Variational Inference with Unnormalized Priors

Anonymous authors
Paper under double-blind review

Abstract

Variational inference typically assumes normalized priors, which can limit the expressiveness of generative models like Variational Autoencoders (VAEs). In this work, we propose a novel approach by replacing the prior $p(z)$ with an unnormalized energy-based distribution $\exp(-E(z))/Z$, where $E(z)$ is an unrestricted energy function and Z is the partition function. This leads to a variational lower bound that allows for two key innovations: (1) the incorporation of more powerful, flexible priors into the VAE framework, resulting in improved likelihood estimates and enhanced generative performance, and (2) the ability to train VAEs with energy priors independent of the intractable normalizing constant, requiring only that the prior estimates the aggregated posterior, which can be achieved via a variety of different objectives. Our approach bridges VAEs and EBMs, providing a scalable and efficient framework for leveraging unnormalized priors in probabilistic models.

1 Introduction

Generative models are essential in unsupervised learning and data generation, with each approach offering unique strengths and facing specific challenges. Among these, Variational Autoencoders (Kingma & Welling, 2022), normalizing flows (Rezende & Mohamed, 2016; Kingma & Dhariwal, 2018), score-based/diffusion models (Song & Ermon, 2020; Sohl-Dickstein et al., 2015; Ho et al., 2020), and energy-based models (Du & Mordatch, 2020; Grathwohl et al., 2020b) represent some of the most influential methods in modern generative modeling. Each of these models brings distinct advantages but also limitations that impact their practical application and effectiveness.

Variational Autoencoders (VAEs) are regarded for their efficiency and scalability. VAEs utilize the variational lower bound (VLB) to approximate complex posterior distributions and have demonstrated considerable success in various tasks such as image generation (Vahdat & Kautz, 2021; Child, 2021) and anomaly detection (Pol et al., 2020). The primary strength of VAEs lies in their ability to efficiently model large datasets through a combination of variational inference and neural network architectures. However, VAEs face a significant challenge due to their use of simple, normalized priors, such as Gaussian distributions. This simplicity can lead to a misalignment between the prior and the posterior, where the model struggles to capture the true complexity and multi-modality of the data. Although efforts to enhance the flexibility of the posterior have been made (Rezende & Mohamed, 2016; Kingma et al., 2017), these methods do not fully resolve other issues pertaining to quality image generation (Dai & Wipf, 2019).

Normalizing flows offer an alternative by applying a series of invertible transformations to a base distribution, allowing for the modeling of complex data distributions with exact likelihood computation. This flexibility makes normalizing flows highly expressive compared to VAEs. However, the challenge lies in designing and training these transformations, which can become computationally demanding and complex, particularly as the dimensionality of the data increases. As a result, while normalizing flows provide powerful modeling capabilities, they may not always be practical for large-scale or real-time applications.

Score-based models and diffusion models (SDMs) represent another innovative approach by learning to model the score function, or the gradient of the log-likelihood, of the data distribution. These models refine noisy data through iterative denoising, leading to high-quality samples and the ability to model intricate data

structures. Despite their impressive performance, SDMs face substantial training and sampling challenges. Training involves optimizing the score function across multiple noise levels, which requires extensive computation. Additionally, the sampling process is typically slow, as generating high-quality samples often involves many iterative refinement steps. These factors can limit the practicality of SDMs for large-scale or real-time generative tasks.

Energy-based models (EBMs) offer a different paradigm by defining probability distributions through an unnormalized energy function. EBMs can capture highly complex and varied data distributions as they parameterize distributions with arbitrary functions, making them extremely flexible compared to other generative approaches. However, the practical application of EBMs is constrained by the need for computationally intensive sampling methods like Markov Chain Monte Carlo (MCMC), which are necessary to approximate the intractable partition function. This reliance on expensive sampling techniques makes EBMs less scalable and efficient compared to other generative models.

Among these approaches, Variational Autoencoders (VAEs) remain our primary focus due to their foundational role in generative modeling and their widespread application in various domains. The core limitation of VAEs lies in their posterior parameterization failing to effectively capture the complexity of the prior distribution. While enhancing the flexibility of the posterior has been explored, this does not fully capture the data distribution.

In this work, we address this limitation by introducing unnormalized energy-based priors into the VAE framework. By incorporating flexible, unnormalized priors, we aim to improve the alignment between the prior, posterior, and even the reconstruction likelihood. This novel approach leverages the expressiveness of energy-based models while maintaining the computational efficiency of VAEs. Our method provides a scalable solution that enhances generative performance and likelihood estimation, positioning unnormalized priors as a powerful tool for advancing VAE capabilities and addressing their core limitations.

