Supplementary Material

The supplementary document is organized as follows:

- Sec. 6 covers additional details of VisDial-Q.
- Sec. 7 shows additional quantitative evaluations beyond Mean Rank and Recal@5 (included in the main paper).
- Sec. 8 shows additional qualitative examples of unrolling question generation and answering modules.

6. VisDial-Q and VisDial comparison

Sec. 3.5 explains the re-purposing of VisDial to VisDial-Q, using correct, plausible, popular and random question options. Here we include a comparison of the distribution of answers and questions. Sentence distribution of target questions are shown in Fig. 10. A steeper slope of answer distribution vs. question distribution shows the challenging nature of question generation. This is supported by entropy (higher 4.71bits for question and 4.52bits for answer). Fig. 10b has examples of popular question candidates.

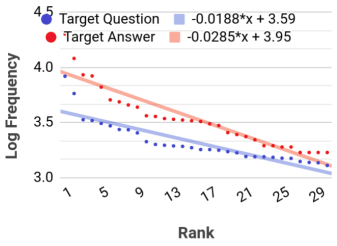
7. Quantitative Results

In the following we present additional quantitative results, some of which were already mentioned in the paper. We report Recall@1, Recall@5, Recall@10, Mean Reciprocal Rank (MRR) and Mean rank for the test sets of both answer prediction task (*VisDial evaluation*) and question prediction task (*VisDial-Q evaluation*). Fig. 11 and Fig. 12 summarize these metrics as training proceeds for the two tasks. Our models perform significantly better than the most complex architectures of [6]. Our models are easy to train, with convergence in under 5 epochs in contrast to a 20-30 epoch pre-training required for the baseline set by generator-discriminator architecture in [23]. Since we introduce a new evaluation protocol for question prediction in visual dialog, there aren't any existing baselines for this task in Fig. 12.

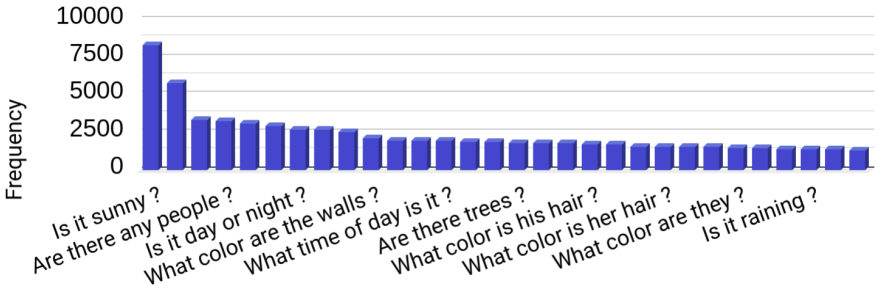
8. Qualitative Results

As mentioned in the paper, we decide to unroll both our question prediction and answer prediction module together to show how these discriminative models can be used to 'generate' dialog. The answer module chooses the best answer option to a given question while the question module chooses the best next question to a given question-answer pair. As mentioned in the paper, a few arrangements are necessary to jointly unroll questioning and answering modules, since answer options and next question options are available for only dataset dialogs, while we are 'generating' (*i.e.*, selecting) new sequences. We create options on the fly, by choosing from a set of questions and answers of nearest neighbor images. Since there are no ground-truth options for these predicted dialog sequences, we can't report quantitative metrics for this dynamic setup where our models communicate with each other.

We test our models in two different visual dialog setups. *Firstly*, we unroll our VQA and VQG modules when there is very little history. A visual dialog system needs to be more inquisitive in such a setup and '*explore*' the image. Fig. 13 shows both short and long dialogs predicted by our models in such a setup. *Secondly*, we also test our models when there is a long history available to build on. Here, the models need to be consistent with existing context - avoid repetitions, and handle co-reference resolution. In such a setup the models '*exploit*' the available history to find finer details about the image. The generations do not repeat questions from the history and reference objects using correct pronouns. Fig. 14 shows both short and long visual dialogs predicted by our discriminative VQA and VQG models.

