We would like to thank all reviewers for your constructive and intriguing comments. We have concluded more related work, compared with more SOTA, based on your suggestions. We have significantly polished our paper in terms of introduction reorganization, grammar correction, fluent expression as well as broader impacts for camera ready.

Reviewer #1 and Reviewer #4

Q1: Comparison with FiLM (AAAI 2018). The experiment is somewhat inadequate.

A1: Feature-wise Linear Modulation, namely FiLM is proposed to influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. Such a novel module has achieved SOTA performance on the CLEVR benchmark. The GRU module in FiLM functions similar to the BiLSTM module in our porposed SIRI, and the FiLM functions similar to the spatial relation guided distillation module. Besides, positional embedding are leveraged in both work. Whereas, the spatial relation guided distillation we introduced induces the network with a gate mechanism while FiLM performs feature-wise affine transformation. To fairly compare our SIRI with FiLM, we adopt the part II and the part III only in SIRI. Our SIRI has an accuracy@80px of 58.33%, much better than FiLM with an accuracy@80px of 52.37%, which demonstrates the effectiveness of our SIRI.

Reivewer #2

Q1: The effect of k (for the number of top-k selected orientation words) is also not studied.

A1: The accuracies@80px are 51.02%, 58.33% and 59.74% when k is 4, 6 and 8, respectively.

Q3: The ablation study is not as thorough as it can be (it adds the components in order). Ideally, it would also show the effect of 2) and 3) without using 1) (with ResNet features instead of GloRe) and effect of 3) without 2).

A3: The accuracy@80px is 56.24% when the component 2) and 3) are adopted without 1). Besides, the accuracy is 38.76% when the component 3) is adopted without 1) and 2).

Q4: It is not clear whether it would be of interest to the broader NeurIPS community.

A4: We believe that our proposed SIRI with strong novelty and promising performance provides a new insight in terms of architecture design for any vision-language tasks, such as VQA, SDR, etc.

Reviewer #3

Q2: The new extended dataset used in 4.2 should be described more, no details are given on this new data and how it was collected.

A2: Due to the page limitation, the details of the new extended dataset are appended in the supplementray materials. We also analyze the word frequency on it and visualize the prediction results on this new dataset.

Reviewer #4

Q2: Although this is a new task, but the solution is compositional and of limited novelty.

A2: In this paper, we design a novel framework to explicitly tackle the SDR task. Each component is carefully designed and well investigated. We believe that such a novel framework can push forward this important task and provides a new insight for any other vision-language task.

Q3: The ablation study in Table 3 only shows three combinations besides pure LingUnet – I, I+II and I+II+III, it will be better if the author could provide the combinations of II + III, III only and II only. Then we can better evaluate the importance of Part II and Part III.

A3: The accuracies@80px are 56.24%, 38.76% and 44.95% for II + III, III only, II only, respectively.

Q4: About the generalization ability of Stage II & III.

A4: Because MAttNet detects objects and scores the RoIs, the Stage II and III in our paper are not suitable to MAttNet. We use YOLO-VG(A Fast and Accurate One-Stage Approach to Visual Grounding, ICCV2019) as our baseline, a one stage method for visual grounding, which can be end-to-end trained. To investigate the generalization ability of Stage II & III, we add the Stage II and III to it. The results show that Stage II and III can improve the performance by 0.9% and 0.7%, respectively. Thus, our proposed modules can also perform well on other tasks and datasets.

Q5: The paper has some typos, such as: [Line 92]: adapt should be adopt; [Line 121]: averaged should be summed; [In figure 2]: VI (in fact 6) should be IV (4).

A5: Thanks for pointing out these typos. We have significantly polished this paper for camera ready.