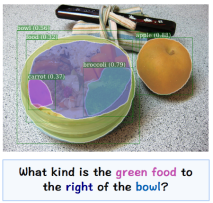

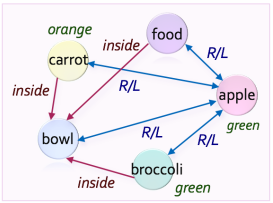

Table 1: Error Analysis

Type	Rate %
Ambiguity / Synonymity	31
Perception	29
Counting	12
External Knowledge	11
Subjective	9
OCR	8

Table 2: Noise Experiments

Alteration	Sem	Sem+Vis
Add Object	24/25	25/25
Remove Object	18/25	21/25
Remove Attribute	21/25	25/25
Remove Relation	19/25	23/25

Dear reviewers and area chairs,

First of all, we would like to thank the reviewers for the feedback, questions and comments! We appreciate a lot the time they took to read our paper and write the reviews. In the following, we discuss specific points made by the reviewers:

Reasoning Skills and Ablations. To explore reasoning skills that are not covered by GQA, we compared our model with the BottomUp model for counting and negative (*not*) questions selected from VQA2. In both cases, our approach achieves better performance, with 52.14% vs 48.64% for counting, and 54.2% vs 38.1% for negation. Since the NSM computes a sequence of soft attentions over the graph, it can distribute its attention probability to capture multiple nodes at the same time, which we believe could help with counting, negations or universal quantification. Following reviewer 1's suggestion, we have also explored the impact of the final-stage question conditioning on the overall performance, yielding an accuracy of 60.41% for GQA and 44.35% for VQA-CP. Introducing such conditioning allows the model to consider more directly question nuances that may be necessary to address it, and thus potentially provides increased flexibility in handling more varied reasoning types (even though it may indeed also leverage training biases). To get further insight, we plan to perform a qualitative analysis to inspect the model's internal behavior and produced attention maps for questions involving different reasoning skills.

Noise Robustness and the Test Set. To evaluate the model's robustness to noise in the predicted scene graphs, we made graph alterations for 100 questions answered correctly by the model, and measured their impact on its performance. We also evaluated a variant of the NSM model where attention for each object is computed both over its semantic representation and its visual features. We can see in table 2 that even when noise is introduced, both model versions still answer most questions correctly, with the new variant being naturally more robust as it relies on additional information. We plan to extend this initial qualitative assessment and explore larger-scale controlled experiments about noise robustness which we will add to the paper. Finally, please note that, as discussed in the official website, the images and scene graphs in the GQA test set have not been made public and do not overlap with Visual Genome, but rather have been collected independently for the GQA task. As such, they have not been used either for training or evaluation of our or other scene graph parsers – namely, our model's performance over the test set is based on graphs predicted over new unused images, providing further evidence for the ability of the model to cope with potentially noisy graphs.

Additional Datasets and Error Analysis. We have begun to explore the model in context of other VQA datasets: Currently, we achieve 68.2% for VQA2 and believe that with tuning we may be able to improve that result. Based on the error analysis summarized in table 1, we can see that many of the questions which are counted as errors (31%) are in fact cases where the model's prediction is semantically equivalent to or synonymous with the labeled answer, e.g. *korean air / korean*, or *flushing / to flush*. Other cases include subjective questions (*Is this a happy home? maybe*) or ones that require external knowledge (*Where is this located? London*) or OCR (*What is written on . . .*), with the rest of the errors arising from imperfect perception (29%) or counting mistakes (12%). We plan to also perform experiments on other datasets such as CLEVR-humans or NLVR^2. Finally, following the strong generalization results obtained for the new GQA splits (section 4.2), we would especially be interested to test the model in terms of data efficiency, e.g. over CLEVR or VQA, comparing the training set size needed by our and other models to achieve the same accuracies.

Graph Generation and Sparsity. We implement our own scene graph parser following common ideas from several public implementations [55, 58, 10]. In particular, following [55], we prune the graphs to make them sparser and reduce the computational load, keeping only the top 300 most likely edges, based on a combination of factors including: class-based prior [58], objects proximity, and graph parser's confidence. We will add to the paper a more detailed description of our scene graph parser, including both further implementation details as well as the training scheme.

Misc and Editing. The decoder (section 3.3) is indeed not supervised directly but rather recurrently computes attention, first over the concepts and then over the updated question word embeddings, to generate a sequence of soft operations. In lines L108–112 all the elements are vectors of the same dimension $d = 300$: the concept embeddings are learned parameters, while the states, edges and instructions are computed throughout the model operation. The sizes of the new generalization splits are: 834k/241k train/test for the content split and 858k/217k for the structure split. Finally, we will extend the Related Work section to discuss the papers mentioned in the reviews, add the reasoning visualization (Supplementary figure 1) to the main paper, and fix the notation and terminology suggestions.

Thank you very much!

– Paper 3180 authors