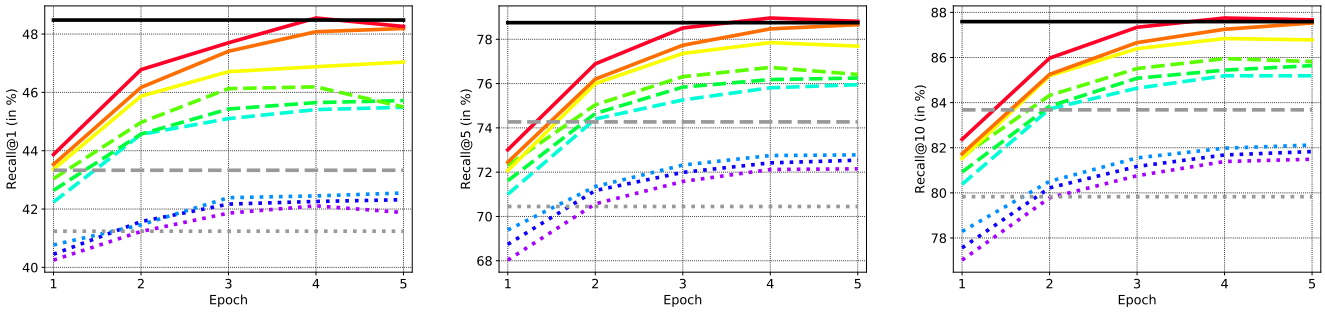


(a) Comparing target answer and target question distributions

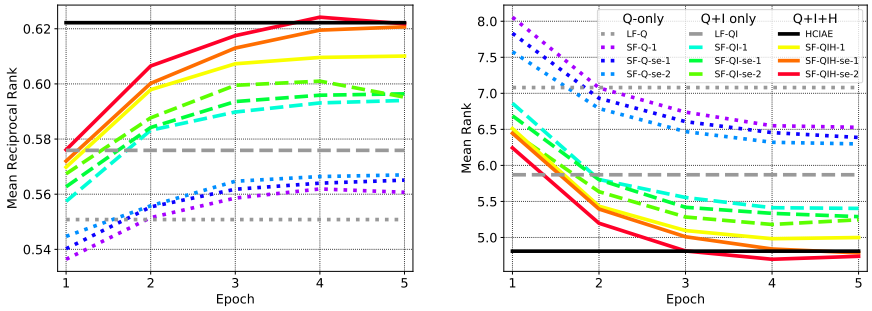


(b) Target question distribution (top 30)

Figure 10: (a) compares target distribution of questions and answers (top 30 ranked targets). Steeper slope of answers indicates higher frequency biases in the answer targets. (b) displays the frequency distribution of questions, analogous to Fig. 15 in [6].

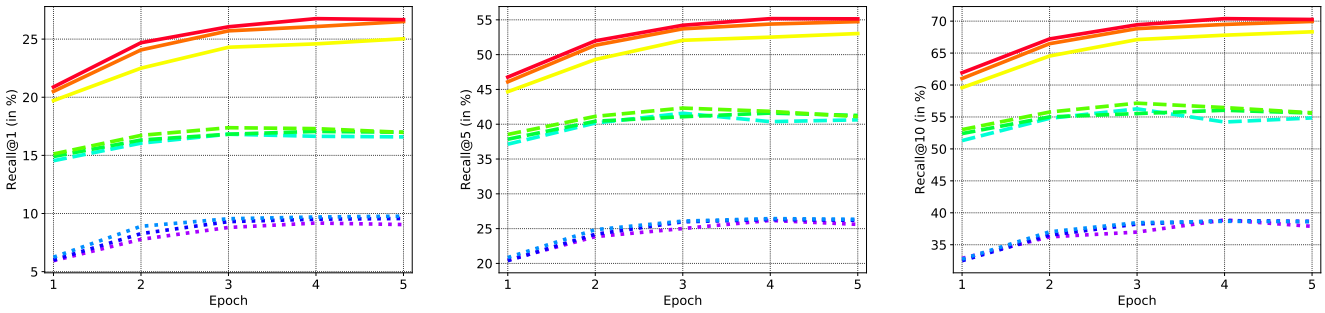


(a) Recall@1 (b) Recall@5 (c) Recall@10

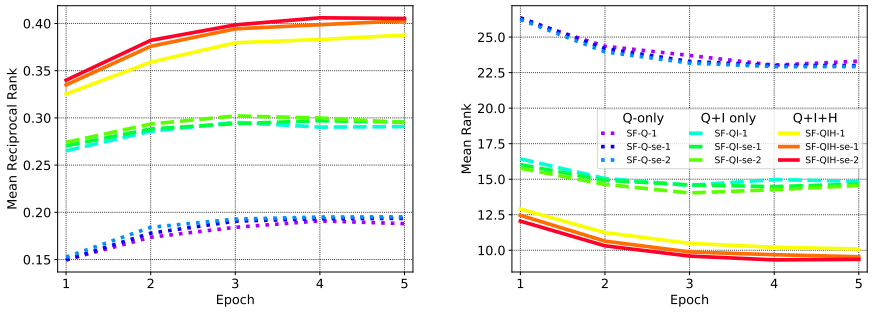


(d) MRR (e) Mean Rank

Figure 11: *VisDial* evaluation protocol: Evaluation metrics for our models and best models from [6, 23] - Late fusion (LF) and HCIAE-D-NP-ATT (abbreviated as HCIAE). '-1' and '-2' refer to one and two hidden layers in our 'similarity learning + fusion net' (SF) model. '-se' refers to shared word embeddings across all LSTM nets. (Legend is same as (e))



(a) Recall@1 (b) Recall@5 (c) Recall@10



(d) MRR (e) Mean Rank

Figure 12: *VisDial-Q* evaluation protocol: Metrics for our models on the newly proposed VisDial-Q evaluation protocol. '-1' and '-2' refer to one and two hidden layers in our 'similarity learning + fusion net' (SF) model. '-se' refers to shared word embeddings across all LSTM nets.

Figure 13: Joint unrolling of VQA and VQG modules for short history (1 QA pair): Short and long dialogs 'generated' by our discriminative models.

Figure 14: Joint unrolling of VQA and VQG modules for long history (5 QA pair): Short and long dialogs 'generated' by our discriminative models.