Reviewer #1: We thank you for appreciating our contributions and providing valuable feedback, which will be taken into account when revising our paper, such as shortening Sect. 3 to free up space for including more details in Sect. 4.

The empirical results comparing parameter tying vs. naive design are in fact reported in Table 3 of Appendix C.2; a reader is referred to these results in lines 161 to 163 of the main paper. Based on your suggestion, we will move such a comparison from the appendix to the main paper.

Regarding how ϵ is shared among all the layers, observe from the left-hand side of Fig. 2d as well as from line 156 that the same ϵ is fed as an input to g_{ϕ_ℓ} in each layer ℓ for $\ell = 1, \ldots, L$.

Reviewer #4: We thank you for providing insightful comments and advice, which will be incorporated into our revised paper. We will include a further discussion on how the other works of implicit VI (Titsias and Ruiz, 2019; Yin and Zhou, 2018) are related to IPVI, as you have suggested.

Reviewer #5: We thank you for providing valuable suggestions and feedback, which we will consider seriously in revising our paper. We would like to address your comments and questions below.

Regarding the necessity of parameter tying, we think overfitting is still an issue to be addressed. To see this, note that the generator's parameters (i.e., variational parameters) are trained to come close to the true posterior that, importantly, is conditioned on only the *training* data and not the entire data distribution (which encompasses the training data as well as the test data). Hence, variational inference (VI) is still subject to overfitting, even though it is regularized by the KL term. This motivates our use of parameter tying. We provide some experimental evidence below, as you have suggested.

Table I reports the train/test mean log-likelihood (MLL) achieved by IPVI with and without parameter tying for 2 small datasets: Boston ($N = 506$) and Energy ($N = 768$). For Boston dataset, it can be observed that no tying consistently yields higher train MLL and lower test MLL, hence indicating overfitting. This is also observed for Energy dataset when the no. of layers exceeds 2. For both datasets, as the no. of layers (hence no. of parameters) increases, overfitting becomes more severe for no tying. In contrast, parameter tying alleviates the overfitting issue considerably. There is no issue of overfitting in the discriminator, as you have correctly pointed out; we will clarify this in our revised paper.

Table I. Train/test mean log-likelihood (MLL) achieved by IPVI with and without parameter tying over 10 runs.

Dataset	Boston ($N = 506$)				
DGP Layers	1	2	3	4	5
No Tying	-1.86/-2.21	-1.76/-2.37	-1.64/-2.48	-1.52/-2.51	-1.51/-2.57
Tying	-1.91/-2.09	-1.79/-2.08	-1.77/-2.13	-1.84/-2.09	-1.83/-2.10
Dataset	Energy ($N = 768$)				
DGP Layers	1	2	3	4	5
No Tying	-0.12/-0.44	0.03/-0.31	0.18/-0.34	0.20/-0.47	0.21/-0.58
Tying	-0.16/-0.32	-0.11/-0.34	-0.02/-0.23	0.10/-0.01	0.17/0.13

About the performance gap of SGPs, the optimal variational posterior is indeed a Gaussian for single-layer SGP regression (Titsias, 2009). However, since the SGP model hyperparameters are not known beforehand, DSVI SGP has to *jointly* optimize its hyperparameters and variational parameters. Such an optimization is not convex. Hence, there is no guarantee that it will reach the global optimum. Thus, the performance gap can be explained by IPVI's ability to jointly find "better" values of hyperparameters and variational parameters. Additionally, in our experiments, the performance of DSVI has already converged within 20000 iterations and is thus not limited.

We have also computed the estimate of ELBO (by, after training, continuing to train the discriminator for longer), like you have suggested. Table II shows the mean ELBOs of DSVI and IPVI over 10 runs for the Boston dataset. IPVI generally achieves higher ELBOs, which agrees with results of the test log-likelihood in Fig. 4 of the main paper. Since SGHMC DGP is not based on VI, no ELBO is computed for that method; a comparison of its expressive power and performance with that of IPVI is already investigated and analyzed in the synthetic experiment in Section 5.1.

Based on your comments for the MNIST experiment, we increase the number of inducing points to 800. Table III reports the mean test accuracy over 5 runs. The results are consistent with that in Table 2 of the main paper where IPVI DGP 4 performs the best. Regarding the code, we will cite GPflow in our revised paper.

Table II. Mean ELBOs for Boston dataset.

Model	DSVI	IPVI
SGP	-956.57	-934.07
DGP 2	-850.54	-846.65
DGP 3	-836.13	-846.45
DGP 4	-787.10	-776.93
DGP 5	-770.67	-758.42

Table III. Mean test accuracy (%) on MNIST datasets.

Dataset	MNIST	
	SGP	DGP 4
DSVI	**97.92**	98.05
SGHMC	97.07	97.91
IPVI	97.85	**98.23**