We are very grateful to all the reviewers for their thoughtful feedback. Below we address the main questions and comments, and identify topics we will expand on in the final paper. All typos and minor points will also be fixed.

Theoretical questions on Prop. 3. The sequence of distributions that converges weakly to π is $\{(\mathfrak{T}_\ell)_\sharp \rho\}_{\ell \geq 1}$ as the number of lazy layers increases, i.e., $\ell \to \infty$. Crucially, condition (6) must apply simultaneously to *all* layers for a given $0 < t \leq 1$, rather than one particular layer. In the future, we plan to generalize our result by allowing $t \equiv (t_\ell)_{\ell \geq 0}$ to be a sequence that goes to zero sufficiently slowly to maintain the weak convergence of $\{(\mathfrak{T}_\ell)_\sharp \rho\}_{\ell \geq 1}$. In the case of $r = 1$, each layer operates on a one-dimensional subspace, but each subspace can be different, and thus we can capture the posterior in the limit. We do not claim any specific rate of convergence for the approximate posteriors based on t, though it is reasonable to suspect that t close to 1 will yield faster convergence. We are currently considering methods for estimating t in future work, which would lead to empirical studies of the convergence rate and t.

Greedy sub-optimality. Prop. 3 implies that any inference problem can be decomposed into a sequence of r-dimensional problems, and that the limit of this sequence is exactly the posterior. Therefore, while our greedy method is certainly not optimal for a given length ℓ, we can't provide an example where a certain rank r leads the approximate posterior astray in the limit $\ell \to \infty$. However, deeply lazy maps can sometimes suffer from a certain greedy sub-optimality with regard to the rank and the choice of transport class. For example, the target distribution in the toy problem of §4.1 can be captured with a single full-dimensional quadratic map. By choosing $r = 1$, we instead need a sequence of maps to capture the target. This is an example where the underlying problem does not have immediate lazy structure, and applying an overly lazy map requires extra work with no pay-off. This naturally raises the question of how to choose r. The spectrum of H is a natural guide: if it decays quickly, then one can fix r using an error threshold as discussed in Prop. 2. Otherwise, the choice of r defines a trade-off between how expensive each layer is to train and how many layers one needs to train. One strategy is to limit r by some r_{max}, which defines a maximum computational budget for each layer. Another consideration, as highlighted by the example of §4.3, is that reducing the dimension of each layer can *improve* training behavior. The missing square in the posterior realizations of the diffusion field in §4.4 is in fact a property of the true posterior; we have verified this fact with MCMC results.

Comparison to MCMC. Finding a fair comparison between MCMC and VI is a universal and interesting question in Bayesian computation, as the two methods have different computational cost patterns. In VI one spends considerable computational effort to construct the approximate posterior, but afterwards has cheap access to (approximate) samples and normalized density evaluations. How well the approximation matches the true posterior depends on the expressiveness of the transport map/flow and the ability to optimize the map—two qualities that the lazy map framework seems to improve. MCMC methods in contrast require continual computational effort (even after tuning), but (asymptotically) generate samples from the exact posterior. We will discuss these trade-offs to additionally frame our contribution in the final paper. The strongest comparison we believe we can make is comparing the ESS of an MCMC method with and without transport preconditioning (as discussed in [23, 41] and in §4.4). We will make these improvements a more central measure of success and will also report the improvements when applied to the examples of §4.2 and §4.3.

Computing H considerations. The cost of forming H scales with the cost of computing the gradient of the unnormalized target log-density at a sample. This is required for each optimization step as well. For the example of §4.2, the cost of computing H with 500 samples is the same as that of 5 optimization steps with 100 samples each, a relatively minor cost compared to 20000 total optimization steps. The cost of identifying the dominant eigenspace is also small in comparison. In the final paper we will include this cost in the training plots in Appendix G, either by plotting against wall clock time or the number of gradient evaluations. We do accept that the cost of map evaluations increases as the lazy map grows deeper, which also occurs when increasing the length of a typical normalizing flow. Empirically, we have found that the dominant subspace U_r of H^B does not differ strongly from that of H, but that the smaller variance of the estimator $\widehat{H}^B$ can yield more reliable results in early iterations. Methods to reduce the variance of an unbiased estimator of H, such as control variates, may certainly be useful in other problems, and we will provide more guidance on this in the final paper. We will also discuss a criterion for switching to the importance sampling estimator of H for deeper layers. As the intermediate targets become closer to Gaussian, the variance of this unbiased estimator naturally reduces. Currently, however, we haven't found problems where the basis derived from H^B has been ineffective.

Other points. In the discussion after Prop. 1, we should have $\mathcal{L}_y(x) \propto f(U^T x)$, i.e. f is implicitly scaled to enforce normalization. The phrase "lack of precision" in §4.4 refers to the finite number of samples drawn from ρ used to resolve the objective and diagnostics. We believe the jump in the diagnostic at $\ell = 9$ is the result of sampling noise when computing H_9, as the diagnostic drops back after one additional lazy step. We agree that the ordering of our map composition is the reverse of that in several methods. This is a result of defining the residual targets π_ℓ as pullbacks, and allows all maps trained to use the Gaussian base distribution. Outside of this work, we are considering the effects of using the forward KL in the error bound and the backward KL in training, though we observe benefits in both directions. We agree with several reviewers that the addition of a dedicated prior work section will improve the organization of the paper, and we will include additional derivations of key results in the appendices to make things self-contained.