We thank the reviewers for their invested time and constructive criticism. We believe that their suggestions will significantly improve the manuscript. In the following, the comments are addressed separately for each reviewer.

Reviewer-1

i) "Should the optimisation problem: $\min\, \mathbb{KL}(q\|\, p)$ s.t. $\mathbb{E}[C(\mathbf{x}, \mathbf{z})] < \kappa^2$":
Yes, indeed our optimisation objective starts from $\min\, \mathbb{KL}(q\|\, p)$ s.t. $\mathbb{E}[C(\mathbf{x}, \mathbf{z})] < \kappa^2$ (line 58). Eq. (6) defines the corresponding Lagrange dual problem. To extend this to a two level stochastic model, we additionally use an IW upper bound on $\log p(\mathbf{z})$ inside the KL—Eq. (8)—to accommodate the hierarchical representation of $\log p(\mathbf{z})$. This leads to the optimisation problem in Eq. (10). We will emphasise this in the manuscript.

ii) "In what way is the interpolations done with VHP+REWO better than...? Is there a way to quantify these results?":
We chose to quantify the graph-based interpolations through the smoothness of the interpolated trajectories in the data space, as it is one of the desired properties of informative latent representations (Bengio et al.; arXiv:1206.5538). For this purpose, we introduced a smoothness factor (line 223). When we compare VHP+REWO to the VampPrior and the standard normal prior, we obtain smoother trajectories (Fig. 7). We will try to expand this part of the paper.

iii) "It would be good also to see whether the Lagrangian update alone already leads to a good performance, or...":
Our preliminary results showed that the update alone does not guarantee a good performance as it still leads to over-regularisation due to the standard normal prior. Hence, we obtain unrealistic interpolations similar to Fig. 4 (bottom). It is the combination of both the Lagrange update (REWO) and the VHP that leads to good performance as shown in Tab. 1. We will point that out more clearly.

iv) "In terms of reporting the results it would be better to do multiple runs and report LL mean + standard error":
We agree on that and we are trying to close this gap. We have been somewhat limited by the number of GPUs we have access to (depending on the dataset, one optimisation takes over a week on a single GPU until it converges).

Reviewer-2

i) "Just one point is to show whether equation 9 which is objective function of our optimization is...":
This is a valid point, we will add "$\log p(\mathcal{D}) \geq \mathcal{L}_{\text{VHP}}(\theta, \phi, \Theta, \Phi; \lambda)$ if $\lambda \geq 1$" before we introduce REWO (line 102).

ii) "My only suggestion is that authors can take a look at the paper Molchanov, Dmitry, et al...":
We thank the reviewer for pointing us to this paper. Indeed, the authors use a similar two-level stochastic model with a combination of implicit and explicit distributions for the encoder and decoder. Inference is done through optimising a sandwich bound of the ELBO, which is specific to the choice of implicit distributions. In our work, however, we address inference using a constrained optimisation approach and our distributions are all explicit. We will definitely cite and discuss Molchanov, Dmitry, et al. (2018) in the related work.

Reviewer-3

i) "I found the paper quite clear though the argumentation is sometimes a bit too informal (see for example, lines 123...":
We thank the reviewer for pointing us to this paragraph. It can indeed be improved in terms of clarity—we will emphasise that the intuitions in this paragraph are mostly based on empirical evidence.

ii) "...though it would have been nice to see an ablation for REWO itself using priors of different complexity.":
That is an interesting question and we have run some selective experiments, where we combined REWO with the VampPrior and the standard normal prior. Generally, we observed that: i) REWO alone makes the optimisation less sensitive to hyperparameters like the network architecture, and ii) it guarantees that the reconstruction is not neglected in favour of a low KL. However, we decided that experiments in these directions would take the focus from our main message: learning informative latent representations. On the other hand, if we want to judge REWO only based on the quality of the latent representation, it is beneficial to use an arbitrary flexible prior (experiments in Tab. 1).

iii) "Similarly, to which extent the modification of the update rule for lambda contributes to results?":
We compared GECO to REWO (modified update rule) on our two-level stochastic model (Sec. 4.1 & 4.3). Apart from obtaining better ELBO values (Tab. 2) at the end of training, REWO led to more informative latent representations, as shown in the graph-based interpolations (Fig. 4) and the OLS regression (Tab. 1).

iv) "In Related work you discuss connections with VampPrior which uses the same inference network q(z|x)...":
Thank you for pointing that out, we will emphasise the need of an additional $q(\zeta \,|\, \mathbf{z})$ in comparison to the VampPrior.

v) "The text also says (line 158–159) "the aggregated posterior is...". I think a better wording would...":
Yes, this is indeed a more accurate description, we will reword the sentence as suggested—and also replace "aggregated posterior" by "prior" in the previous sentence (line 156).