A. Bi-level optimization

A.1. Metrics for lower-level subproblem

IPMs. We start from the gradient of standard objective for EBLVM *i.e.* Eq. (3):

$$\omega^*(\psi) = \arg\min_{\omega\in\Omega} \left\|\mathbb{E}_{p(v)p_\psi(h|v)}\left[\nabla_\psi \mathcal{E}_\psi\right] - \mathbb{E}_{p(v)q_{\omega_1}(z|a)}\left[\nabla_\psi \mathcal{E}_\psi\right]\right\|_\infty \quad (16)$$

$$+ \left\|\mathbb{E}_{p_{\omega_2}(v,h)}\left[\nabla_\psi \mathcal{E}_\psi\right] - \mathbb{E}_{p_\psi(v,h)}\left[\nabla_\psi \mathcal{E}_\psi\right]\right\|_\infty. \quad (17)$$

Then we bound two infinite norms from above by general integral probability metrics (IPMs), we only show derivation about term (17) and the derivation about term (16) can be obtained by the same way:

$$\begin{aligned}
&\left\|\mathbb{E}_{p_{\omega_2}(v,h)}\left[\nabla_\psi \mathcal{E}_\psi\right] - \mathbb{E}_{p_\psi(v,h)}\left[\nabla_\psi \mathcal{E}_\psi\right]\right\|_\infty \\
\leq &\sup_{f:\mathcal{V}\times\mathcal{H}\to\mathbb{R}, f\in\mathcal{F}} \left|\mathbb{E}_{p_{\omega_2}(v,h)}\left[f(v,h)\right] - \mathbb{E}_{p_\psi(v,h)}\left[f(v,h)\right]\right| \\
= &D_{\mathcal{F}}(p_{\omega_2}(v,h), p_\psi(v,h)),
\end{aligned} \quad (18)$$

where $\mathcal{F}$ is a class of scalar function over space $\mathcal{V}\times\mathcal{H}$, $D_{\mathcal{F}}$ denotes the IPM induced by $\mathcal{F}$. Notice $\mathcal{F}$ depends on the gradient of energy function $\nabla_\psi \mathcal{E}_\psi(v,h)$, we thus propose some assumptions about $\nabla_\psi \mathcal{E}_\psi(v,h)$ for special cases:

(1) Assume the infinite norm of each component, *i.e.* $\|\nabla_\psi \mathcal{E}_\psi(v,h)_i\|_\infty$, is bounded by a constant C, then

$$(17) \leq C \cdot D_{\mathrm{TV}}(p_{\omega_2}(v,h), p_\psi(v,h)),$$

where D_{TV} denotes the total variation distance corresponding to $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$.

(2) Assume each component $\nabla_\psi \mathcal{E}_\psi(v,h)_i$ is C-Lipschitz, then

$$(17) \leq C \cdot W_1(p_{\omega_2}(v,h), p_\psi(v,h)),$$

where W_1 denotes the 1-Wasserstein distance corresponding to $\mathcal{F} = \{f : f \text{ is 1-Lipschitz}\}$.

(3) Assume the norm of each component defined in reproducing kernel Hilbert space $\mathcal{H}$, i.e. $\|\nabla_\psi \mathcal{E}_\psi(v,h)_i\|_{\mathcal{H}}$, is bounded by a constant C, then

$$(17) \leq C \cdot \mathrm{MMD}(p_{\omega_2}(v,h), p_\psi(v,h)),$$

where MMD denotes the maximum mean discrepancy corresponding to $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$.

Practical choices. Assumption (2) is hard to verify because both $p_\psi(h|v)$ and $p_\psi(v,h)$ are intractable in our setting, so that we can not directly use the metric for optimization. In practice, we resort to moderate metrics under mild assumptions:

1. Notice that assumption (1) is typically mild in practice and we can consider the generally adopted KL divergence by Pinsker's inequality

$$2D_{\mathrm{TV}}(p_{w_2}, p_\psi)^2 \leq D_{\mathrm{KL}}(p_{w_2} \| p_\psi),$$

where the equality holds only if p_{w_2} equals p_ψ. KL divergence has a stronger convergence than many other divergence metrics and we show that it is feasible in our work.

