We thank all the reviewers. Based on the valuable comments, we provide the following rebuttal to address the concerns of each reviewer, respectively.

Response to Reviewer 1 Thanks for your constructive feedbacks! Regarding your comments:

Comments 1 & 4: It turns out that RNP (The baseline in "Rationalizing Nueral Predictions" proposed by Lei et. al.) with class-specific generators will converge to a set of degenerated solutions, where, rather than highlighting the informative words for each class, the generators will communicate the class information using trivial symbols (*e.g.* punctuation) or other uninformative features (*e.g.* word position). In one degeneration example, the class-0 generator will always highlight the first word and the class-1 generator will always highlight the last word. As a result, the predictor always has 100% prediction accuracy, which achieves the global optimum of the collaborative target function of RNP. Our empirical experiments verify that RNP with class-specific generators quickly converges to degeneration.

In fact, even the original RNP suffers from the degeneration problem. The additional class input merely deteriorates the problem. The problem primarily results from the collaborative nature of the RNP framework. Furthermore, we are able to show that the adversarial game in CAR can fundamentally resolve the degeneration problem. This is because any attempt to communicate class information using uninformative symbols will be easily mimicked by the counterfactual generator, and therefore would not be a good strategy for the factual generator, whose goal is to prevent mimicking. This is another major advantage of CAR, which we did not have enough space to uncover in the paper. Nevertheless, we will add the above discussion to the paper.

Comment 2: Max-pooling followed by a feed-forward classifier is a commonly-used and well-performed sequence classifier in the sentiment classification community (*e.g.,* Yoon Kim, EMNLP 2014). Therefore we simply follow the usage. Moreover, in our experiments when GRUs are used, max-pooling makes the training converge faster compared to using the final hidden state, which is beneficial for the game training.

Comment 3: Eq. (10) is slightly different from [18] in the first term as we hope to better control of the lengths for apple-to-apple comparison. We will add the reference in the updated version.

Comment 5: Directly highlighting top-K positions during the testing phase can be undesirable because during training we used Eq. (10) to constrain on the *average* sparsity level, not the *per-passage* sparsity level. The theoretic guarantee of CAR is also contingent on the average sparsity level. Per-passage sparsity stipulates that each passage has exactly the same proportion of rationale, which is less theoretically and empirically sound, and which also creates a training-test discrepancy if applied to testing only. Using top-K projection for training is not desirable either because the top-K ball is non-convex and does not lead to good convergence.

Comment 6: RNP and POST-EXP come with their respective strengths and weaknesses. RNP is good at finding the aspects that correlate most closely with the output labels, but lacks class-specific capabilities; whereas POST-EXP is good at finding class-specific words and phrases, but does poorly in finding the right aspect. That is why the two baselines perform differently in Tables 1 and 2. Amazon reviews (Table 1) mostly contain single aspect, but has a lot of mixed reviews, and thus POST-EXP works better. Beer and hotel reviews (Table 2) contain multiple aspects, but within each aspect, the reviews are less mixed, and thus RNP performs better. This can be verified by Tables 4 and 5.

Response to Reviewer 2 We greatly appreciate your acknowledgment of the paper. Regarding your concerns:

Direct comparison to Lei et. al. paper: The experiment in the Lei et. al. paper is a regression task, where the output scores are real values within $[0, 1]$. In our experiment, we convert the task to a classification task (because class-wise rationales are only well-defined on classification tasks so far) by quantizing the real-valued scores into binary scores. The quantization operation squeezes out information, and deteriorates the performance of the RNP baseline. We explained this difference in the paper (line 213), but we will emphasize it more in the updated version.

Inter-annotator agreement: Each HIT was originally assigned to one crowd worker. Therefore, we rerun the experiment and obtain a second set of scores. We then compute the agreement between the two sets of scores. Please note that despite that the agreement score can be high or not so high, the new set of accuracy is consistent with the one reported in the paper. For our method on factual rationales, the sentiment agreement is 0.71 on beer reviews, 0.79 on hotel reviews and 0.81 on Amazon reviews. The aspect agreement is 0.58 on beer reviews and 0.64 on hotel reviews. The aspect agreement is lower because for many cases people are making random guesses among the three options, which can be confirmed by the low accuracy in some aspects (beer aroma and hotel cleanliness) in Tables 4 and 5.

Response to Reviewer 4 We greatly appreciate your positive feedback! Please note that although there is only one generator network for each class in our implementation, there are still two generator players for each class. The two players are differentiated by feeding the ground truth label as an additional input to the generator network, and they still have completely distinct target functions in the game. We will make this point clearer in our updated version.