We thank the reviewers for their constructive feedback and unanimous acceptance of our theory on general f-VI. We notice that *a big concern among all reviewers is about our experiments, as the reviewers commented that the experiments i) did not interpret the improvements of f-VI against the well-established VI methods, ii) need to be tested on more challenging datasets, and iii) should include more custom f-functions.* We would like to mention that our main contribution is to provide a solid theoretical foundation and two (stochastic and mean-field) optimization schemes for VI subject to all f-divergences, and our experiments then serve to verify the correctness and feasibility of these results. While proposing a specific f-divergence VI that surpasses the existing VI methods is a significant task, the workload of it deserves an independent and thorough study as in [16, 17], and the current paper can guide and facilitate this task by offering a general f-VI framework. Meanwhile, our experiments (Section 4.2 and 4.3) suggest that the performance of different f-divergence VIs varies by the training model as well as the dataset; *e.g.* the custom f-divergence VI showed some quantitative improvements against other well-established f-divergence VIs in the BNN experiment on Fish Toxicity and Stock datasets, while it slightly underperformed in the other empirical examples; similar results were also reported in [2, 3], which compared the KL-VI and some Rényi's α-VIs. Hence, a (custom) f-divergence VI that consistently dominates other well-studied f-VIs is almost unforeseeable. However, as per your request, we will add the following new examples:

1) Two new custom f-functions, $f^*(t) = |t-1|$ for total variation distance and $f^*(t) = -\log(2t+1) - \log t + \log 3$, are tested in the BNN experiment of Section 4.2 and the VAE experiment of Section 4.3. The results are appended in the Supplementary Material (SM). Some strategies for generating a valid custom f-function are also supplemented.
2) Three additional datasets, Frey Face, Caltech 101 Silhouettes and Omniglot, are tested in the VAE experiment of Section 4.3. Reconstruction errors and some reconstructed and generated images are added and compared.

A minor comment from all reviewers is about the introduction section, as the reviewers pointed out that the introduction i) did not have a dedicated related work section or background on VI, and ii) lacked a deeper discussion when reviewing prior work on VI and f-VI, such as [15, 17, 18]. We referred the readers without any experience in VI to two recent survey papers [11, 12] for VI background. Prior work on f-VI either investigated only a specific family of f-divergence [16, 17] or circumvented the fundamental setting of f-VI [18], *i.e.* minimizing $D_f(q(z)|p(z))$. Hence, none of them can unify the existing VI methods and also be applicable to all f-divergences. More deeper discussions on the prior work and VI background will be added in either the introduction or the SM. *Other comments with regard to the Boarder Impact and writing* will be addressed in the camera-ready paper. The following is our response to the individual questions raised by the reviewers:

***R2**: Although useful for completeness, some parts of Section 3 (mostly 3.3 and the first half of 3.2) repeat previous results.* - The general f-VI algorithms presented in Section 3 are new and have not been reported before, although with proper f- or f^*-functions assigned, they restore many well-known Bayesian approximation algorithms, *e.g.* examples in Section C and D. New Bayesian approximation or VI algorithms can also be generated from the results in Section 3, if we assign the f-VI algorithms in Section 3 with new (custom) f- or f^*-functions.

***R2**: What is the intuition regarding mathematical results that could help to broaden the potential impact of f-VI?*
***R3**: Why and how to use the proposed f-VI? In the experiments, it seems the commonly used KL-VI works the best in general. So, what are the advantages of f-VI over existing methods, in addition to being general?* - We will answer the preceding two comments collectively. An explicit advantage (or impact) of f-VI is that it allows to perform Bayesian approximation or VI with more variety of divergences, which could potentially bring us sharper variational bounds, faster convergence rates, smaller RMSE and reconstruction error as in our experiments. The empirical performance of different f-divergence VIs varies by the training model and datasets, while the logarithmic KL-function makes the KL-VI more numerically stable and accurate when $p(z, \mathcal{D})/q(z)$ is tiny, which can be a guideline for picking f or f^*-functions.

***R3**: Based on Eq. (2), to derive Eq. (8) is straightforward, as revealed in Line 405 of the SM. Why spent so much effort to circle around from a dual space?* - Eq. (8) is derived by minimizing the reverse f-divergence, while Eq. (18) is from the forward f-divergence. Eqs. (8) and (18) are different, once you expand the dual functions in Eq. (8) and consider the constraints on f- and f^*-functions. Since the existing VIs based on the reverse divergences generally have better statistical properties [18, 20], we derived the f-variational bound Eq. (8) in a dual space, which also makes our results more compatible and consistent with the existing VI algorithms. However, both Eq. (8) and (18) are essential parts of our f-VI framework, as the mean-field update rules Eq. (15) and (16) are respectively derived from Eqs. (8) and (18).

***R3**: Line 179, why treating the parameters ϕ as latent variables z helps reduce variance?* - Treating the parameter ϕ in $p_\phi(z, \mathcal{D})$ as latent variables does not necessarily help reduce the variance. We will withdraw this statement in the camera-ready paper.

***R3**: Line 200. The word "two-dimensional" is confusing, because in general, noise samples need not be 2-dimensional.* - "Two-dimensional" describes the size of noise samples $\{\varepsilon_{k,1:L}\}_{k=1}^K$ in the importance-weighted estimators, Eq. (14) and Line 143. We will withdraw this word in the camera-ready paper.