1: 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. 1b 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

	# QA	# Images	Q Length	A Length	A Length > 2	Top-1000 A	Human Accuracy
DAQUAR [8]	12,468	1,447	11.5 ± 2.4	1.2 ± 0.5	3.4%	96.4%	-
Visual Madlibs [12]	56,468	9,688	4.9 ± 2.4	2.8 ± 2.0	47.4%	57.9%	-
COCO-QA [11]	117,684	69,172	8.7 ± 2.7	1.0 ± 0	0.0%	100%	-
Baidu [5]	316,193	316,193	-	-	-	-	-
VQA [3]	614,163	204,721	6.2 ± 2.0	1.1 ± 0.4	3.8%	82.7%	✓
Visual7W [14]	327,939	47,300	6.9 ± 2.4	2.0 ± 1.4	27.6%	63.5%	✓
VisDial (Ours)	1,232,870	123,287	5.1 ± 0.0	2.9 ± 0.0	37.1%	63.2%	✓

Table 1: Comparison of existing image question answering datasets with VisDial

in VQA cover ∼83% of all answers, while in VisDial that figure is only ∼63%. There is a significant heavy tail of answers in VisDial – most long strings are unique, and thus the coverage curve in Fig. 1b becomes a straight line with slope 1. In total, there are 337,527 unique answers in VisDial (out of the 1,232,870 answers currently in the dataset).

1.2. 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, 19% of answers, and *nearly all* (98%) 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. 2, 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.

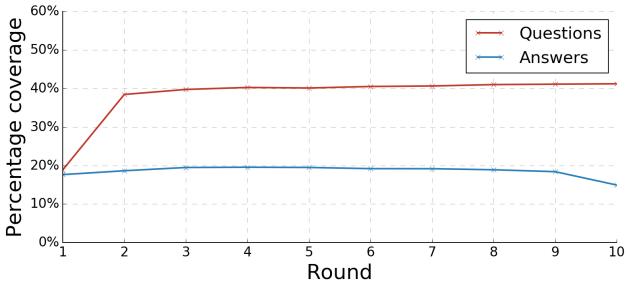


Figure 2: Percentage of QAs with pronouns for different rounds. In round 1, pronoun usage in questions is low (in fact, almost equal to usage in answers). From rounds 2 through 10, pronoun usage is higher in questions and fairly consistent across rounds.

1.3. 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* [3].

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 ± 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 ± 0.05 topics while VQA has 2.53 ± 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 ± 0.02. In contrast, VQA has a topic transition probability of 0.80 for in-order *vs*. 0.83 ± 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.

1.4. 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 [4]. 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 [6] 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.

Tab. 2 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. 2 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.

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 2: 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.

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. 2 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.

1.5. VisDial eliminates visual priming bias in VQA

One key difference between VisDial and previous image question answering datasets (VQA [3], Visual 7W [14], Baidu mQA [5]) 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 [13], 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 [13]. 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.

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. 3). 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 [10] for a related, but complementary effort on question relevance in VQA.

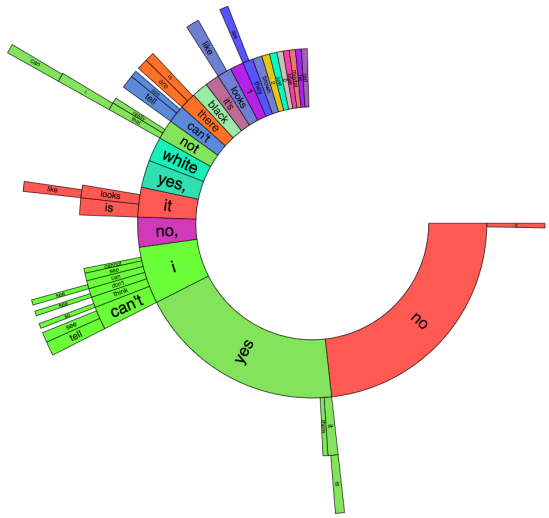


Figure 3: Distribution of answers in VisDial by their first four words. The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

Binary Questions $\neq$ Binary Answers in VisDial. In VQA, binary questions are simply those with 'yes', 'no', 'maybe' as answers [3]. 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' [3, 13] – 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.