2 Likelihood Estimator for Unnormalized Priors

Consider the following formulation of the variational lower bound (VLB):

$$\ln p_\theta(x) \geq \mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[\ln p_\alpha(x|z) + \ln p_\beta(z) - \ln q_\phi(z|x)] \tag{1}$$

Where $\ln p_\alpha(x|z)$ is the reconstruction likelihood, $\ln p_\beta(z)$ is the prior and $\ln q_\phi(z|x)$ is the approximate posterior. We can represent the prior $p_\beta(z)$ in terms of a Boltzmann distribution $\exp(-E_\beta(z))/Z$, where $E_\beta(z)$ is the energy function and $Z = \int \exp(-E_\beta(z))dz$ is the partition function or normalizing constant. The VLB then becomes:

$$\mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x)] - \ln Z \tag{2}$$

The main issue here pertains to the partition function as it is generally intractable to compute. When training pure energy-based models, samples from the model are required to be generated during training to approximate its gradient, which is a difficult endeavour in and of itself as it requires a high quality sampler. To make it more practical, we can exploit the approximate posterior to our advantage to estimate the partition function through self-normalized importance samples, leading to the following upper-bounded estimator of the VLB:

$$\mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x)] - \ln \left(\mathop{\mathbb{E}}_{q_\phi(z|x)}[\exp(-E_\beta(z) - \ln q_\phi(z|x))] \right) \tag{3}$$

After model training, VAEs are often also evaluated using the importance-sampled negative log-likelihood, which has a tighter bound over the VLB. We may compute this also using self-normalized importance sampling as such:

$$\ln p_\theta(x) = \ln \left(\frac{\mathbb{E}_{q_\phi(z|x)}[\exp(\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x))]}{\mathbb{E}_{q_\phi(z|x)}[\exp(-E_\beta(z) - \ln q_\phi(z|x))]} \right) \tag{4}$$

One can also train energy-based importance-weighted autoencoders (Burda et al., 2016) by directly optimizing the above bound, which in theory should result in a better generative model.

The gradient of the VLB with respect to the energy function is as such:

$$\nabla_\beta \mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[-\nabla_\beta E_\beta(z)] + \frac{\mathbb{E}_{q_\phi(z|x)}[\nabla_\beta E_\beta(z) \exp(-E_\beta(z) - \ln q_\phi(z|x))]}{\mathbb{E}_{q_\phi(z|x)}[\exp(-E_\beta(z) - \ln q_\phi(z|x))]} \tag{5}$$

The above gradient resembles the typical log likelihood gradient of an energy based model, with the first term being the positive gradient against the approximate posterior samples. The second term, while not immediately apparent, is in fact the negative gradient under the model, re-expressed as a self-normalized importance-weighted estimate by reusing the true posterior samples. As this is a ratio estimator, the gradient estimator is biased, but consistent; that is, it is asymptotically unbiased in the limit of infinite samples. More importantly, the variance of this estimator estimator decreases with increasing number of samples (Burda et al., 2016). The corresponding standard estimator is as such:

$$\nabla_\beta \mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(z|x)}[-\nabla_\beta E_\beta(z)] + \mathop{\mathbb{E}}_{p_\beta(z)}[\nabla_\beta E_\beta(z)] \tag{6}$$

Compared to the traditional gradient estimators, the second term in Equation (5) does not require generating samples from the energy model through MCMC sampling, which can be expensive but also lack convergence guarantees; this makes the importance weighted gradient estimate appealing, since we reuse the true posterior samples in an asymptotically unbiased estimator. However, it is known that finite-sample estimates can result in malformed energy function estimates. In fact, the above importance-sampled estimator severely underestimates the partition function, as the approximate posterior proposal distribution concentrates most of its energy to a small region in the latent manifold. A consequence of this is that the log likelihood estimates are massively upper-bounded, so alternative optimization methods that circumvent direct computation of the partition function's gradient are favoured.

