We thank the reviewers for their enthusiastic feedback and insightful suggestions. We appreciate the recognitions of contribution in the reviews, such as “the methodology is meticulously detailed, and the work’s contributions to research and development of better generative models, an issue of major importance in machine learning," that HYPE as a “benchmark for generative models can be very useful for the community to enable more consistent evaluation of new methods," and that there was “[s]olid writing and justification for HYPE as a benchmark." We also appreciate the questions and concerns raised in the reviews, as well as the requests and opportunity for clarification in our author feedback response. We address them as follows and in our corresponding revision.

**[R1, R2] Proposed diversity incorporation.** R1 and R2 ask how diversity may be incorporated into a full ranking of the GANs. As mentioned in our submission and acknowledged by R1 and R2, the current scope of HYPE is limited to perceptual realism, though we agree that diversity measurement is worthwhile. Similar to HUSE (Hashimoto 2019), we suggest that diversity can be computed using the automated recall score (Sajjadi 2018). Recall measures diversity independently from precision ($F_{1/8}$), the corresponding measure of fidelity in the paper. We will include this comment in our paper, citing both works.

**[R3] Related Work.** We will update our draft with an extended discussion on related work, stating how HYPE differs, beyond FID, KID, and $F_{1/8}$. Broadly, HYPE measures human perceptual judgments of generative outputs using humans directly in an inexpensive, widely accessible method. This contrasts with Neuroscore (Wang 2019) because it can be widely accessible by researchers rather than depend on humans with EEGs properly worn and measured. Additionally, the automated metrics 1-NN and Kernel MMD, suggested for evaluation by Xu et al., show success only in a limited architectural case (the convolutional space of an ImageNet-pretrained ResNet) and nevertheless indirectly assess human judgment.

Specifically, we address limitations of automatic metrics. Prior work has asserted that there exists coarse correlation of human judgment to FID and IS, leading to their widespread adoption. However, both metrics depend on an ImageNet-pretrained Inception v3 Network to calculate statistics on the generated output (for IS) and on the real and generated distributions (for FID). The validity of these metrics when applied to other datasets has been repeatedly called into question (Barratt 2018, Rosca 2017, Borji 2018, Ravuri 2018). Perturbations imperceptible to humans alter their values, similar to the behavior of adversarial examples (Kurakin 2016). Finally, because FID is a biased estimator, there is inherent variance to the metric depending on the number of images and which images were chosen—in fact, there exists a correlation between accuracy and budget (cost of computation) in improving FID scores, because spending a longer time and thus higher cost on compute will yield better FID scores (Lucic 2018). KID addresses this as an unbiased estimator (Bińkowski 2018), but otherwise strongly correlates with FID and not human judgment, as we report in the paper.

**[R1] Impact of number of samples per evaluator.** Similar to our earlier evaluation on increasing the number of evaluators, we find that the CI decreases when increasing the number of samples that each evaluator assesses. Computed via a bootstrap, find that the CI width decreases monotonically when increasing the number of images, specifically from $10.5$ to $8.5$ when evaluating $20$ versus $100$ images. We choose $100$ images for HYPE infinity because we find it to be a good combination of reliability and efficiency for the GANs we evaluated. This number can be treated as a hyperparameter, and may be increased or decreased based on the desired reliability and quality of GANs evaluated.

**[R2] Meaning of HYPE scores.** As R2 points out, HYPE is designed to rank GANs; deltas in HYPE scores may not reflect an equal delta in visually perceived realism. However, we note that HYPE scores are directly correlated with perceived realism, though that correlation may not be linear. Additionally, an absolute HYPE infinity score above $50$ meaningfully indicates “hyper-realistic” images, or images that appear more realistic than real images.

**[R2] Discussion on differences from automated metrics.** R2 requests an extended discussion on results from automated metric differences. We will include comments in the paper on qualitative differences that we believe exist between HYPE scores and $\{$FID, KID, $F_{1/8}\}$ scores.

**[R2] Dataset and code release.** We plan to release the dataset and code for researchers to analyze after publication.

**[R2] Explicit 50:50 real:fake ratio.** R2 mentions that revealing the ratio to evaluators may bias them. We made this design decision intentionally and carefully. Amazon Mechanical Turk forums would enable evaluators to discuss and learn about this distribution over time, thus altering how different evaluators would approach the task (this discussion occurs often). Thus, we decided to make this ratio explicit such that evaluators would have the same prior entering the task. We also make this decision to evaluate a threshold for which fake images appear more realistic than the real ones.

**[R2] Hyper-realism with unrealistic real datasets.** R2 mentions the unrealistic appearance of some real datasets such as CIFAR-10. We acknowledge this, and it enables varying levels of difficulty across datasets. Thus, “hyper-realism” is relative to the real dataset on which a model is trained. Some are easier because of lower resolution and/or lower diversity of images.