2. Qualitative Examples from VisDial

Fig. 4 shows random samples of dialogs from the VisDial dataset.

	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 3: Human-machine performance comparison on VisDial v0.5, 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.

3. Human-Machine Comparison

We conducted studies on AMT to quantitatively evaluate human performance on this task for all combinations of {with image, without image}$\times${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. 3. 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)[2] 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. 3 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

[2] 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

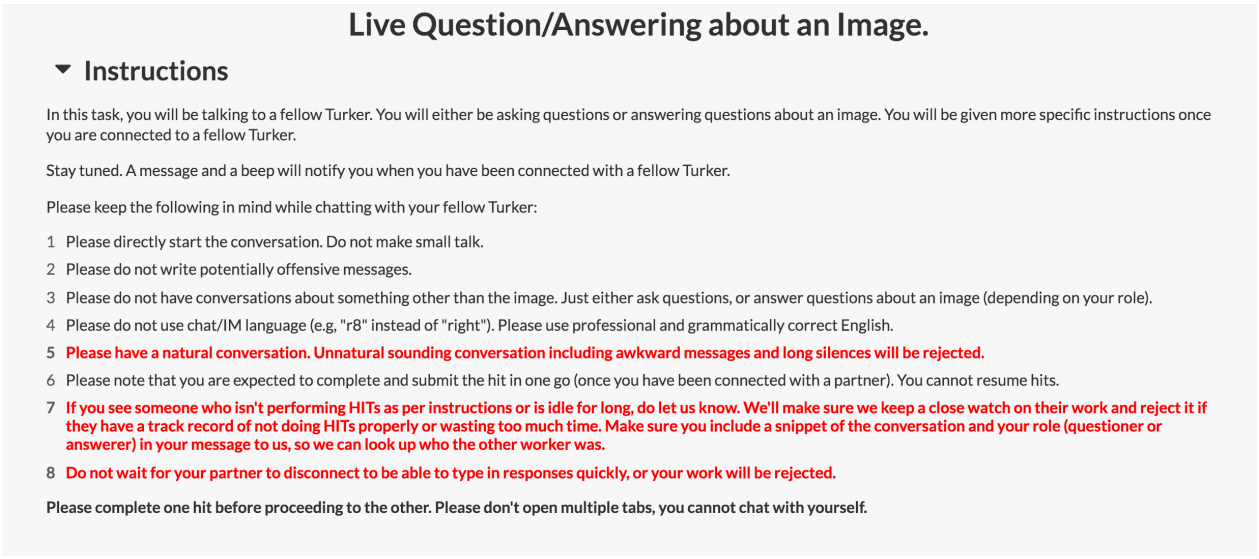
(e)

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

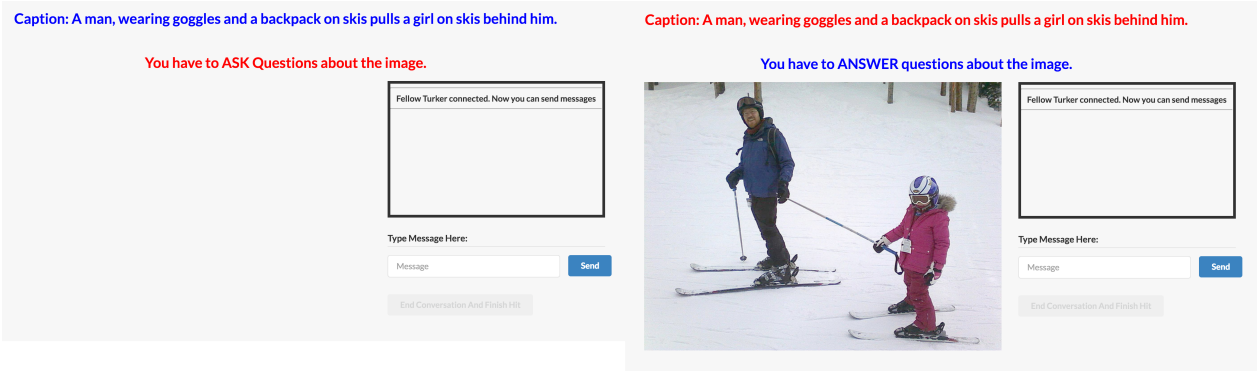
(f)

Figure 4: Examples from VisDial

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%).



