Thank you all for the thoughtful reviews! We will respond in plural form to maintain blindness. We will first address issues brought up by multiple reviewers, followed by individual queries.

Regarding the definition of reproducibility: We regarded a paper as reproducible if most (75%+) of the claims in the paper could be reproduced. If a claimed improvement was measured in orders-of-magnitude, being within the same order-of-magnitude was sufficient (e.g., a paper claims 700x faster, but we observer 300x, still qualifies). When compared to other algorithms, we consider a paper reproduced if most (90%+) of the new algorithm's rankings correspond to what was in the paper (e.g., new method was most accurate on 95% of tasks compared to 4 other models, we want to see our reproduction be most accurate on at least 95%*90%=81% of the same tasks, compared to the same models). As a last resort, we considered getting within 10% of the numbers reporting in the paper (or better), or in the case of non-quantitative results (e.g., GAN sample quality), we subjectively compare our results with the paper to make a decision. We will include a version of the above in the revised paper.

Regarding Venue: While a "science of science" type venue would be appropriate, we feel NeurIPS is a more appropriate venue. We are more concerned with the nature of reproducibility within our specific field, rather than broader genres such as computer science or science generically. We feel the discussion appropriate given the current high and growing discussion of these issues within the field, and necessary to build communal momentum around larger organized efforts to track and quantify reproduction (as we discussed regarding the ICLR Reproducability Challenge). While we consider our work valuable, issues in study bias will persist until we get a larger pool of implementers under consideration.

Regarding Author/reproducer details: We are trying to maintain double-blindness as much as possible in our replies. The camera ready will detail this and reproducer background extensively.

To R#1: We would have also liked to do a textual analysis of papers, and hope to do so in the future. However, first attempts produced significant issues with respect to parsing reliability (some papers are photo-copies, some are OCRed, some are poorly formated PDFs, some are nicely formtated PDFs, etc) that would confound results and increase analysis difficulty, not to mention take considerable time to get working reliably. This also prevented us from automated the equation counting.

Re line 212: you make an excellent point, and we will add caution to the final manuscript regarding that statement.

"*Furthermore, the reader would also benefit from a rough classification of reasons for non-reproducible papers*": We did not record this in detail for each paper, but agree would be valuable! We will add a discussion of the general types of issues we encountered in the final version (though unfortunately non-quantified). Subjectively, the below list would be the primary issues that gave us reproduction problems, which we will elaborate further in camera ready. None of the below issues were mutually exclusive.

1. Unclear notation or language. A component of the algorithm is explained, but not in a way easily understood by the reproducers, or was ambigiously specified.
2. Missing algorithm step or details. A step was completely left out of description.
3. Results left as an exercise to the reader: many papers would specify loss functions or other equations for which the gradient needed to be taken, but then not detail the resulting gradients. Depending on the functions and math involved re-deriving was non-trivial.
4. Missing hyper-parameters, or similar nuance details. We appear to have an implementation accurate to what was described, but some minor detail was not specified and makes a big difference in results.

We understand knowing the number of reproducers / backgrounds is intrinsically valuable for our paper, but also violates double blind review. Reproduction attempts/effort was approximately uniformly distributed between reproducers.

To R#2: The effort was indeed significant, and only possible because we used software early on that recorded much relevant information. Between efforts done as part of education, job, and for fun, back-of-envelop guesstimation puts our effort at $\approx$10,710 hours per author.

To R# 3: On computing "*how long it too to reproduce*", this is a very good idea, thank you! We think we could do this for *most* of our papers that were reproduced, as most where made open source and we can compare our start date to the commit date to get an approximation. We can't do it for all though as some remain closed source, and we would be unable to compare to unreplicated papers since we did not track that information.

We can't get this done by the end of rebuttal time, but will consider it for camera ready - we aren't sure how long this would take us (given day jobs as well!). We also want to give it the same thought and consideration we gave the factors already in the paper, and not rush analysis without considering confounding factors.