2. Under assumption (3), $\mathrm{MMD}^2(p,q) = \mathbb{E}[k(x,x') - 2k(x,y) + k(y,y')]$, where x, x' and y, y' are i.i.d. drew from p and q, respectively. However, it is impossible to obtain its gradient w.r.t. w by Monte Carlo estimation. The kernelized Stein discrepancy (KSD) is a special MMD with a kernel $u_q(x,x')$ depending on q:

$$\begin{aligned}
\mathrm{KSD}(p,q) &= \mathbb{E}_{x,x'\sim p}[u_q(x,x')] \\
u_q(x,x') &= \mathbf{s}_q(x)^\top k(x,x')\mathbf{s}_q(x') + \mathbf{s}_q(x)^\top \nabla_{x'} k(x,x') \\
&+ \nabla_x k(x,x')^\top \mathbf{s}_q(x') + \mathrm{trace}(\nabla_{x,x'} k(x,x')),
\end{aligned}$$

where $\mathbf{s}_q(x) = \nabla_x \log q(x)$ is known as the score function. Score function is a gradient w.r.t. x, so it eliminates the intractable partition function which is independent of variable x. We also refer to Fisher divergence as a moderate metric in our setting which is further stronger than KL, total variation and KSD. Using Fisher divergence in fact corresponds to score matching widely used in learning generative models.

A.2. Derivations

In this subsection, we provide some supplemental derivations. We start from the Eq. (5) with focus on the second part in Eq. (5) and the first part can be derived in a same way:

$$\begin{aligned}
&\frac{\partial D_{\mathrm{KL}}(p_{\omega_2(\psi)}(v,h) \| p_\psi(v,h))}{\partial \psi} \\
= &\frac{\partial D_{\mathrm{KL}}(p_{\omega_2}(v,h) \| p_\psi(v,h))}{\partial \psi}|_{\omega_2=\omega_2(\psi)} \\
+ &\left(\frac{\partial \omega_2(\psi)}{\partial \psi^\top}\right)^\top \frac{\partial D_{\mathrm{KL}}(p_{\omega_2}(v,h) \| p_\psi(v,h))}{\partial \omega_2}|_{\omega_2=\omega_2(\psi)},
\end{aligned}$$

where $\frac{\partial \omega_2(\psi)}{\partial \psi}$ is the Jacobian. We then look into its first term as follow where we use ∇_ψ for clarity:

$$\begin{aligned}
&\nabla_\psi D_{\mathrm{KL}}(p_{\omega_2}(v,h) \| p_\psi(v,h))|_{\omega_2=\omega_2(\psi)} \\
= &\nabla_\psi \left[\int p_{\omega_2}(v,h) \log \frac{p_{\omega_2}(v,h)}{p_\psi(v,h)} \mathrm{d}h\mathrm{d}v\right]_{\omega_2=\omega_2(\psi)} \\
= &\nabla_\psi \left[-\int p_{\omega_2}(v,h) \log \frac{\exp(-\mathcal{E}_\psi)}{\mathcal{Z}(\psi)} \mathrm{d}h\mathrm{d}v\right]_{\omega_2=\omega_2(\psi)} \\
= &\mathbb{E}_{p_{\omega_2(\psi)}(v,h)}\left[\nabla_\psi \mathcal{E}_\psi\right] + \nabla_\psi \log \mathcal{Z}(\psi),
\end{aligned}$$

where the partition function $\mathcal{Z}(\psi)$ is independent of variables v and h. At last, we obtain Eq. (5) by

$$\begin{aligned}
&\frac{\partial D_{\mathrm{KL}}(q(v)q_{\omega_1}(h|v)\|p_\psi(v,h))}{\partial\psi}|_{\omega_1=\omega_1(\psi)} \\
&-\frac{\partial D_{\mathrm{KL}}(p_{\omega_2}(v,h)\|p_\psi(v,h))}{\partial\psi}|_{\omega_2=\omega_2(\psi)} \\
&=\mathbb{E}_{q(v)q_{\omega_1(\psi)}(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]+\nabla_\psi\log\mathcal{Z}(\psi) \\
&-\mathbb{E}_{p_{\omega_2(\psi)}(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\nabla_\psi\log\mathcal{Z}(\psi) \\
&=\mathbb{E}_{q(v)q_{\omega_1(\psi)}(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{p_{\omega_2(\psi)}(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right].
\end{aligned}$$