One natural alternative is to optimize a second lower bound on the normalizing constant $\ln Z$ by exploiting Jensen's inequality, amortizing sampling in the energy model with another tractable parameterized density:

$$\ln Z = \ln(\mathop{\mathbb{E}}_{q_\zeta(z)}[\exp(-E_\beta(z) - \ln q_\zeta(z)]) \geq \mathop{\mathbb{E}}_{q_\zeta(z)}[-E_\beta(z) - \ln q_\zeta(z)] \tag{7}$$

Leading to the following upper bound on the VLB:

$$\mathcal{L}_{vae} \leq \mathop{\mathbb{E}}_{q_\phi(z|x)}[\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x)] - \mathop{\mathbb{E}}_{q_\zeta(z)}[-E_\beta(z) - \ln q_\zeta(z)] \tag{8}$$

With this, we end up with an adversarial objective in which the variational prior $q_\zeta(z)$ requires a tractable likelihood. This limits the choices of models vastly, requiring either powerful enough normalizing flows (Kingma et al., 2017; Papamakarios et al., 2018), whose expressiveness is inherently constrained by the computational considerations of computing the Jacobian determinant; or indirect methods that approximate its gradient (Grathwohl et al., 2021). Some methods end up resorting to MCMC sampling (Dieng et al., 2019), which reintroduces the computational difficulty back into the training scheme. Moreover, the additional adversarial objective within the variational framework can introduce possible instability, which is undesirable.

Instead, we take advantage of the fact that neither the inference model $q_\phi(z|x)$ nor the generative model $p_\alpha(x|z)$ depend on the normalization constant of the prior:

$$D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \int q_\phi(z|x) \ln \left(\frac{q_\phi(z|x)}{p_\theta(z|x)} \right) dz$$

$$D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \int q_\phi(z|x) \ln \left(\frac{q_\phi(z|x) \int p_\alpha(x|z) p_\beta(z) dz}{p_\alpha(x|z) p_\beta(z)} \right) dz$$

$$D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \int q_\phi(z|x) \ln \left(\frac{q_\phi(z|x) \int p_\alpha(x|z) \exp(-E_\beta(z)) dz}{p_\alpha(x|z) \exp(-E_\beta(z))} \right) dz$$

$$D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \int q_\phi(z|x) \ln \left(\frac{q_\phi(z|x) Z(\eta) dz}{p_\alpha(x|z) \exp(-E_\beta(z))} \right) dz$$

$$D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \mathop{\mathbb{E}}_{q_\phi(z|x)} [\ln q_\phi(z|x) + E_\beta(z) - \ln p_\alpha(x|z)] + \ln Z(\eta)$$

$$\ln Z(\eta) - D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\theta(z|x)) = \mathop{\mathbb{E}}_{q_\phi(z|x)} [\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x)]$$

The expectation on the right hand side forms a lower bound on $\ln Z(\eta)$, which is analogous to the log likelihood of the original VAE objective. Yet, maximizing this analogous lower bound (ALB) decreases the KL divergence between the approximate and ground-truth posterior distributions, making this a valid variational lower bound to train both $q_\phi(z|x)$ and $p_\alpha(x|z)$ without the dependence on the energy prior's normalizing constant. This independence from the normalizing constant is especially apparent when taking the gradient of the original variational lower bound with respect to the approximate posterior and the generator. This is appealing because the energy prior can be arbitrarily complex and still be computationally efficient, allowing us to obtain an unbiased estimate of this ALB with as little as one posterior sample.

Knowing this, it is easy to separate the optimization of the generative-inference pair, which is learned using the ALB, from the energy prior, giving us more freedom in the choice of its optimization strategy. From here, the EBM prior only requires approximating the aggregated posterior $q_\phi(z) = \int q_\phi(z|x)p(x)dx$ by reusing latent samples obtained from training the VAE. This comes from the fact that one of the terms in an alternative formulation of the VLB is a KL divergence between aggregate posterior and prior:

$$\mathcal{L}_{vae} = \mathop{\mathbb{E}}_{q_\phi(x,z)} [\ln p_\alpha(x|z)] - \mathbb{I}_{q(x,z)}[z;x] - D_{\mathrm{KL}}(q_\phi(z) \parallel p_\beta(z)) \tag{9}$$

With $\mathbb{I}_{q(x,z)}[z;x]$ being the mutual information between the observed data and latent variable. The negative KL divergence term represents an explicit objective, and the only objective, of minimizing the discrepancy between the aggregated posterior and reference prior. Additionally, this KL divergence corresponds to maximum likelihood learning of the energy prior on the aggregate posterior, meaning that existing explicit and implicit EBM objectives can also be employed here, resulting in a flexible framework for training energy-prior VAEs.