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



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

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 (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 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.

4. 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. 5a. Since the workers do not know their roles before starting the study, we provide instructions for both questioner and answerer roles.

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. 5b. Observe that the questioner does not see the image while the answerer does have access to it. Both questioner and answerer see the caption for the image.

5. Additional Analysis of VisDial

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

5.1. Question and Answer Lengths

Fig. 6 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. 7 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.

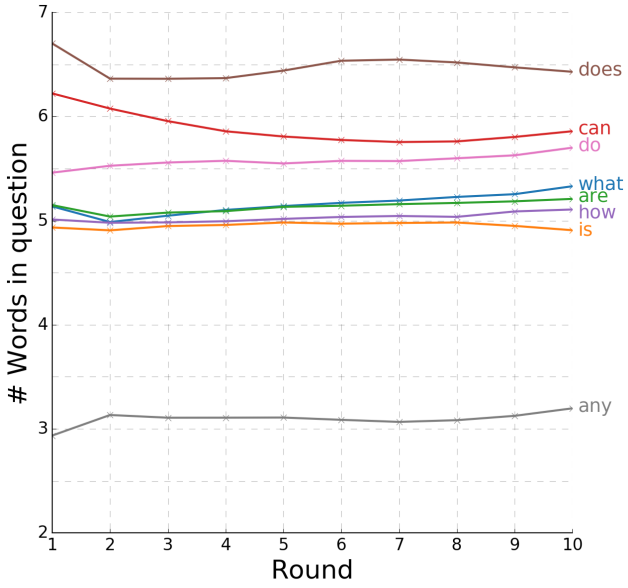


Figure 6: Question lengths by type and round. Average length of question by type is fairly consistent across rounds. Questions starting with 'any' ('any people?', 'any other fruits?', etc.) tend to be the shortest.

5.2. Question Types

Fig. 8 shows round-wise coverage by question type. We see that as conversations progress, 'is', 'what' and 'how' questions reduce while 'can', 'do', 'does', 'any' questions occur more often. Questions starting with 'Is' are the most popular in the dataset.

6. Performance on VisDial v0.5

Tab. 4 shows the results for our proposed models and baselines on VisDial v0.5. A few key takeaways – First, as expected, all learning based models significantly outperform

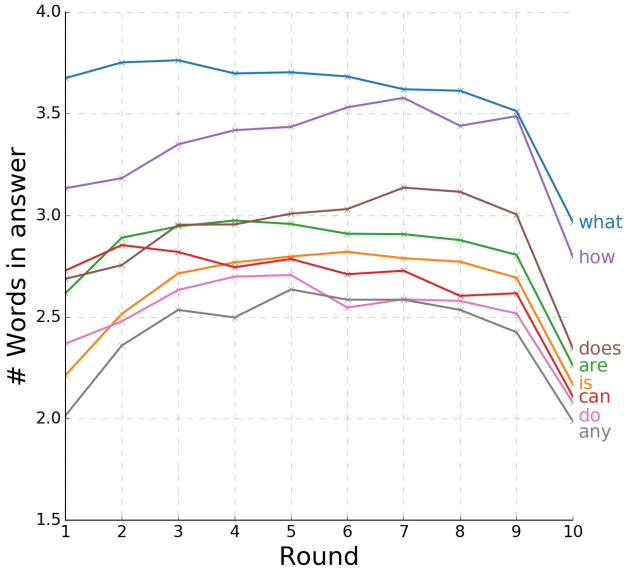


Figure 7: Answer lengths by question type and round. Across question types, average response length tends to be longest in the middle of the conversation.

