We thank the reviewers for their useful comments and suggestions. We are glad that the reviewers found our approach to be novel (R2, R3, R4), general and significant (R4), a valuable contribution (R2), appreciated its superior performance (R2, R3, R4), and found our paper to be clear (R2, R3). We now address their requests and concerns.

Answers to R2:

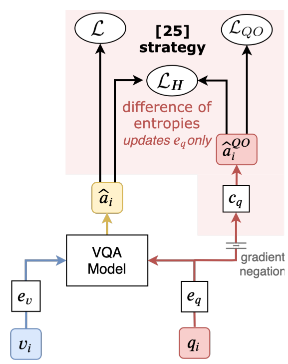
- **Q1 Additional baselines:** We ran this baseline of sampling answers from a uniform distribution. This gets an accuracy of 40.25% (compared to 47.11% with our approach using the same baseline architecture). As a recall, our current baseline gets 38.46%. Inspired by this suggestion, we also tested sampling answers from a uniform distribution per question-type. This gets an accuracy of 42.11%. We will add these two new baselines in Table 1.

- **Q2 Grounding ability, interpretability and future works:** We ran new experiments on the VQA-HAT dataset to quantitatively validate that models trained with the RUBi strategy on VQA 1.0 improves the ability to *attend to the "right" regions of the image*. We report 0.4551 in rank-correlation (higher is better) with our baseline architecture and 0.4671 when trained with RUBi (see Table 2 in VQA-HAT paper for reference; recall that we use image features from [15]). Interestingly, our approach improves the grounding ability without being designed to do so explicitly. We will add a new table of results on VQA-HAT including different architectures, as well as qualitative results similar to the attention maps from Figure 6 of the VQA-HAT paper. These visualizations will allow us to discuss about interpretability and grounded/symbolic reasoning. Also, we will add details about future works in the conclusion.

Answers to R3:

- **Q1 Significance of c'_q:** We ran new experiments to evaluate the usefulness of c'_q. First, we fixed c'_q to be the identity (i.e. we removed c'_q while c_q receives gradients from L_{QO}). We report an accuracy of 5.38% on VQA-CP v2 with our baseline architecture. This low performance is expected since c_q is designed to output a 0-1 mask using the sigmoid, and not to output logits. We agree that the term "classifier" to define c_q was unclear. We will change it. Secondly, we removed both c'_q and the question-only loss L_{QO}. We report a slightly lower accuracy of 46.08% (-1.03 compared to a training with the full RUBi strategy) for the baseline architecture. Intuitively, the 0-1 masks produced by c_q must be good enough to reduce the importance of biases early during training. c'_q and L_{QO} provides an additional supervision to c_q helping it to generate better masks, earlier in the training. We will add a new table of results about c'_q. We will also improve the discussion about c'_q and L_{QO}.

- **Q2 Comparison with other candidate models:** We experimented with different fusion techniques to combine the output of c_q with the output from the VQA model. For instance, a ReLU instead of a sigmoid gets 40.02% (compared to 47.11% with our approach using the same baseline architecture). Other classical fusions such as an element-wise sum lead to more significant performance drop than what was previously reported with ReLU. Upon acceptance, we will add a detailed discussion about these fusions in the final paper.



Answers to R4:

- **Q1 Visual comparison to [25]:** We will add to Figure 2 an *"apples-to-apples" comparison to [25]* as depicted in the figure of this rebuttal. Similarly to the "gradient negation" illustration, we will improve Figure 2 to indicate *when the backpropagation is not happening* in e_q. We will also clarify the comparison with [25], from line 113 to 122.

- **Q2 Clarification about c_q and c'_q:** We will clarify that c_q receives gradients from L_{QM} and L_{QO}. See the answer Q1 to R3 for further information about c_q and c'_q.

- **Q3 Evaluation on VQA-CP v1 and detailed evaluation breakdown:** We ran new experiments on VQA-CP v1 and report state-of-the-art results regardless of the architecture trained with RUBi. Our approach consistently leads to significant gains over the classical learning strategy. We report improvements of +9.80 in overall accuracy with our baseline architecture, +10.46 with UpDn, +19.23 with SAN. We will add a new table of results on VQA-CP v1 similarly to Table 1. We will also include the accuracy for each answer types for the UpDn and SAN architectures in Table 2.

- **Q4 Discussion about [A,B,C] and prior approaches:** We will add [A,B,C] to the related works section to highlight the importance of biases reducing methods in the multimodal context. Finally, we will introduce [15,41,19,16] from Table 1 in the state-of-the-art comparison paragraph. Note that these previous approaches do not focus on biases reduction contrary to [25].

Model	Overall
GVQA [10]	39.23
SAN [26]	26.88
+ [25]	43.43
+ RUBi	**46.11**
UpDn [15]	37.15
+ RUBi	**47.61**
Baseline	37.13
+ RUBi	**46.93**