First, we thank all reviewers for their time and valuable feedback. We reply to each reviewer individually and then comment on the significance of the work.

To **Reviewer 1**: Yes, you understood the paragraph on lines 56-59 correctly. Optimizing the lower bound does not imply maximization of the log marginal likelihood; thus we should seek to close this gap as much as possible by choosing a flexible approximate posterior $q(z|x)$. If $q(z|x)$ is limited in its expressivity, the true posterior would also have to be simplistic, since the KL divergence gap forces the two to be close. We will rewrite this paragraph to make it more clear for the camera-ready version (This paragraph is also relevant to **Reviewer 4**).

We will add the missing citation of Naesseth et al. (2018) and fix the legend and the typo.

To **Reviewer 3**: yes, the key idea is to learn the auxiliary variational distribution $\tau(\psi|z)$ so that the samples $\psi_{1:K}$ are coming from the high-probability region of the optimal distribution $q(\psi|z)$. In contrast, SIVI uses samples from $q(\psi)$, which are uninformed about the particular z and thus much more samples are needed to achieve the same quality, as we have shown empirically (please see the paragraph on significance for more details).

Molchanov et al. (2018): the scheme of their proof of the Theorem 1 of DSIVI is not directly applicable to IWHVI. Also, they require the bound to be averaged over $q(z)$, whereas the IWHVI gives a valid upper bound for any fixed z.

We are very thankful for the alternative proof of the Lemma C.2. While the existing Lemma is valuable as it provides an insight into the underlying generative process (Self-Normalized Importance Sampling), we will add the suggested proof for a reader's convenience.

To **Reviewer 4**: we introduced the term log marginal density to avoid confusion with the log marginal likelihood which is usually assumed to be $\log p(x)$ – the model's likelihood of the observed data, even though both terms mean logarithm of a marginal distribution of some joint distribution $\log \int p(\alpha, \beta|\gamma) d\beta$. Although the IWHVI bound could indeed be used to give an upper bound on the log marginal likelihood $\log p(x)$, it would require intractable and impractical posterior sampling. The paragraph starting on line 205 indeed meant the upper bound on the log marginal density, and we are thankful for pointing out the typo.

Missing related work: we will add citations, but they are outside of the scope of the paper: the Nonparametric VI only works for finite mixtures, and Likelihood-Free VI uses GAN-like estimation and loses lower bound guarantees in case of a suboptimal discriminator.

K schedule: we observed it was beneficial to use as large K as is affordable, however large number of samples seemed to cause computational instabilities in the early stages of training. 90% of training is done with $K = 50$.

UIVI: we omitted it because it did not scale to our VAE experiment setup. From the algorithmic point of view, MCMC-based methods are inherently sequential and thus are not amenable to parallelization. In particular, UIVI uses 5 samples from HMC with 5 leap-frog steps, and 5 burn-in steps, which results in 5*10 backward and 5*10 + 10 forward (including MH correction) passes done sequentially. Even if one runs 5 independent chains to obtain 5 samples in parallel, the burn-in is still required to decorrelate ε and ε', limiting the potential speed-up to 2x. In contrast, IWHVI/SIVI are much more parallel, as samples are independent and can be processed simultaneously. Originally UIVI was benchmarked on CPU in which case all methods indeed perform comparably time-wise (UIVI: 0.18sec/iter, IWHVI K=100: 0.14s/it, SIVI K=100: 0.13s/it). However, in our implementation in TensorFlow (with TensorFlow Probability for MCMC) for GPU we observed that IWHVI (0.02s/it) / SIVI (0.02s/it) are almost 8-9 times faster than UIVI (0.13s/it), and 7 times faster if one runs 5 parallel chains (0.11 s/it). We can add such experiments by the camera-ready deadline though.

Run times: are listed in the previous paragraph and are in absolute agreement with our discussion of complexity in the sec. 4.1, which predicts the actual SIVI/IWHVI run times to be indistinguishable.

Finally, we would like to address the **significance of the work**. We believe that our method significantly improves upon SIVI for several reasons. ①: it provides an upper-bound analog of the IWAE lower bound, whereas SIVI only gives an upper-bound analog of the IWAE lower bound for a special case of *the proposal distribution* $q(z|x)$ *fixed to be the prior* $p(z)$. This is not how IWAE bounds are typically used. ②: using the prior as a proposal forces the generative model of x to adapt to such choice, that is, the true posterior $p(z|x)$ has to be close to the prior $p(z)$ (because of the gap), which negatively affects the mutual information (MI) between x and z and thus diminishes effectiveness of latent variable modeling (this is because each z carries little information about x). Similar reasoning holds for SIVI/IWHVI, where we consider a (conditional) generative model with joint $q(z, \psi|x)$, true intractable "posterior" $q(\psi|x, z)$ and the "prior" $q(\psi|x)$. We validate this intuition with empirical results in fig. 1b: IWHVI with 5 samples has MI between ψ and z two times higher than SIVI with 50 samples. ③: we also note that despite IWHVI's similarity to IWAE lower bound and relations to SIVI/HVM, it is not obvious how to obtain the IWHVI bound in the absence of our theorem. We also render HVM and SIVI as special cases of a more general technique.