

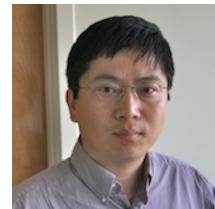
# Learning Energy-based Models by Diffusion Recovery Likelihood

Ruiqi Gao<sup>1</sup>, Yang Song<sup>2</sup>, Ben Poole<sup>3</sup>, Ying Nian Wu<sup>1</sup>, Diederik P. Kingma<sup>3</sup>

<sup>1</sup>UCLA, <sup>2</sup>Stanford University, <sup>3</sup>Google Brain

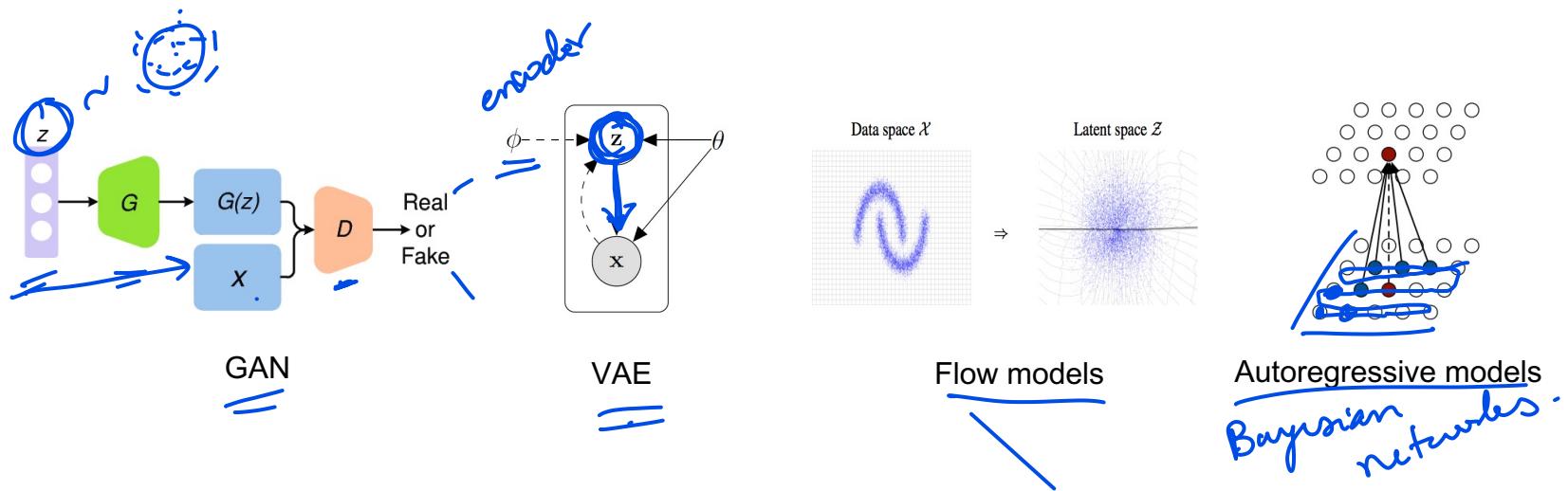
arXiv: 2012.08125

 ruiqigao@ucla.edu  
 @RuiqiGao



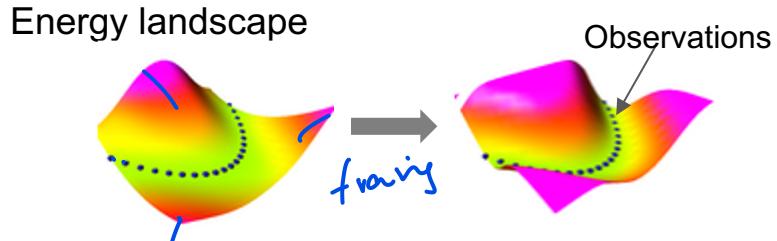
# Generative models

- Learning representations from data without labels



# Energy-based models

# Energy-based model (EBM)



$$p_{\theta}(\mathbf{x}) = \frac{1}{Z_{\theta}} \exp(f_{\theta}(\mathbf{x}))$$

$$\begin{aligned} f_{\theta} &= -E_{\theta} \\ f_{\theta}(\mathbf{x}) &\rightarrow R_{\text{geo}}^{\text{image}} \end{aligned}$$