Furthermore, recall Eq. (7), we have:

$$\begin{aligned}
&\frac{\partial\mathcal{J}_{\mathrm{UL}}(\psi,\omega)}{\partial\psi}|_{\omega=\omega^*(\psi)}= \\
&\mathbb{E}_{q(v)q_{\omega_1^*(\psi)}(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{p_{\omega_2^*(\psi)}(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right]
\end{aligned}\tag{19}$$

$$\begin{aligned}
&\left(\frac{\partial\omega^*(\psi)}{\partial\psi^\top}\right)^\top\frac{\partial\mathcal{J}_{\mathrm{UL}}(\psi,\omega)}{\partial\omega}|_{\omega=\omega^*(\psi)}= \\
&\left(\frac{\partial\omega_1^*(\psi)}{\partial\psi^\top}\right)^\top\frac{\partial D_{\mathrm{KL}}(q(v)q_{\omega_1}(h|v)\|p_\psi(v,h))}{\partial\omega_1}|_{\omega_1=\omega_1^*(\psi)} \\
&-\left(\frac{\partial\omega_2^*(\psi)}{\partial\psi^\top}\right)^\top\frac{\partial D_{\mathrm{KL}}(p_{\omega_2}(v,h)\|p_\psi(v,h))}{\partial\omega_2}|_{\omega_2=\omega_2^*(\psi)},
\end{aligned}\tag{20}$$

A.3. Proof and properties

We next proof the equivalence of BLO problem (4,6) and the original one, under the unparametric assumption.

Proof of Theorem 1. Suppose for $\psi\in\Psi$ we have $\hat{\omega}\in\Omega$ such that $D(q(v)q_{\hat{\omega}_1}(h|v),q(v)p_\psi(h|v))=0$ and $D(p_{\hat{\omega}_2}(v,h),p_\psi(v,h))=0$, then $0\le\mathcal{J}_{\mathrm{LL}}(\psi,\omega^*(\psi))\le\mathcal{J}_{\mathrm{LL}}(\psi,\hat{\omega})=0$, thus $\mathcal{J}_{\mathrm{LL}}(\psi,\omega^*(\psi))=0$. In other words, $\omega^*(\psi)$ satisfies $q_{\omega_1^*(\psi)}(h|v)=p_\psi(h|v),p_{\omega_2^*(\psi)}(v,h)=p_\psi(v,h)$. Then we have

$$\begin{aligned}
&|\mathcal{J}_{\mathrm{UL}}(\psi,\omega^*(\psi))-\mathcal{J}(\psi)| \\
&=|D_{\mathrm{KL}}(q(v)q_{\omega_1^*(\psi)}(h|v)\|p_\psi(v,h)) \\
&-D_{\mathrm{KL}}(p_{\omega_2^*(\psi)}(v,h)\|p_\psi(v,h))-D_{\mathrm{KL}}(q(v)\|p_\psi(v))| \\
&=\left|\mathbb{E}_{q(v)q_{\omega_1^*(\psi)}(h|v)}\left[\log\frac{q(v)q_{\omega_1^*(\psi)}(h|v)}{p_\psi(v)p_\psi(h|v)}\right]\right. \\
&\left.-D_{\mathrm{KL}}(p_{\omega_2^*(\psi)}(v,h)\|p_\psi(v,h))-\mathbb{E}_{q(v)}\left[\log\frac{q(v)}{p_\psi(v)}\right]\right| \\
&=\left|\mathbb{E}_{q(v)q_{\omega_1^*(\psi)}(h|v)}\left[\log\frac{q_{\omega_1^*(\psi)}(h|v)}{p_\psi(h|v)}\right]\right. \\
&\left.-D_{\mathrm{KL}}(p_{\omega_2^*(\psi)}(v,h)\|p_\psi(v,h))\right| \\
&=|D_{\mathrm{KL}}(q(v)q_{\omega_1^*(\psi)}(h|v)\|q(v)p_\psi(h|v)) \\
&-D_{\mathrm{KL}}(p_{\omega_2^*(\psi)}(v,h)\|p_\psi(v,h))| \\
&\le D_{\mathrm{KL}}(q(v)q_{\omega_1^*(\psi)}(h|v)\|q(v)p_\psi(h|v)) \\
&+D_{\mathrm{KL}}(p_{\omega_2^*(\psi)}(v,h)\|p_\psi(v,h))=0,
\end{aligned}\tag{21}$$

where the last equation holds because $D_{\mathrm{KL}}(P\|Q)=0$ is equivalent to $P=Q$. On the other hand, by (3,6,19,20), we have

