We thank the reviewers for their recognition of our work and helpful comments.

Q1/Reviewer#1: "More related work on GraphVLAD front". **A:** We will modify our paper to cite more related work.

Q2/Reviewer#1: "Throw some light on the failure modes". **A:** In Fig. 4, we have shown two failure examples and given some discussions (Line 220-227). Taking the 2nd row as an example, when the inference of correct rationale requires a strong ability of reasoning with the commonsense, the model might fail. The insights are: Firstly, when the visual reasoning involves more commonsense, the task of interpreting the answer is more difficult. Secondly, though the model fails, the wrong rationale indeed matches the visual content, which shows GraphVLAD module is helpful for obtaining an effective visual representation.

Q3/Reviewer#1: "On the modification of Fig. 2". **A:** Thanks for your advice. We will rectify Fig. 2.

Q4/Reviewer#1: "On the denominator of Eq. 3". **A:** Thank you. In the denominator, it should be $b_{j'}$, instead of $b_{k'}$.

Q5/Reviewer#1: "On the initialization of centers". **A:** We use random uniform initialization on the interval [0, 1).

Q6/Reviewer#1: "On the extraction of object features and feature map". **A:** The visual feature of each object is Roi-Aligned from its bounding region [2]. And we use ResNet50 as the backbone in Mask R-CNN [2]. By performing a Max-pooling operation on the output of the third block of ResNet50, we obtain X to compute GraphVLAD.

Q7/Reviewer#1: "On the difference between the representation Y and $\widetilde{Q}$". **A:** In this paper, Y is a feature vector which indicates the output of LSTM at the last time step. $\widetilde{Q}$ is an output matrix which contains the output of LSTM at each time step. And the first dimension of $\widetilde{Q}$ indicates the length of the query.

Q1/Reviewer#2: "The connection between our model and the brain". **A:** In Line 37, recent studies [28] on brain networks have suggested that brain function or cognition can be described as the global and dynamic integration of local neuronal connectivity. Since we used GCN in each module, our CCN could be regarded as a hierarchical GCN. The bottom layer is used to capture local relations. As the layer increases, our model could use dynamic connections to integrate local relations. Finally, the top layer performs reasoning based on the global integration of all the relations. This process is similar to that of the brain cognition.

Q2/Reviewer#2: "On the explanation of NetVLAD". **A:** We will explain more details of NetVLAD in our paper.

Q3/Reviewer#2: "The performance of VQA-CPv2". **A:** We test our CCN on the VQA-CPv2 dataset [1]. We do not change the hyper-parameters tuned on VCR. The accuracy of our method is 39.44%, which outperforms the baseline method [1] by 8%. Recently, [32] specially designed a multimodal fusion module for VQA. It used a pairwise modeling component to further update the multimodal representation with multiple iterations. We obtained comparable results to [32] (39.54%) in VQA-CPv2. This shows our method could be readily applied to standard VQA task.

Q4/Reviewer#2: "The result of the method based on word embedding and LSTM". **A:** We conducted the baseline that used "word embedding + LSTM". The LSTM encoded the query and response. In $Q \rightarrow A$, $QA \rightarrow R$, and $Q \rightarrow AR$ mode, the performance on the validation set is 57.3%, 60.1%, and 36.5%. Our CCN significantly outperforms it.

Q5/Reviewer#2: "More interpretations about the t-SNE visualization". **A:** In this paper, we want to use conditional centers to capture the characteristics of the current input data. When an image contains rich content and its corresponding query is complex, e.g., Fig. 5(b), in order to capture rich visual information to answer the query, these centers will learn to spread further apart from each other. Meanwhile, when the image content and its corresponding query contain relatively less information, e.g., Fig. 5(a), in order to focus on visual information which is related to the query, these centers will adaptively adjust to being more concentrated. In this way, we can obtain an effective visual representation, which is helpful for the following contextualization and reasoning.

Q1/Reviewer#3: "On notations". **A:** We will modify our paper to make notations much more concise and consistent.

Q2/Reviewer#3: "On the use of BERT". **A:** We will modify our paper to explain the details of the other models clearly. Besides, we tried some VQA models, e.g., MUTAN [5] and MLB [18], and used BERT as word embedding. However, we found R2C model (baseline in our paper) outperformed them by around 3% in terms of $Q \rightarrow AR$ performance.

Q3/Reviewer#3: "On the explanation in some experimental settings". **A:** In Line 193-196, we have given the setting details of each mode. For the case of $QA \rightarrow R$, it uses the question and **corresponding ground-truth answer** to predict rationale. And the answer is concatenated with the question in textual form. In this case, P indicates the length of the concatenation of the question and the answer. And J indicates the length of the rationale.

Q4/Reviewer#3: "On the generalized performance". **A:** Here, we test our method on the VQA-CPv2 dataset [1] which is the improved version of VQA 2.0. The accuracy of our method is 39.44%, which outperforms the baseline method [1] by 8%. This shows our method could be readily applied to standard VQA task. Please refer to our response to Q3/Reviewer#2 for details.

References

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pages 4971–4980, 2018.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.