Submission 180: Author Response

We thank the reviewers for their thoughtful comments. We are encouraged to receive positive feedback from all reviewers regarding the importance of investigating the value of out-of-distribution testing and discussion of the issues associated with it. Reviewers have described our work as "extremely important in that it provides a reality check for prior works on VQA-CP (R1)" and "targets on an important aspect of current machine learning methods (R2)". The reviewers also unanimously agreed on the clarity and correctness of the paper and our empirical evaluation. Reviewers have described our proposed baseline models and experimental validation as: "Interesting and novel" (R2), "simple and effective" (R3), "reasonable and convincing" (R2), and that the "experiments sufficiently validate the hypotheses"(R1).

The reviewers have also raised some concerns and issued some constructive suggestions about our work, which we address in the comments below. Reviewers' comments have been paraphrased for brevity.

R3: *It looks like the random image regularizer hurts in-domain performance.*

Response: Firstly, the random image regularizer is designed to showcase how exploiting the fact that the test-set follows approximately the inverse distribution can be exploited and has no practical use case. As discussed in L258-260, it is possible to tune the trade-off between in-domain and out-of-domain performance by tuning the hyperparameter λ for the random image regularizer. A part of the problem we want to highlight is that most algorithms do not even report the in-domain performance before retraining and our results clearly show that it is possible to obtain really high performance on the OOD split by sacrificing in-domain performance.

That being said, our random image regularizer works on-par or better than existing methods on in-domain performance reported before retraining [1, 2]. Table 4 only shows lambda = 5 and 12, which indeed show that it lags behind in in-domain setup (while far surpassing the OOD setup), but lower values of lambda (Figure 4a) shows that it is possible to have higher in-domain performance. E.g., at $\lambda = 2$, our results exceed [1] and are on par with [2] on both in-domain and out-of-domain splits.

R3: *Do other VQA datasets (e.g., GQA, VCR) have the same problem?*

Response: Neither GQA nor VCR contain an OOD test split and therefore they are not capable of measuring OOD performance. However, if a split of the dataset was made in a manner similar to VQA-CP (lacking in-domain holdout set, OOD test set approximately inversely distributed as train) and the algorithms made similar assumptions (access to knowledge about the construction of test set, evaluating in-domain performance after retraining), our findings on the pitfalls of OOD testing would readily apply to any other VQA dataset. As discussed in our recommendation section, alleviating these concerns calls for a radically different approach to constructing datasets for OOD testing for VQA that is not present in any of the current VQA datasets to our knowledge.

R2: *Do other datasets for OOD evaluation have similar problems like VQA-CP?*

Response: The pitfalls we identified with VQA-CP and the methods evaluated on it are relevant to OOD testing in general. The problems stem from the three key issues we discussed in the paper and are largely model and dataset agnostic. After our submission, some other papers [3] have pointed out similar issues with other datasets. We will summarize these observations in our final version.

R1: *I encourage the authors to report some notion of statistical significance.*

Response: We agree with the spirit of the suggestion. In the final version, we will report statistical tests for variations of our baseline algorithms as well as comparison models, wherever appropriate.

References

[1] Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. Don't take the premise for granted: Mitigating artifacts in natural language inference. *arXiv preprint arXiv:1907.04380*, 2019.

[2] Anton van den Hengel Damien Teney, Ehsan Abbasnejad. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020.

[3] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.