[Related work] We agree our work is related to MC methods on augmented spaces and will add a more discussion of this to the paper. One distinction is that augmented MC methods traditionally use particle-based estimators where the *choice* of coupling is "obvious", but the *proof* much less so (as we briefly allude on lines 41-44, 112 and 138-145), and require case-by-case derivations if a new estimator is introduced. We start with an arbitrary estimator R, for which is is not clear *a priori* that a coupling *exists*, and provide a systematic approach to finding estimator-coupling pairs.

In more detail, the closest MCMC work is "Particle Independent Metropolis Hastings" (Andrieu et al. JRSSB). It says (in our terminology): take an estimator for $p(x)$ defined by $Q(\omega)$ and $R(\omega)$ that also comes with an "obvious" distribution $a(z|\omega)$ for sampling z (i.e., select one particle or trajectory in proportion to its weight). Define the extended proposal distribution $Q(z, \omega) = Q(\omega)a(z|\omega)$ and the (unnormalized) extended target $P^{\text{MC}}(z, \omega, x) = R(\omega)Q(\omega)a(z|\omega)$. Run independent Metropolis-Hastings (MH) with extended target and proposal to sample from $P^{\text{MC}}(z, \omega|x)$. The acceptance probability is $\min(1, R(\omega')/R(\omega))$, so can be computed as simple byproduct of generating the proposal. Further, one can show using properties of particle-based estimators that the "obvious" distribution $a(z|\omega)$ is indeed a coupling, i.e., $p(z|x)$ is a marginal of $P^{\text{MC}}(z, \omega|x)$, so by discarding the ω variables we have a valid MCMC sampler for $p(z|x)$. The underlying reasoning is the same as in our Theorem 2: R is the ratio of the extended proposal and target densities used within MH. Our work can be viewed as the "VI side" of this work — what happens if we use extended proposals and targets within VI? In MCMC, dropping auxiliary variables automatically yields a valid marginal sampler. In VI, it introduces the conditional divergence in the ELBO decomposition (Theorem 2).

R2 also mentions the recent ICLR workshop paper of Lawson et al. (which appeared in early May and is concurrent to our work) and augmented VI more broadly. Our paper has strong roots in augmented VI and we certainly think of it as such. We hint at this in a few places (lines 8, 11, 122, 156) this but can (and will) make the point more explicitly. We build most closely on ref [7], which clearly articulates the special case of Theorem 2 for IWAEs (i.e., "IID Mean" in our Table 2) as augmented VI. Lawson et al. arrive at a decomposition of $\log p(x)$ analogous to our Theorem 2 but from a very different standpoint. They assert that "many unbiased estimators can be justified as performing simple importance sampling on an extended state space", and assume knowledge of the relevant extended and conditional distributions (cf. the form of $\hat{p}(x)$ at the end of Sec 2.2) — essentially a coupling in our terminology. They then *derive* the distributions (only) for IWAEs, as was done in ref [7] (also [6], [14]). The details for other objectives are left unstated and require case-by-case derivations. In contrast, we provide general tools to find estimator-coupling pairs without case-by-case derivations. We also believe that our framework of estimator-coupling pairs much more explicitly and clearly articulates the ingredients needed for such an approach to work.

[Experiments] We accept the view of the reviewers that our experiments did not make our points convincingly enough. We will start by reformatting the presentation. Our primary point was that better likelihood bounds ($\mathbb{E} \log R$) correspond to better posterior approximations, thus the two axes in Fig. 8. However, due to the large amount of model-to-model variability in these two metrics, it is difficult to see the differences due to methods. We plan to move Fig. 8 to the appendix and instead show results like Fig. 6 for more models — several examples are shown below (for likelihood bounds). These show differences due to sampling strategies and numbers of replicates. We would like to emphasize that iid sampling consistently improves on naive VI (equivalent to M=1) and that the alternative sampling methods offer a further consistent improvement at near-zero computational cost. (Often, M=8 or M=16 with iid sampling is needed to equal the performance of antithetic sampling with M=2.) We will also create analogous plots that aggregate across models by normalizing/standardizing the two metrics (likelihood bound improvement, posterior error).

[Writing] The reviewers made many helpful points regarding the organization and writing of the paper, as well as pointing out typos. We will revise the paper with these in mind, focusing especially on providing details and summary of experimental findings, and de-emphasizing some of the current content in Sec. 5.