- $Z_{\theta} = \int \exp(f_{\theta}(\mathbf{x}))d\mathbf{x}$  : partition function, analytically intractable
- Energy function:  $-f_{\theta}(\mathbf{x})$
- $f_{\theta}: \mathbb{R}^D \rightarrow \mathbb{R}$  : free-form. Easy to incorporate structure knowledge
- a generative version of a discriminator

$$p_{\theta_k}(x) = \frac{1}{Z(\theta_k)} \exp[f_{\theta_k}(x)] \longleftrightarrow P(k|x) = \frac{\exp(f_{\theta_k}(x) + b_k)}{\sum_{l=1}^K \exp(f_{\theta_l}(x) + b_l)}$$

# Two challenges for training EBMs

- Maximum likelihood estimation (MLE): gradient approximately follows

$$\frac{\partial}{\partial \theta} \mathbb{E}_{p_{\text{data}}} [\log p_{\theta}(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[ \frac{\partial}{\partial \theta} f_{\theta}(\mathbf{x}) \right]$$

Expected gradient of energy  
w.r.t. data distribution

$\frac{\partial}{\partial \theta} \log p_{\theta}$

Expected gradient of energy  
w.r.t. model distribution

contrastive  
training

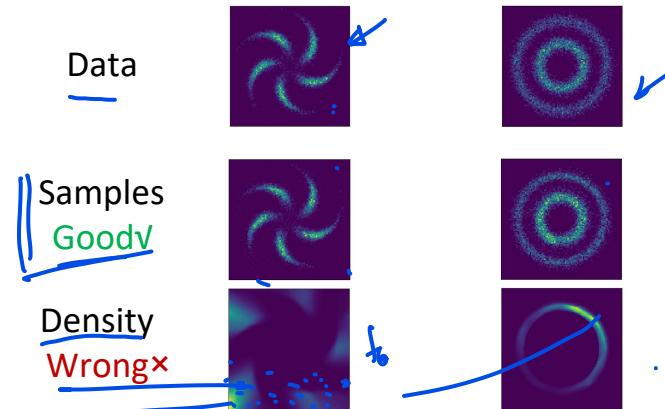
**Challenge 1:** requires MCMC to sample from model. **Extremely expensive** for high dimensional and multi-modal distributions. Difficult to converge.

- E.g. Langevin dynamics

$$\mathbf{x}^{\tau+1} = \mathbf{x}^{\tau} + \sqrt{\frac{\delta^2}{2}} \nabla_{\mathbf{x}} f_{\theta}(\mathbf{x}^{\tau}) + \delta \epsilon^{\tau}, \quad \epsilon^{\tau} \sim \mathcal{N}(0, \mathbf{I}).$$

# Two challenges for training EBMs

**Challenge 2:** density function learned with non-convergent MCMC can be **malformed**.



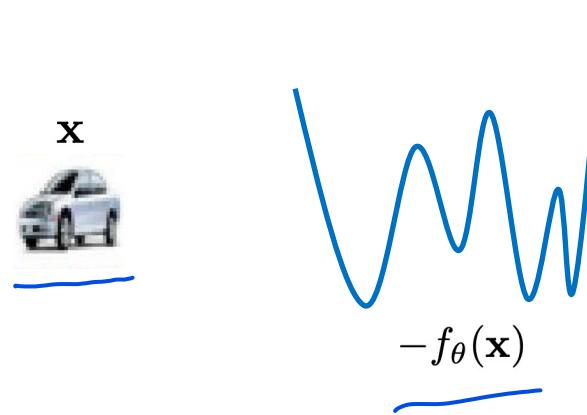
- No even long-run MCMC samples remain realistic. (Nijkamp et al. 2019)



# Our method

# From marginal to conditional

Q: Can we simplify the energy so that MCMC becomes easier?