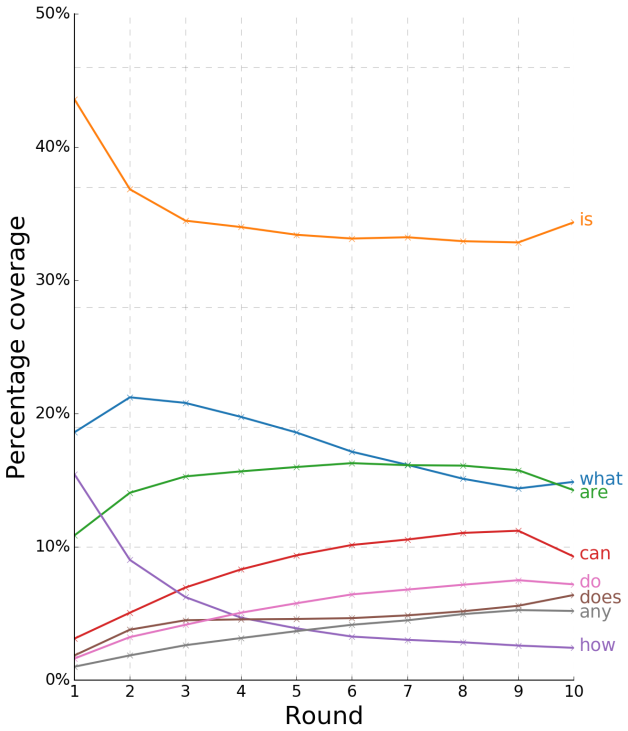


Figure 8: Percentage coverage of question types per round. As conversations progress, 'Is', 'What' and 'How' questions reduce while 'Can', 'Do', 'Does', 'Any' questions occur more often. Questions starting with 'Is' are the most popular in the dataset.

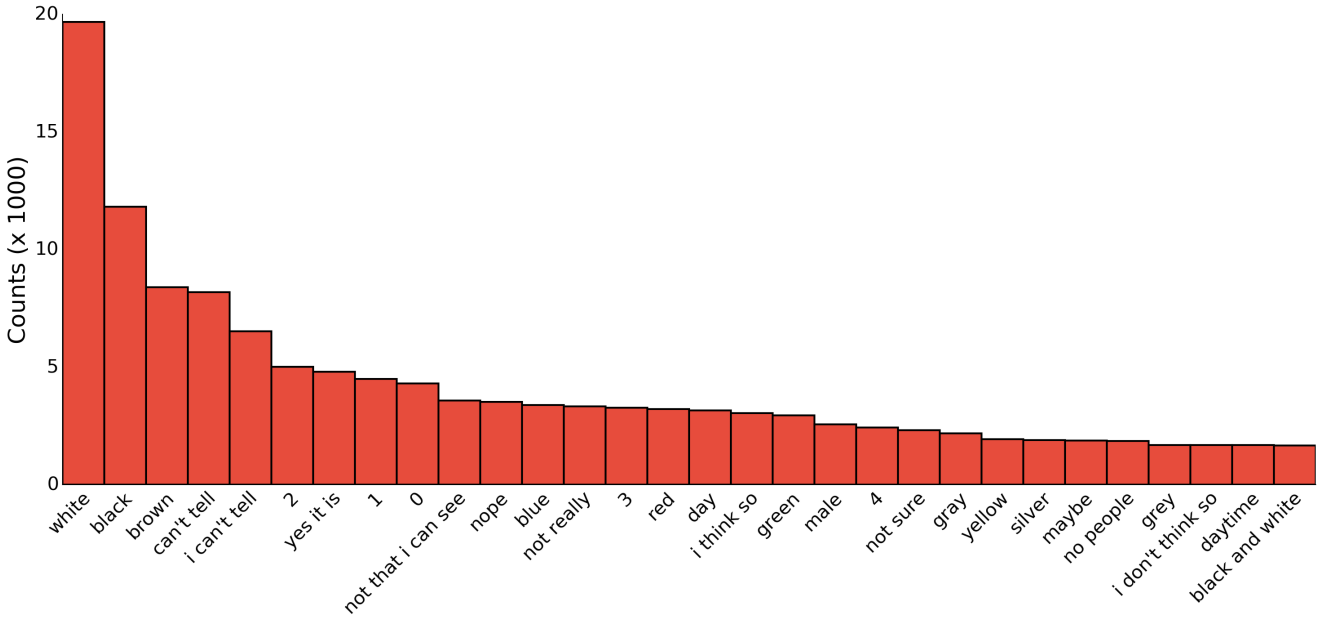
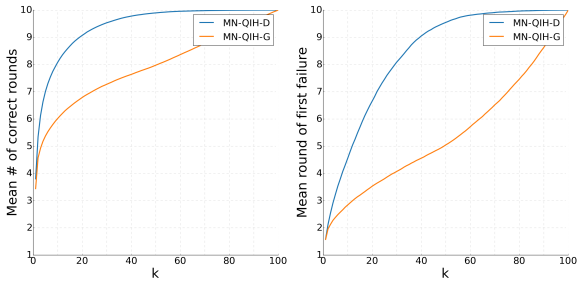