$$\begin{aligned}
&\|\nabla_\psi\mathcal{J}(\psi)-\nabla_\psi\mathcal{J}_{\mathrm{UL}}(\psi,\omega^*(\psi))\|_\infty \\
&=\left\|\mathbb{E}_{q(v)p_\psi(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{q(v)q_{\omega_1^*(\psi)}(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]\right. \\
&+\mathbb{E}_{p_{\omega_2^*(\psi)}(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{p_\psi(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right] \\
&\left.-\left(\frac{\partial\omega^*(\psi)}{\partial\psi^\top}\right)^\top\frac{\partial\mathcal{J}_{\mathrm{UL}}(\psi,\omega)}{\partial\omega}|_{\omega=\omega^*(\psi)}\right\|_\infty \\
&\le\|\mathbb{E}_{q(v)p_\psi(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{q(v)q_{\omega_1^*(\psi)}(h|v)}\left[\nabla_\psi\mathcal{E}_\psi\right]\|_\infty \\
&+\|\mathbb{E}_{p_{\omega_2^*(\psi)}(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right]-\mathbb{E}_{p_\psi(v,h)}\left[\nabla_\psi\mathcal{E}_\psi\right]\|_\infty \\
&+\left\|\left(\frac{\partial\omega^*(\psi)}{\partial\psi^\top}\right)^\top\frac{\partial\mathcal{J}_{\mathrm{UL}}(\psi,\omega)}{\partial\omega}|_{\omega=\omega^*(\psi)}\right\|_\infty.
\end{aligned}\tag{22}$$

Because $\omega^*(\psi)$ satisfies $q_{\omega_1^*(\psi)}(h|v)=p_\psi(h|v)$ and $p_{\omega_2^*(\psi)}(v,h)=p_\psi(v,h)$, under nonparametric assumption. Thus we know that $\omega_1=\omega_1^*(\psi),\omega_2=\omega_2^*(\psi)$ are the stationary points of $\min_{\omega_1}D_{\mathrm{KL}}(q(v)q_{\omega_1}(h|v)\|q(v)p_\psi(h|v))$ and $\min_{\omega_2}D_{\mathrm{KL}}(p_{\omega_2}(v,h)\|p_\psi(v,h))$, respectively. Due to

$$\begin{aligned}
&\nabla_{\omega_1}D_{\mathrm{KL}}(q(v)q_{\omega_1}(h|v)\|q(v)p_\psi(h|v)) \\
&=\nabla_{\omega_1}D_{\mathrm{KL}}(q(v)q_{\omega_1}(h|v)\|p_\psi(v,h)),
\end{aligned}$$

we have $\frac{\partial\mathcal{J}_{\mathrm{UL}}(\psi,\omega)}{\partial\omega}|_{\omega=\omega^*(\psi)}=0$. Consequently we bound Eq. (22) by

$$\begin{aligned}
&\le C\cdot D(q(v)q_{\omega_1^*(\psi)}(h|v),q(v)p_\psi(h|v)) \\
&+C\cdot D(p_{\omega_2^*(\psi)}(v,h),p_\psi(v,h)) \\
&=C\cdot\mathcal{J}_{\mathrm{LL}}(\psi,\omega^*(\psi))=0,
\end{aligned}$$

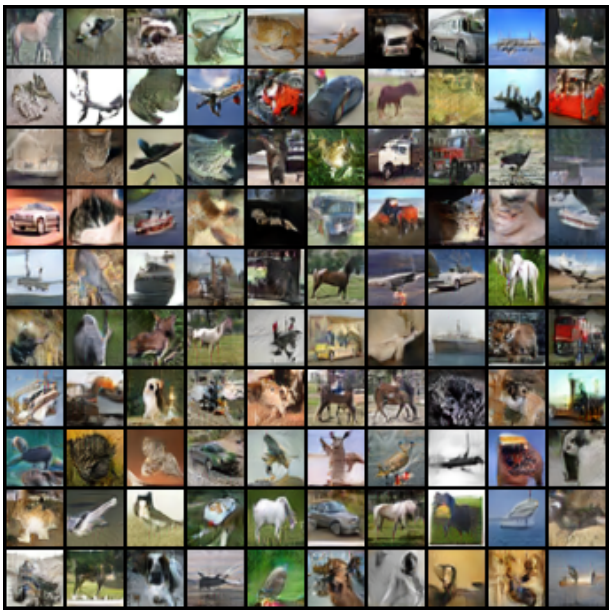

Figure 5. Randomly generated images on CIFAR-10.

which is derived from Eq. (18), C depends on the assumption about the gradient of energy function. At last we have $\nabla_\psi \mathcal{J}(\psi) = \nabla_\psi \mathcal{J}_{\mathrm{UL}}(\psi, \omega^*(\psi))$. □