In this paper, we use sliced score matching (Song et al., 2019) to match the EBM prior to the aggregate posterior, which avoids expensive MCMC sampling by ignoring the intractable normalizing constant. However, we emphasize that any arbitrary objective that pushes the EBM prior towards the aggregated posterior can be used, such as Stein discrepancies (Grathwohl et al., 2020a), adversarial training (Grathwohl et al., 2021), and even self-normalizing and MCMC sampling; the latter two corresponding to optimizing the original VLB. With this, we now have a framework with which we can train arbitrarily complex generative models within rigorously defined theoretical grounding. Algorithm 1 shows the simple process.

Thanks to the generality of this framework, the choice of $E_\beta(z)$ can be arbitrary, ranging from simple restricted Boltzmann machines to large ResNets for higher-dimensional datasets. For simplicity, we will be focusing mainly on Gaussian-Bernoulli RBMs as the energy prior for the remainder of the paper. The Gaussian-Bernoulli RBM is a specific formulation of restricted Boltzmann machines in which the visible units parameterize a Gaussian distribution, while the hidden units parameterize a Bernoulli distribution, realizing a universal approximator of mixture models (Krause et al., 2013; Gu et al., 2022). The marginal energy of a Gaussian-Bernoulli RBM is as follows (Liao et al., 2022):

Algorithm 1 Training

1: **repeat**
2: $x \sim D(x)$
3: $z \sim q_\phi(z|x)$
4: Update $\nabla_{\alpha\phi}\, \mathbb{E}_{q_\phi(z|x)}[\ln p_\alpha(x|z) - E_\beta(z) - \ln q_\phi(z|x)]$
5: Train $E_\beta(z)$ on reused z using algorithm of choice
6: **until** converged

Algorithm 2 Variational Lower Bound Estimation

1: $x \sim D(x)$
2: $z \sim q_\phi(z|x)$
$T(x, z)$ is a function estimating $-D_{\mathrm{KL}}(q_\phi(z|x) \parallel p_\beta(z))$
3: $\mathcal{L}_{vae} = \mathbb{E}_{q_\phi(z|x)}[\ln p_\alpha(x|z) + T(x, z)]$
4: **return** $\mathcal{L}_{vae}$

$$E_\beta(z) = \frac{1}{2}(\frac{z-\mu}{\sigma})^\top(\frac{z-\mu}{\sigma}) - \mathrm{Softplus}(W^\top \frac{z}{\sigma^2} + b)^\top \mathbf{1} \tag{10}$$

Where μ and σ are the per-visible unit mean and standard deviation vectors, b is the hidden bias vector and W is the weight matrix of the RBM.

To sample from the RBM prior, we can use either the general block Gibbs sampling technique, or take advantage of the continuous nature of the GRBM and utilize Langevin dynamics:

$$z_0 \sim p_0(z), \quad z_{k+1} = z_k - \sigma \nabla_z E_\beta(z_k) + \sqrt{2\sigma}\epsilon_k, \quad k = 1, ..., K \tag{11}$$

Post-training evaluation of the VAE through maximum likelihood estimates is rather challenging, since we cannot entirely avoid the prior's normalizing constant. To that end, there exists approaches to estimating the partition function efficiently, such as annealed importance sampling (Salakhutdinov & Murray, 2008). In this paper, we take advantage of classifier-based density ratio matching (Mohamed & Lakshminarayanan, 2017) to estimate the KL divergence between the prior and approximate posterior implicitly, allowing us to obtain an estimate of the VLB without requiring an expensive and potentially non-convergent sampling procedure. To do so, we follow Mescheder et al. (2018) by training a classifier to contrast a pair of samples x and z between the joint distribution $p(x, z)$ and $p(x)p(z)$.

3 Related Work

Incorporating flexible priors such as energy-based, score/diffusion-based, and mixture priors into variational autoencoders (Vahdat et al., 2021; Han et al., 2020; Lee et al., 2023; Rombach et al., 2022) and also regular autoencoders (Ghosh et al., 2020; Jing et al., 2020) is not a new concept, and has seen some considerable success. However, many of these attempts have approached this idea from a fundamentally different perspective that divorces the objective from theoretically sound probabilistic frameworks (e.g. variational inference), resulting in what is essentially just a "packaging" of two different models. This can have potentially suboptimal effects, especially for regular autoencoders due to their non-probabilistic nature. As a result, attempts to enhance the prior distribution for these models (Ghosh et al., 2020; Jing et al., 2020) have not yielded substantial improvements.

