Rebuttal for "Invertible Gaussian Reparameterization" (ID 10306): We thank the reviewers for their time and valuable feedback. We will incorporate all minor points in the manuscript, and answer the major ones below:

• A few key misconceptions we want to strongly rebut (R1):

− We disagree that our empirical improvements are not significant. For example, REBAR obtains a 1.65% relative improvement over the GS, and RELAX 0.22% over REBAR; while we obtain 6.2% over the GS.

− There is no error in eq.13. While q is not Gaussian, the $\mathbb{KL}$ between q's is equivalent to a $\mathbb{KL}$ between Gaussians. The $\mathbb{KL}$ is invariant to invertible transformations. We introduced the softmax$_{++}$ precisely to use this fact.

− We note that the GS does not transform α into an approximate one-hot sample as claimed by R1. Instead, it transforms Gumbel noise into the sample, and the transformation depends on α. Similarly, IGR transforms Gaussian noise into the sample, and the transformation depends on (μ, σ).

• Tasks being too simple (R1 and R4): The tasks we use are the de-facto standard for benchmarking continuous relaxations of discrete distributions in the literature, see [8, 9, 10, 16, 22, 25] from the main manuscript and [3] from the appendix to reference a few papers whose experiments we are consistent with. Note also that while the VAE problems we consider involve (albeit partially factorizable) 10^{20}-dimensional discrete distributions, the main goal of our experiments is to compare IGR and GS, not to obtain state-of-the-art VAE performance. That being said, we will include experiments on CIFAR-10 with more complex architectures by publication time.

• Reporting results on test set (R3): Thank you, this is an important point to discuss. From an ML standpoint, we believe test scores on discretized models are the most relevant metric. An ML practitioner wanting to train a discrete VAE (or another ML model) will care about the recovered discrete model's performance; regardless of the optimization objective's value. We agree that training losses (on the discrete objective) are also interesting to look into though, and so we have included them in the table below. Note that IGR also outperforms the GS here.

• On flexibility (R3): To show that at least to a certain degree our empirical improvements do come from added flexibility, we re-ran our experiments while fixing $\sigma = 1$ and only learning μ, resulting in the same number of degrees of freedom as the GS. While this version of IGR still outperformed the GS, it was by a much smaller margin: test log likelihoods were -102.81 and -43.83 for MNIST and FMNIST, respectively, and -140.37 for Omniglot, where the GS slightly outperformed (compare with Table 1 from the main manuscript).

• Unclear storyline (R3): We believe the conceptual simplicity of not having to involve the arguably exotic and not extensible Gumbel distribution, along with more straightforward density derivation, closed-form $\mathbb{KL}$, and strong empirical results do form a coherent storyline. We will convey this in a clearer way in the manuscript.

• Straight-through estimator (R2): This is a great point. Proposition 1 enables discretizing the sample, so that we can carry out a straight-though gradient estimate. We will also include an experiment with this estimator by publication.

• Using statically binarized MNIST (R3): We are actually using this version of MNIST. The odd numbers in the appendix table are a consequence of numerical instabilities for the quadrature we used to estimate the integral in proposition 2, which we have corrected since submission. This issue only affects the IGR-I+RELAX entries of tables 2 and 3 in the appendix, which should be -81.9481 and -102.214, respectively. Note also that IGR still outperforms.

• On VIMCO and VQ-VAE (R4): Thank you for mentioning these papers, we will reference them. However, we do point out that VIMCO is outperformed by USPGBL ([3] from the appendix), against which we do compare and outperform in the appendix; and that VQ-VAE does not endow the latent variables with a "true" distribution as they use point masses and is different to IGR and the GS in this regard.

• Stick-breaking and orderings (R2): Note that IGR-SB need not place more mass around the first vertices of the simplex, as by choosing (μ, σ) appropriately mass can be shifted towards the right (e.g our Poisson experiments).

• R1: Equation 14 is valid, while p_0 is a distribution, we are thinking of it as a probability vector. We will further clarify this in the manuscript.

• R1: We also point out that we did mention that IGR and the GS have the same running times (lines 253-255).

• R2: Note that $\bar{\alpha}$ is a probability vector, not a scalar; and is thus treated as a distribution in eq. 8.

Model	**MNIST**	**FMNIST**	**Omniglot**
IGR-I	-131.86	-66.74	**-143.09**
IGR-Planar	**-126.44**	**-53.65**	-161.78
IGR-SB	-130.99	-71.87	-154.23
GS	-147.82	-85.74	-160.46
IGR-I + SL	-128.79	-65.77	-145.66
IGR-Planar + SL	**-126.22**	-66.39	**-139.56**
IGR-SB + SL	-128.08	**-65.21**	-157.11
GS + SL	-147.60	-83.24	-155.17

Table 1: Discretized Train ELBO (not log-likelihood) on MNIST, FMNIST and Omniglot for IGR and GS. Higher is better.