Table A: FID on CIFAR10. † means averaged by 5 runs. Methods with ‡ use comparable networks.

Method	FID ↓	FID-ES ↓
Flow-CE [1*]	37.30	-
VAE-EBLVM[2*]	**30.1**	-
MDSM‡ [34]	-	31.7
MDSM‡ (our code)	39.12	$30.19 \pm 2.60^\dagger$
BiMDSM‡ (20)	**34.55**	$\mathbf{26.62} \pm 1.52^\dagger$
BiMDSM‡ (50)	38.82	$29.43 \pm 2.76^\dagger$
BiMDSM‡ (100)	-	$\mathbf{26.90} \pm 2.14^\dagger$

Table B: Test log-likelihood (LL) results on Frey face. We use 2,000 chain samples in AIS.

Method	LL ↑
DSM	129.23
BiDSM (N=0)	107.59
BiDSM (N=1)	110.65
BiDSM (N=5)	124.00
BiDSM (N=10)	125.72

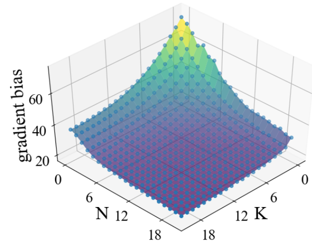

Figure A: The gradient bias (i.e., the left hand side of Thm. 2) w.r.t. N and K in GRBM on Frey face.

We thank all reviewers for their valuable comments. We **update the FID results in Tab. A**, **validate Thm.2 in Fig. A**, **add AIS results in Tab. B**, **add three baselines (VNCE [42], [1*] and [2*]**), and clarify other issues (e.g., computation time). Below, we first address the common concerns and then answer the detailed questions.

Common concern (CC) 1 from R#2&R#4 **Validate Thm. 2:** In Fig. A, the gradient bias decays (approximately) exponentially w.r.t. N, as proved in Thm. 2. Besides, we find that as K increases from 0 to 20, $||\phi^0 - \hat{\phi}^*||$ decreases from 1.38 to 0.87. It leads to smaller bias (see Fig. A), which also agrees with Thm. 2. **CC2 from R#2&R#4&R#5** **Benefits of EBLVM:** First, introducing latent variables can improve the sample quality (w.r.t. FID) in a fair comparison. Indeed, we update Tab. 2 and obtain Tab. A, whose second column shows the FID with early stopping (ES) according to the results on 1,000 samples. As stated in L290, a similar protocol is adopted in MDSM [34]. We also implement MDSM in our code for a fair comparison. The reproduced MDSM is slightly better than the original paper [34] and serves as a stronger baseline. Our result outperforms MDSM and [1*]. We mention that [2*] generate samples from a VAE instead of an EBM and is less comparable. Second, the deep EBLVM is not suitable for conditional generation because $p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta})$ is multimodal. The results (Fig.4) and analysis are shown in Appendix C.2.2. We expect that the model can serve as a benchmark and inspire new model design. **CC3 from R#3&R#4** **Computation time:** The time complexity per gradient estimate is $O(N{+}K)$. However, empirically, we don't need arbitrarily large N and K. Indeed, in the default setting, the training time per 100 iterations is 8.61s for BiDSM, 1.59s for CD-5, 4.36s for VNCE, 1.33s for DSM in GRBM on Frey face. The training time of 300,000 iterations is 48h for BiMDSM and 32h for MDSM in deep EBLVM on CIFAR10 (see L118 in Appendix B.2). Thus, BiSM can learn general EBLVMs without a prohibitive cost.

To R#2. **Typos:** We'll correct it in the final version. **Strongly convex assumption in Thm. 2:** Currently the assumption is necessary. **Dependence on the batch size:** An infinite batch size is not necessary. Actually, the constants (A, B, C and κ) and the learning rate α can be made independent of the batch size by applying assumptions 2 and 3 in Thm. 2 to $\mathbb{E}_{q(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\phi})}\mathcal{F}(\cdots)$ (Eqn. (8)) and $\mathcal{D}(\cdots)$ (Eqn. (9)) instead of $\hat{\mathcal{J}}_{Bi}$ and $\hat{\mathcal{G}}$. **Update $\boldsymbol{\phi}$ for K times on the same minibatch:** It is a special design. According to Thm. 2, we should minimize $||\phi^0 - \hat{\phi}^*||$, where $\hat{\phi}^*$ is optimal on a given minibatch. We update $\boldsymbol{\phi}$ multiple times on the same minibatch to obtain $\boldsymbol{\phi}^0$ that approximates $\hat{\boldsymbol{\phi}}^*$. We'll make it clearer. **Validating Thm. 2 and ablation study:** See the common concern 1. We'll add the ablation study of K. **Practical usefulness:** See the common concern 2. **CelebA 128x128:** We obtain promising generation results on CelebA 64x64 and are working on 128x128 data. We'll include the results. **Noise annealing on the images:** It is necessary. Indeed, MDSM uses the annealed noise in its objective (Eqn. (5)). **Recent work:** Thanks. We'll discuss it.

To R#3. **Compare to [1*]:** Thanks, we compare to [1*] in Tab. A. **Higher dimension of $\boldsymbol{h}$ is worse:** We add a new experiment with $\boldsymbol{h}$ dimension ($d_{\boldsymbol{h}}$) of 100 in Tab. A, which is comparable to $d_{\boldsymbol{h}} = 20$. The relatively worse results of $d_{\boldsymbol{h}} = 50$ may be caused by the variance of training on different initial seeds. **Unstable learning:** The lower level optimization can be slightly unstable because the distribution of EBLVM is moving during training. The higher level optimization is stable. We'll plot it in the final version.

To R#4. **Compare to VNCE [42]:** We implement VNCE. On the toy data, its log likelihood is 0.303 nats, which is worse than 0.319 nats of BiDSM. We'll add the curve of VNCE to Fig. 2 in the final version. **Missing reference:** Thanks. We'll discuss this work in the final version. **Motivation of deep EBLVM:** See the common concern 2. **Likelihood estimate:** See Tab. B. BiDSM gets closed to DSM as N increases and $N \geq 5$ is sufficient. **Trades-offs and future work:** Taking less than twice computation time of the regular SM (see common concern 3), BiSM can learn deep EBLVMs. We'll discuss the future work in the final version. **Inference model:** Thanks. The inference model is similar to the one used in VAE, as described in Appendix B.1 and B.2 (also see the code in the anonymous link in Page 6). We will add more details in the main text. **Correctness:** See common concern 1. **Clarity:** The algorithm is consistent to what you believe and we'll improve the clarity.

To R#5. **Experimental results and compare to [2*]:** We use the widely adopted FID (Tab. A) metric for evaluation and compare to strong baselines [34][1*][2*]. Our updated results outperform baselines with comparable architectures (see common concern 2). **Missing references:** We will include the missing references mentioned in the comments.

[1*] Flow contrastive estimation of EBMs. [2*] Joint Training of Variational Auto-Encoder and Latent EBM.