The idea of matching a learnable prior to the aggregated posterior is motivated by the pursuit of safely maximizing mutual information between the latent code and observed data — resulting in better reconstruction quality and representation learning — while simultaneously learning a good generative model (Alemi et al., 2018). A special case of this is the VampPrior (Tomczak & Welling, 2018), which parameterizes a constrained mixture model with the approximate posterior via a finite collection of pseudo-inputs to capture a rough approximation of the true aggregated posterior. In our case, we use an energy prior to allow for arbitrary flexibility to balance quality and generalization.

Our proposed method of incorporating unnormalized, flexible priors into the VAE framework is orthogonal to the use of normalizing flows for improving posterior estimates (Rezende & Mohamed, 2016; Grathwohl et al., 2018). While both strategies aim to make models more expressive, they address considerably different

limitations. Normalizing flows focus on improving the posterior distribution $q_\phi(z|x)$ by applying a series of invertible transformations to a simple base distribution (typically Gaussian), introducing more flexibility into the posterior and allowing it to better approximate the true latent distribution. However, normalizing flows are subject to important design constraints in order to ensure computational efficiency, which places a natural limit on how complex or expressive the posterior can be. EBMs do not have such a restriction, allowing for arbitrary flexibility with considerably less restrictions.

Another similar approach is adversarial Variational Bayes (Mescheder et al., 2018), where the posterior is matched to an arbitrary prior *implicitly* via an adversarial objective. Unlike our energy-based approach which can be trained with any arbitrary objective with a malleable prior, the AVB objective is based on black-box density ratio matching that requires samples from a predefined prior.

Noise Contrastive Priors (Aneja et al., 2021) are very similar in principle to both AVB as well as our framework in that ratio matching used to learn a prior. The discriminator in this case represents an exponential tilting of a base distribution, whose optimal value tilts the base distribution to the aggregated posterior. However, whereas our framework allows joint training of a VAE alongside the EBM, thus actively shaping the VAE, the discriminator in NCP is learned ex-post (Ghosh et al., 2020; Dai & Wipf, 2019), i.e. on latent variables of a pretrained VAE. The latter approach falls short of the key benefits of learning a VAE through an expressive prior, such as improved reconstruction quality and representation learning, as the VAE is otherwise a standard Gaussian VAE.

Most closely related to our work is the generator with a latent-space EBM in Pang et al. (2020). Here, the authors address the exact same problem of learning an energy-based prior through variational inference, making our work orthogonal to theirs. Unlike our work, where we provide a general framework for training VAEs with energy priors, the authors focused on a specific instantiation that uses MCMC sampling to learn the EBM. Recognizing that sampling in a pure EBM prior is not guaranteed to converge, they instead resort to modeling the prior as an exponentially-tilted distribution, allowing them to train their model reasonably efficiently through short-run MCMC correction of samples from a base distribution. Additionally, albeit for simplicity, the authors also sample directly from the true posterior; we amortized this costly procedure by jointly learning an approximate posterior, making clever use of the fact that it does not depend on the EBM prior's normalizing constant. The fully amortized model that is proposed, but not empirically verified, in their paper is the adversarial instantiation discussed in our framework.

4 Experiments

We demonstrate the validity and effectiveness of the unnormalized prior VLB through density estimation on image datasets.

4.1 MNIST

For this experiment, we train a fully linear VAE on dynamically binarized MNIST for 100 iterations. The VAE encoder is comprised of sizes 784-512-512-128 with ReLUs in between, and the decoder similarly is comprised of sizes 64-512-512-784 with ReLUs in between. The output of the encoder is split into the mean and log-variance of the Gaussian approximate posterior, and the output of the decoder are Bernoulli logits. The prior is a GRBM with 64 visible units and 128 hidden units, which is trained using sliced score matching. The Adam optimizer with learning rates of 0.001 and 0.0002 are used for the VAE and prior respectively. A batch size of 512 is used for faster training. Reconstructions and samples from the prior are shown in Figure 1. Negative log likelihood is reported in Table 1. Extended samples are available in Appendix A1.

Results

Reconstructions in the VAE are of very high quality, which is expected as the prior and approximate posterior are well suited for each other. Samples from the EBM prior are also very high quality, albeit with noticeable artifacts and quality that is slightly worse than competing models, reinforcing the idea that good likelihood is not necessarily consistent with good sample generation (Theis et al., 2016). The variational lower bound is exceptionally low, outperforming the state-of-the-art generative models, demonstrating that EBM priors can

massively improve the overall quality of the probabilistic model. This is especially intriguing considering that the choice of neural network is vastly simpler — just a stack of dense layers — than the massive hierarchical models that are utilized in modern VAEs.

