We gratefully thank all reviewers for their valuable comments. We will try our best to address them in the revision.

Comments on Clarity for R #1.
(1) Thanks for pointing out the missing details in HINT paper, we will try to add a description of how the gradient-based sensitivity score effects the model parameters through second order gradients when defining our proposed objectives.
(2) For the identifying attributes case, even though the influential region should support both correct and incorrect answers, it should *contribute more* to the correct one. The self-critical loss helps update model parameters to increase the right answer's sensitivity and decrease the wrong ones to the influential object until the model is right.

Comments on Measuring the Frequency of Wrong Answers with Valid Attention for R #2 & R #4.
We agree that a global measurement makes our claims stronger. In particular, we think our false sensitivity rate defined in Eq. 6, which computes the fraction of false sensitivity where the predicted incorrect answer's sensitivity to the influential object is greater than the correct answer's sensitivity, can partially help indicate that the false sensitivity is a common problem in VQA.

Comments on the discrepancy of the influential objects extracted from different explanations for R #2.
Thanks for the proposal of using Jaccard Distance to compare the object sets extracted with visual cues and text description in VQA-X validation and test set. We will include this measurement in our revision.

Comments on Significant tests for R #2 & R #4.
Thanks for mentioning the significance-test issue, we believe it would make our results more convincing. We will add error bars on the VQA scores to show the stabilities of our proposed objectives.

Comments on Various Ablation Study Issues for R #2 & R #4.
(1) *Influence Strengthening Loss and Self-Critical Loss*: Currently, we only show the results of using and not using the influence strengthening loss due to the page limits and focus more on the self critical loss, since the former is more intuitive and mentioned in HINT paper (Selvaraju et al., arXiv 2019). We hope to add the detailed ablation study on the influence strengthening loss to the paper in the final version for both completeness and further analysis. Also, as pointed out by Reviewer # 4, in order to analyze the behavior of both losses, we will measure the intersection over union (IOU) between the sets of QA pairs that are correctly answered trained with one of the two objectives but incorrectly answered using simply the original VQA objective.
(2) *Proposal Set Size* $|\mathcal{I}|$. Thanks for pointing out the missing details on $|\mathcal{I}|$. We set it to 6 in all of our experiments in the submission. Table. 1 reports the ablation results with various set sizes indicating the two objectives are fairly robust. We use VQA-HAT visual explanations to construct the influential object sets and both losses to fine-tune our model.

$\lvert\mathcal{I}\rvert$	4	5	6	7	8	10
VQA-CP v2 test	48.8%	49.1%	49.2%	49.1%	48.7%	48.3%

Table 1: Ablation study on the size of proposal influential object set.

Comments on Generic Questions for R #2.
(1) Modified Version of GradCAM. We use the same methods to compute the sensitivity score adopted in the HINT paper and we also empirically find that this modified GradCAM, which removes the ReLU, works better than original one. Our interpretation is that before ReLU, the negative values could provide counter-factual meanings to the prediction. Therefore, there is no need to filter them out using ReLU. However, in the original GradCAM visualization paper was more focuesd on visualizing the regions that positively contribute to the predictions.
(2) Model Complexity and Computation Cost. First, our model does not introduce parameters to the original VQA systems, and therefore the model parameter complexity remains the same. During forwarding and backwarding, original VQA system takes about 20min per epoch and with the two objectives the system takes about 25min using a TITAN V.
(3) False Negatives. Thanks for mentioning the false negative issue in the model. We think the impact of this issue is hard to evaluate since there are no gold standard object annotations. However, it is partially revealed by the ablation study on the proposal set size $|\mathcal{I}|$ in Table. 1, $e.g.$ larger size reduces false negatives but increases false positives.
(4) Glove Embedding Similarity. We did not do k-cross-validation as we do not have gold standard annotations on which word should be associated with which object and we simply inspect some examples manually.
(5) Other Influential Loss. We tried "mean" as the influential loss and found "min" works better. Our interpretation is "min" only requires at least *one* object to be influential while "mean" encourages all of the objects in the proposal object set to be influential, which is more vulnerable to false positives in the proposal set.

Comments for R #4.
(1) Quantification of L111-112. For visual explanations, we can ensure that the proposal set contains the regions that humans find most important. When using textual explanations and QA pairs, we will use Jaccard Distance (mentioned by R #2) to measure the similarity between the proposal sets in VQA-X validation and test set in our final version.
(2) Usage of Proposal Set. We currently use the three kinds of proposal sets separately denoted under "Expl." in tables.