We want to thank the reviewers greatly for the time and effort put into these reviews. Each of you has provided a very detailed critique, and we appreciate your help in presenting our work as best as possible.

We are encouraged that reviewers found Gradient Boosted Normalizing Flows (GBNF) to be novel and significant (R3, R4), and that the "conceptual difference to previous research is a big strength." (R4). Reviewers appreciated that our "wider" approach to building normalizing flow-based models is more than just a way to improve performance, noting that sampling is trivial to parallelize across boosting components (R2, R4). Reviewers (R1, R3, R4) found the writing and derivations to be overall clear and concise (R4 kindly provides a few notes to improve clarity), and we are pleased that R3 found the proposed method easy to grasp but recognized GBNF as non-trivial to formulate or implement.

[R1] "Contrast your method with Boosting Density Estimation, Boosted Generative Models, and boosting methods for variational inference" Our work uncovers challenges that are unique to boosting on normalizing flows. Specifically, for the variational inference setting we address three challenges in augmenting the VAE with a GBNF approximate posterior:

1. Only analytically invertible flows can be boosted for variational inference (Section 5.1, and Figure 2)
2. The "decoder shock" phenomena occurring from sudden changes to the VAE's posterior as new components are added (Section 5.2), and mitigating strategies.
3. Accelerating training by stochastically approximating the GBNF variational posterior (lines 148-152).

In regards to R1 and R3's critique on further differentiating our work with boosted density estimation [Rosset-Segal, '02] and generative models [Grover-Ermon, '18]: We show that the change-of-variables formula can be recursively computed in GBNF, and GBNF's mixture formulation remains invertible. Further, our work clarifies the narrative by providing a derivation for the boosting component updates and establishes a connection with Frank-Wolfe, whereas [Grover-Ermon, '18] merely state their choice of surrogate loss.

[R2] "No increase in theoretical flow expressivity" We felt that proofs of boosting's expressiveness to be outside the scope of our paper. [Guo, '16] address the limit as the number of components $c \to \infty$, and [Cranko-Nock, '19] analyze properties of boosting for density estimation. While more work is needed in understanding these algorithms (R4 also suggests an important direction for future work), we focus on the *unique algorithmic challenges* of boosting normalizing flows for density estimation and variational inference.

We do, however, believe the composite model learned by GBNF is a fundamentally different class of transformation relative to the single component model. R2 writes "there are two bottlenecks in NF expressivity—the base distribution and the class of transformation function [Papamakarios et al., '19]—and the proposed method does not fundamentally change either of these." R2 references recent work [Dinh et al., '19, Cornish et al., '20] using mixture formulations (thank you, we will include these in our related works) that address the NF bottlenecks. GBNF also takes the form of a mixture, and offers a number of advantages over a standard mixture formulation:

1. GBNF is easier to train—we only need updates to the newest component (lines 102–109).
2. Easy to add more components (and flexibility) to an existing model, costing only additional training time (lines 95–97).
3. GBNF can learn a variational posterior, unlike [Cornish et al., '20].
4. GBNF optimizes a different objective that fits new components to the residuals of previously trained components (Eq. (6), lines 143-147), which can refine the *mode covering* behavior of maximum likelihood (similar to [Dinh et al., '19]) and VAEs (Figure 1).

[R1] Correctness of (10) We appreciate reviewers taking the time to check for correctness. We stand by Eq. (10), and clarify that we are taking the gradient of the loss (9) with respect to $G^{(c)}$ at the point $\mathbf{z}$ in the function space, as opposed to the full distribution over random variable $\mathbf{z}$. We are left with the fixed components $G_K^{(c-1)}$ in the denominator because for a small step-size ρ_c we have $G^{(c)}|_{\rho_c \to 0} = (1-\rho_c)G^{(c-1)} + \rho_c g^{(c)}|_{\rho_c \to 0} = G^{(c-1)}$.

[R1, R3, R4] A dedicated Related Work section We agree that a separate Related Work section will improve the readability of our manuscript. At present, comparisons to related work are scattered: relevant boosting approaches are listed in the Background section (lines 75–81), and in Section 5.1 many state-of-the-art flows are discussed and assessed based on compatibility with GBNF. We will reformat our manuscript to highlight these related works, and clearly differentiate our work from boosted density estimators [Rosset-Segal, '02, Grover-Ermon, '18], and normalizing flows using mixture formulations [Dinh et al., '19, Cornish et al., '20].