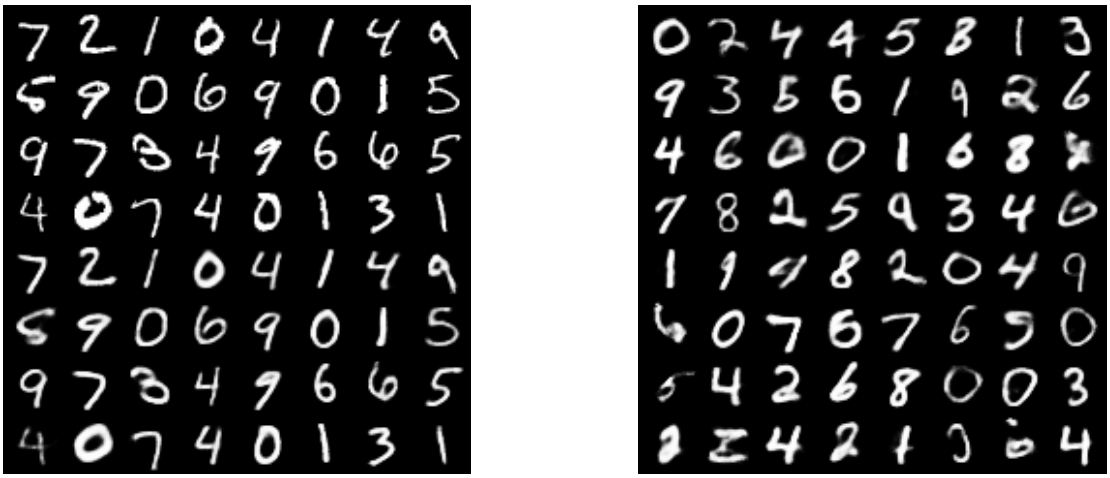

Figure 1: Left: VAE reconstructions (bottom half) of test images (top half). Right: Samples generated from GRBM prior using 100 Langevin steps.

4.2 CIFAR-10

For this experiment, we train a convolutional VAE on 8-bit CIFAR-10 for 100 iterations. The VAE encoder is composed of four 4x4 convolutions of stride 2 and padding 1, with channel sizes of 32, 64, 128 and 256 respectively. ReLU activations are placed after each convolution in the encoder. A linear layer mapping to 512 latent units for both the mean and variance of the Gaussian approximate posterior is the final layer of the encoder. The decoder is composed of four transposed convolutions of channel sizes 128, 64, 32 and 3, all 4x4 kernels with stride of 2 and padding of 1. ReLU activations are placed after all transposed convolutions except the last layer, which maps to the mean of the Laplace distribution. The scale of the distribution is computed analytically using the L1 loss, a technique known as decoder calibration (Rybkin et al., 2021). The prior is a GRBM with 512 visible units and 128 hidden units, which is trained using sliced score matching. The Adam optimizer with learning rates of 0.001 and 0.0002 are used for the VAE and prior respectively. A batch size of 1024 is used for faster training. Reconstructions and samples from the prior are shown in Figure 2. Negative log likelihood in bits-per-dimension (BPD) is reported in Table 1.

To facilitate calculating BPD, we use the same preprocessing step as (Papamakarios et al., 2018). In short, we first dequantize the images with uniform noise and rescaling to the interval $[0, 1]$ (Theis et al., 2016). Then, we change domains to $(-\inf, \inf)$ by applying a logit transform.

Results

Compared to binarzed MNIST, the CIFAR-10 dataset is significantly more difficult for generative models to solve due to the massive variation that exists in the training samples. The VAE gives very good reconstructions, but the samples are poor in comparison to the state-of-the-art. This is especially so considering that the BPD of the energy VAE is the worst. This is expected with the inexpressive network architecture and training steps, as our goal is to not achieve state-of-the-art but to emphasize the relative strength of this approach. As the results demonstrate, the energy-prior VAE is very good even in comparison to a dense masked autoregressive flow (Papamakarios et al., 2018), demonstrating promising performance for VAEs with energy priors.

Figure 2: Left: VAE reconstructions (bottom half) of test images (top half). Right: Samples generated from GRBM prior using 50 Langevin steps.