In fact, the unparametric assumption typically does not hold when we take neural networks as the variational approximators, thus the optima of ω may not lie in the parameter space Ω. The proof (21) characterizes that the bias of upper-level objective $\mathcal{J}_{\mathrm{UL}}(\psi, \omega^*(\psi))$ can be bounded by KL divergences. It means that if we choose KL divergence or certain stronger metric, *e.g.* Fisher divergence, in lower-level subproblem, we can obtain bounds on both sides as $|\mathcal{J}_{\mathrm{UL}}(\psi, \omega^*(\psi)) - D_{\mathrm{KL}}(q(v)\|p_\psi(v))| \leq C' \cdot \min_{\omega\in\Omega} \mathcal{J}_{\mathrm{LL}}(\psi, \omega)$, where C' depends on the metric we choose and the assumptions about the gradient of energy function $\nabla_\psi \mathcal{E}_\psi(v, h)$. To ensure the constraint on $\nabla_\psi \mathcal{E}_\psi(v, h)$, we introduce spectral normalization into the energy network in our experiments. On the other hand, Eq. (22) indicates the upper and lower bound (black curves) in Fig. 1, and the lower-level optimization forces the gradient estimate (the gradient of upper level) to fit the real one.

B. Additional experiments

In the main paper, we demonstrate experiments on 32×32 and 64×64 images, both are small scale, because of the weak expressiveness of convolutional layers based structure. To improve the expressiveness, we borrow the Resnet-based structure of SNGAN [33] to build our marginal EBM, inference model and IGM. And then, we use proposed compact BiDVL to train models on CIFAR-10 and CelebA-128.

In this part, CIFAR-10 dataset consisting of 32×32 scale images, is to compare the generative performance of the

Figure 6. Reconstruction images on CelebA-128.

Figure 7. Randomly generated images on CelebA-128.

Resnet-based model with the simple-structured model in the main paper. While the CelebA-128 dataset constructed by resizing images in CelebA to 128×128, is for evaluating the adaptability to higher-dimensional data. They go through the same pre-processing as in main experiments.

We borrow the structure of the discriminator and the generator of SNGAN to build the marginal EBM and the IGM for both 32×32 and 128×128 datasets. But for the inference model, which is not included in SNGAN, we cascade spectral normalized Resblocks in the same way as in SNGAN discriminator, except that, the last linear layer is replaced with two linear layer to output the mean and the log-variance, respectively. Moreover, the log-variance is followed by a Softplus function to be bound within $(-\infty, 0)$, corresponding to bound the output variance within $(0, 1)$, which significantly helps the training in early stage.

For both CIFAR-10 and CelebA-128, we apply the offset compact BiDVL (13,14) and alternatively optimize the

upper-level objective with one step and the lower-level objective with one step, *i.e*. set the number of lower-level steps N to one.

Unfortunately, we found the complex structure exacerbates another learning problems resulting in instability of the marginal EBM. Since the output of the EBM is an unbounded real value, training with the Monte Carlo estimate of the upper-level gradient (13) makes the energy easily to reduces to $-\infty$. It may because adding a constant to the whole energy landscape will not change the probability distribution:

$$p(v) = \frac{\exp(-\mathcal{E}(v))}{\int \exp(-\mathcal{E}(v))\mathrm{d}v} = \frac{\exp(-\mathcal{E}(v)+c)}{\int \exp(-\mathcal{E}(v)+c)\mathrm{d}v},$$

however, largely influences the training stability. To handle the numerical problem efficiently, we turn to adopt a restricted version of Eq. (13):

$$\begin{aligned}\nabla_{\psi'}\widetilde{\mathcal{J}}_{\mathrm{UL}}(\psi') = \mathbb{E}_{q(v)}[\nabla_{\psi'}\mathrm{ReLU}(1+\mathcal{E}_{\psi'}(v))] \\ + \mathbb{E}_{p_{\omega_2}(v,h)}[\nabla_{\psi'}\mathrm{ReLU}(1-\mathcal{E}_{\psi'}(v))],\end{aligned} \tag{23}$$

which can prevent EBM from outputting numerical unstable value. Furthermore, Eq. (23) may be regarded as a strong constraint on $\nabla_{\psi'}\mathcal{E}_{\psi'}(v)$ as discussed in Appendix A.1, since it clips the objective directly.

The Resnet-based model gets 16.37 FID on CIFAR-10 and demonstrates better performance than the best model shown in Tab. 1. The generated images are presented in Fig. 5. For CelebA-128, we experimentally find that models simply cascading convolutional layers fails to generate meaningful images, but the Resnet-based model can generate good images as demonstrated in Fig. 7. Some reconstruction images are shown in Fig. 6.