Figure 9: Most frequent answer responses except for 'yes'/'no'

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. Fifth, models looking at I outperform both the blind models (Q, QH) by at least 2% on recall@1 in both decoders. Finally, models that use both H and I have best performance.

Dialog-level evaluation. Using R@5 to define round-level 'success', our best discriminative model MN-QIH-D gets 7.01 rounds out of 10 correct, while generative MN-QIH-G gets 5.37. Further, the mean first-failure-round (under $R@5$) for MN-QIH-D is 3.23, and 2.39 for MN-QIH-G. Fig. 10a and Fig. 10b show plots for all values of k in $R@k$.



(a) (b)

Figure 10: Dialog-level evaluation

7. Experimental Details

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

7.1. 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 nonlinearity to a 512-d vector, which is used to decode the response. Fig. 11a 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

	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 4: Performance of methods on VisDial v0.5, 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.

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 dialog-level LSTM, whose output state at t is used to decode the response to Q_t. Fig. 11b shows the model architecture for our HRE.

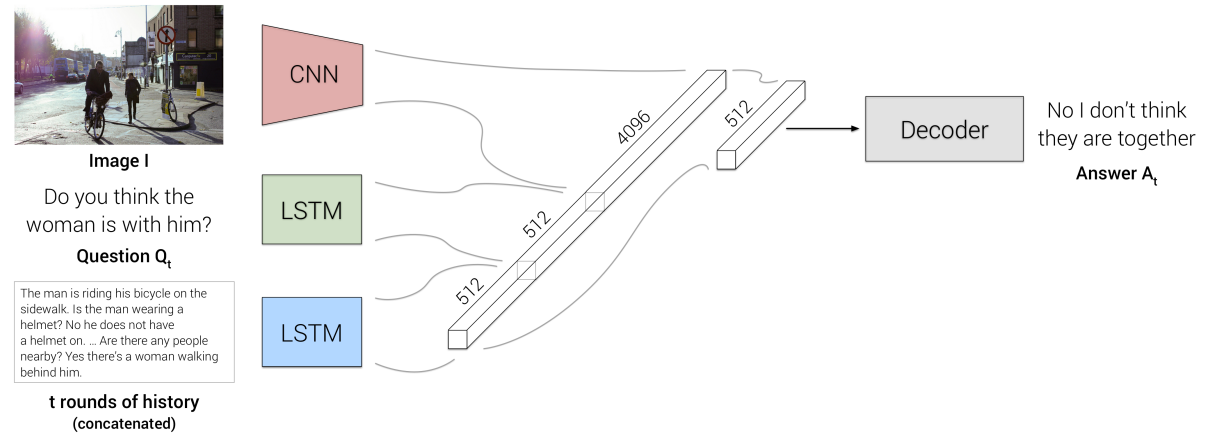
Memory Network. The image is encoded with a VGG-16 CNN and question with an LSTM. We concatenate the representations and follow it by a fully-connected layer and tanh non-linearity to get a 'query vector'. Each caption/QA-pair (or 'fact') in dialog history is encoded independently by an LSTM with shared 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. 11c. Fig. 12 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.