Model	MNIST	CIFAR-10
NVAE w/o flow (Vahdat & Kautz, 2021)	78.01	2.93
IAF-VAE (Kingma & Dhariwal, 2018)	79.10	3.11
CR-NVAE (Sinha & Dieng, 2022)	76.93	**2.51**
BFN (Graves et al., 2024)	77.87	2.66
MAF (10) (Papamakarios et al., 2018)	—	4.31
VAE w/ GRBM prior	**71.77**	5.98

Table 1: Model comparison on binarized MNIST and CIFAR-10 test data. Scores highlighted in bold means best.

5 Conclusion

In this paper, we proposed a novel variational inference framework that integrates unnormalized energy-based priors into the Variational Autoencoder (VAE) model. By replacing the traditional normalized prior with a more flexible energy-based distribution, we addressed key limitations of VAEs, particularly their inability to model complex, multimodal data distributions. Our method demonstrated both theoretical and practical advantages, including improved likelihood estimation and generative performance, as well as scalable training of energy-based models without relying on expensive Markov Chain Monte Carlo (MCMC) sampling. We empirically validated our approach image datasets, showing that energy-based VAEs (EVAEs) perform well in terms of capturing complex data distributions and producing high-quality generative models. Although our experiments primarily focused on Gaussian-Bernoulli RBM priors, the framework is versatile and can be applied to a wide range of unnormalized priors.

6 Discussion

The introduction of unnormalized priors into VAEs offers a new perspective on generative modeling by bridging the gap between VAEs and energy-based models (EBMs). Unlike prior work that seeks to enhance generative models by combining different techniques without a unified probabilistic foundation, our approach maintains the rigor of variational inference. This not only ensures the tractability of likelihood-based training but also leverages the expressiveness of energy-based models to enrich the latent space of the VAE. By doing so, we have effectively addressed one of the core limitations of VAEs—namely, the mismatch between simple priors and complex posterior distributions.

Our experiments on the datasets highlight the capability of our model to better capture multimodal and intricate data distributions compared to standard VAEs. The results suggest that, provided good combination of architecture and training scheme, energy-prior VAEs can be very competitive generative models.

Future work could explore alternative strategies for posterior optimization, including hybrid approaches that combine energy-based priors with more expressive posterior distributions like normalizing flows. Using unnormalized distributions as hierarchical priors could also potentially improve the representation of complex data structures by conserving energy across layers. Additionally, applying this framework to more diverse types of data, such as audio and text, would provide further insights into the generalizability and performance of this framework.

References

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken elbo, 2018. URL https://arxiv.org/abs/1711.00464.

Jyoti Aneja, Alexander Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors, 2021. URL https://arxiv.org/abs/2010.02917.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders, 2016. URL https://arxiv.org/abs/1509.00519.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images, 2021. URL https://arxiv.org/abs/2011.10650.

Bin Dai and David Wipf. Diagnosing and enhancing vae models, 2019. URL https://arxiv.org/abs/1903.05789.

Adji B. Dieng, Francisco J. R. Ruiz, David M. Blei, and Michalis K. Titsias. Prescribed generative adversarial networks, 2019. URL https://arxiv.org/abs/1910.04302.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models, 2020. URL https://arxiv.org/abs/1903.08689.

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders, 2020. URL https://arxiv.org/abs/1903.12436.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models, 2018. URL https://arxiv.org/abs/1810.01367.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling, 2020a. URL https://arxiv.org/abs/2002.05616.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one, 2020b. URL https://arxiv.org/abs/1912.03263.

Will Grathwohl, Jacob Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No mcmc for me: Amortized sampling for fast and stable training of energy-based models, 2021. URL https://arxiv.org/abs/2010.04230.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks, 2024. URL https://arxiv.org/abs/2308.07037.

Linyan Gu, Lihua Yang, and Feng Zhou. Approximation properties of gaussian-binary restricted boltzmann machines and gaussian-binary deep belief networks. *Neural Networks*, 153:49–63, 2022. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.05.020. URL https://www.sciencedirect.com/science/article/pii/S0893608022001940.

Tian Han, Erik Nijkamp, Linqi Zhou, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. Joint training of variational auto-encoder and latent energy-based model, 2020. URL https://arxiv.org/abs/2006.06059.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Li Jing, Jure Zbontar, and Yann LeCun. Implicit rank-minimizing autoencoder, 2020. URL https://arxiv.org/abs/2010.00679.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018. URL https://arxiv.org/abs/1807.03039.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow, 2017. URL https://arxiv.org/abs/1606.04934.

Oswin Krause, Asja Fischer, Tobias Glasmachers, and Christian Igel. Approximation properties of DBNs with binary hidden units and real-valued visible units. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 419–426, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/krause13.html.