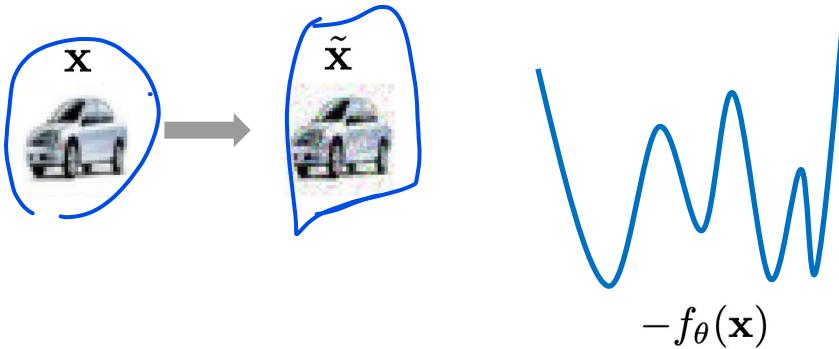


# From marginal to conditional

Q: Can we simplify the energy so that MCMC becomes easier?

A: Yes! Switch attention from marginal to conditional distributions.

Assume  $p_\theta(\mathbf{x}) = \frac{1}{Z_\theta} \exp(f_\theta(\mathbf{x}))$ . Let  $\tilde{\mathbf{x}} = \mathbf{x} + \sigma\epsilon$



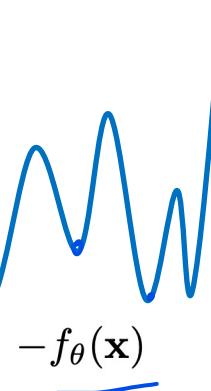
# From marginal to conditional

Q: Can we simplify the energy so that MCMC becomes easier?

A: Yes! Switch attention from **marginal** to **conditional** distributions.

Assume  $p_\theta(\mathbf{x}) = \frac{1}{Z_\theta} \exp(f_\theta(\mathbf{x}))$ . Let  $\tilde{\mathbf{x}} = \mathbf{x} + \sigma\epsilon$ , then we have

$$p_\theta(\mathbf{x}|\tilde{\mathbf{x}}) = \frac{1}{\tilde{Z}_\theta(\tilde{\mathbf{x}})} \exp \left( f_\theta(\mathbf{x}) - \frac{1}{2\sigma^2} \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 \right)$$



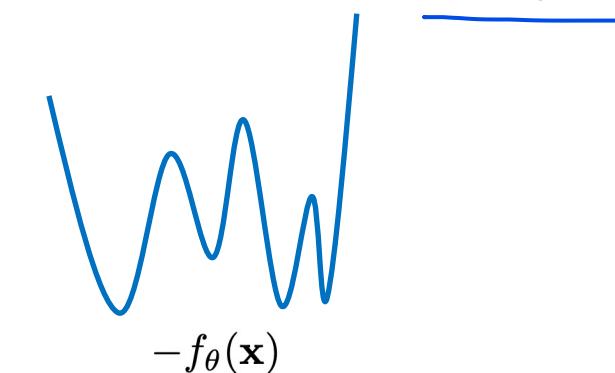
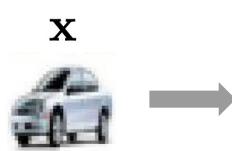
# From marginal to conditional

Q: Can we simplify the energy so that MCMC becomes easier?

A: Yes! Switch attention from **marginal** to **conditional** distributions.

Assume  $p_\theta(\mathbf{x}) = \frac{1}{Z_\theta} \exp(f_\theta(\mathbf{x}))$ . Let  $\tilde{\mathbf{x}} = \mathbf{x} + \sigma\epsilon$ , then we have

$$p_\theta(\mathbf{x}|\tilde{\mathbf{x}}) = \frac{1}{\tilde{Z}_\theta(\tilde{\mathbf{x}})} \exp \left( f_\theta(\mathbf{x}) - \underbrace{\frac{1}{2\sigma^2} \|\tilde{\mathbf{x}} - \mathbf{x}\|^2}_{\text{Localize the energy landscape}} \right)$$