7.2. Training

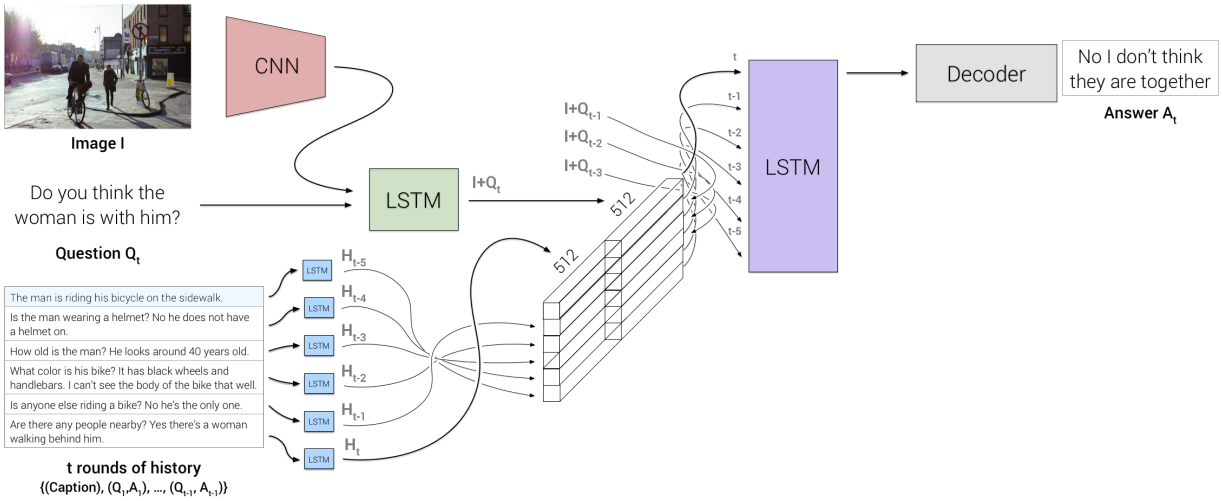
Splits. Recall that VisDial v0.9 contained 83k dialogs on COCO-`train` and 40k on COCO-`val` images. We split the 83k into 80k for training, 3k for validation, and use the 40k as test.

Preprocessing. We spell-correct VisDial data using the Bing API [9]. 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.

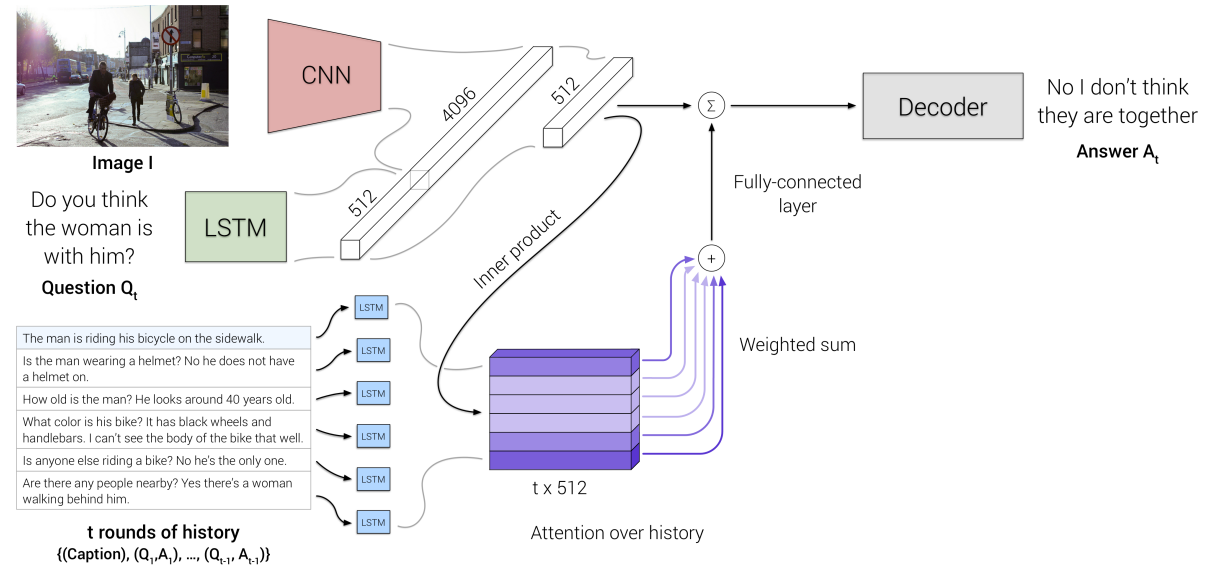
Hyperparameters. All our models are implemented in Torch [2]. Model hyperparameters are chosen by early stopping on `val` based on the Mean Reciprocal Rank (MRR) metric. All LSTMs are 2-layered with 512-dim hidden states. We learn 300-dim embeddings for words and images. These word embeddings are shared across question, history, and decoder LSTMs. We use Adam [7] with a learning rate of 10^{-3} for all models. 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