Hankook Lee, Jongheon Jeong, Sejun Park, and Jinwoo Shin. Guiding energy-based models via contrastive latent variables, 2023. URL https://arxiv.org/abs/2303.03023.

Renjie Liao, Simon Kornblith, Mengye Ren, David J. Fleet, and Geoffrey Hinton. Gaussian-bernoulli rbms without tears, 2022. URL https://arxiv.org/abs/2210.10318.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks, 2018. URL https://arxiv.org/abs/1701.04722.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models, 2017. URL https://arxiv.org/abs/1610.03483.

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model, 2020. URL https://arxiv.org/abs/2006.08205.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, 2018. URL https://arxiv.org/abs/1705.07057.

Adrian Alan Pol, Victor Berger, Gianluca Cerminara, Cecile Germain, and Maurizio Pierini. Anomaly detection with conditional variational autoencoders, 2020. URL https://arxiv.org/abs/2010.05531.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016. URL https://arxiv.org/abs/1505.05770.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective vae training with calibrated decoders, 2021. URL https://arxiv.org/abs/2006.13202.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*, 2008. URL https://api.semanticscholar.org/CorpusID:458722.

Samarth Sinha and Adji B. Dieng. Consistency regularization for variational auto-encoders, 2022. URL https://arxiv.org/abs/2105.14859.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation, 2019. URL https://arxiv.org/abs/1905.07088.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, 2016. URL https://arxiv.org/abs/1511.01844.

Jakub M. Tomczak and Max Welling. Vae with a vampprior, 2018. URL https://arxiv.org/abs/1705.07120.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder, 2021. URL https://arxiv.org/abs/2007.03898.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space, 2021. URL https://arxiv.org/abs/2106.05931.

A Appendix

A.1 Binarized MNIST Linear VAE Extended

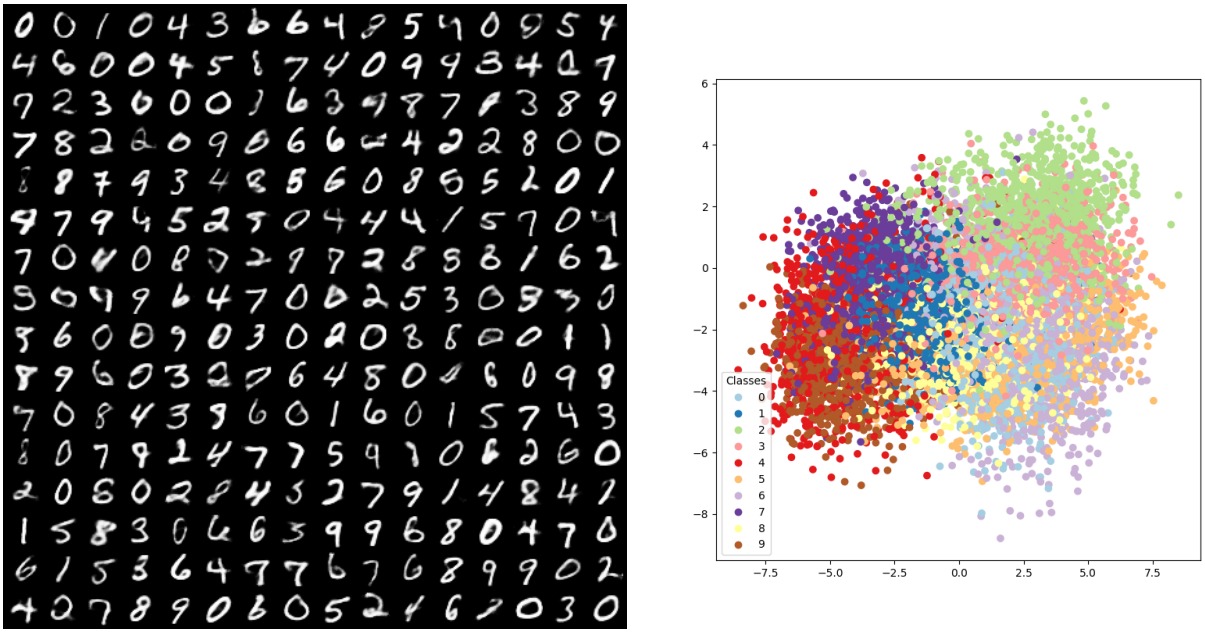

Figure 3: Left: Extended samples from the linear VAE. Right: Latent space (aggregated posterior) induced by the VAE encoder.