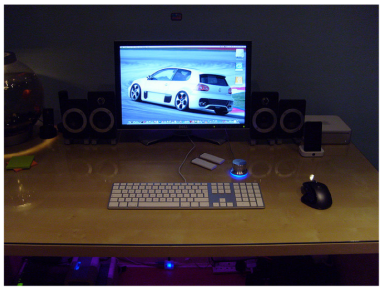


(c) Memory Network Encoder

Figure 11

What color are kites?

A lot of people stand around flying kites in a park.
Are these people children? It looks like a mixture of families
Is this field trip you think? Just family outing
Is there lot of grass? Yes
Are there lot of trees? No
Any vehicles around? No



Can you see street signs?

The computer on the desk shows an image of a car.
What color is car? White
Do you know make? Volkswagen
Are there people? Probably driving car
Is it in office? It's close up of desk so can't tell
Do you see desk? Yes
Is it laptop? No, desktop
What color is computer? You can't see actual computer just screen and keyboard
Can you see brand? It's Mac
Is picture of car taken outside? Yes

Are there people on carriage?

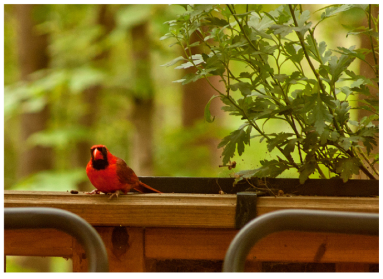
A street scene with a horse and carriage.
Is it real? Yes
What color is horse? Dark brown
What color is carriage? Red

Is anyone on bus?

A large yellow bus parked in some grass.
Are there any black stripes? Yes 3 black stripes
Is there any writing? Yes it says "moon farm day camp"
Is grass well-maintained? No it's all weeds

What color is his board?

A surfer wiping out on an ocean wave.
Is it man or woman? Man
Are they wearing wetsuit? No



Is it fairly close up shot?

A nice bird standing on a bench.
Gazing at? Camera I think
Can you tell what kind of bird it is? No it's bright red bird with black face and red beek
Is it tiny bird? Yes
What sort of area is this in? Looks like it could be back deck

Figure 12: 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/. 10

[2] Torch. http://torch.ch/. 10

[3] 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, 4, 5

[4] 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. 3

[5] 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. 2, 4

[6] Q. V. L. Ilya Sutskever, Oriol Vinyals. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014. 3

[7] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 10

[8] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014. 2

[9] Microsoft. Bing Spell Check API. https://www.microsoft.com/cognitive-services/en-us/bing-spell-check-api/documentation. 10

[10] 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. 4

[11] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015. 2

[12] 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. 2

[13] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *CVPR*, 2016. 4, 5

[14] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 2, 4