# From marginal to conditional

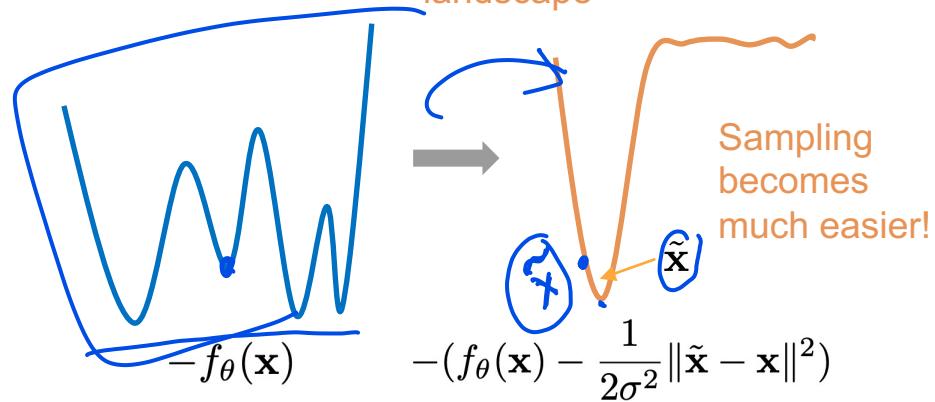
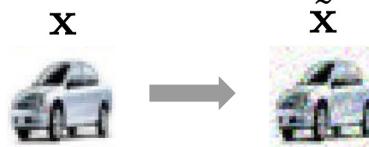
Q: Can we simplify the energy so that MCMC becomes easier?

A: Yes! Switch attention from **marginal** to **conditional** distributions.

Assume  $p_\theta(\mathbf{x}) = \frac{1}{Z_\theta} \exp(f_\theta(\mathbf{x}))$ . Let  $\tilde{\mathbf{x}} = \mathbf{x} + \sigma\epsilon$ , then we have

$$p_\theta(\mathbf{x}|\tilde{\mathbf{x}}) = \frac{1}{\tilde{Z}_\theta(\tilde{\mathbf{x}})} \exp \left( f_\theta(\mathbf{x}) - \frac{1}{2\sigma^2} \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 \right)$$

Localize the energy landscape



# Maximizing recovery likelihood

Define recovery log-likelihood function

$$\mathcal{J}(\theta) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(\mathbf{x}_i | \tilde{\mathbf{x}}_i)$$

# Maximizing recovery likelihood

Define recovery log-likelihood function

$$\mathcal{J}(\theta) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(\mathbf{x}_i | \tilde{\mathbf{x}}_i)$$

- Same learning gradients as MLE, which approximately follows

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[ \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{x} \sim p_\theta} \left[ \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) \right]$$

- Only need to run Langevin dynamics sampling from the conditional distribution, starting from  $\tilde{\mathbf{x}}$

$$\mathbf{x}^{\tau+1} = \mathbf{x}^\tau + \frac{\delta^2}{2} (\nabla_{\mathbf{x}} f_\theta(\mathbf{x}^\tau) + \frac{1}{\sigma^2} (\tilde{\mathbf{x}} - \mathbf{x}^\tau)) + \delta \epsilon^\tau$$

# Maximizing recovery likelihood

Define recovery log-likelihood function

$$\mathcal{J}(\theta) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(\mathbf{x}_i | \tilde{\mathbf{x}}_i)$$

- Same learning gradients as MLE, which approximately follows

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[ \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{x} \sim p_\theta} \left[ \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) \right]$$

- Only need to run Langevin dynamics sampling from the **conditional distribution**, starting from  $\tilde{\mathbf{x}}$

$$\mathbf{x}^{\tau+1} = \mathbf{x}^\tau + \frac{\delta^2}{2} (\nabla_{\mathbf{x}} f_\theta(\mathbf{x}^\tau) + \frac{1}{\sigma^2} (\tilde{\mathbf{x}} - \mathbf{x}^\tau)) + \delta \boldsymbol{\epsilon}^\tau$$

- Maximizing recovery likelihood gives a **consistent estimator** of  $\theta$  in  $p_\theta(\mathbf{x})$ !

## Diffusion recovery likelihood

$$\overset{\circ}{x}_t \times \sqrt{\sigma_t} = \tilde{x}_t$$

We propose to learn a sequence of recovery likelihoods on diffusion data

$$\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x}); \quad \mathbf{x}_{t+1} = \sqrt{1 - \sigma_{t+1}^2} \mathbf{x}_t + \sigma_{t+1} \boldsymbol{\epsilon}_{t+1}, \quad t = 0, 1, \dots, T-1.$$

Let  $\mathbf{y}_t = \sqrt{1 - \sigma_{t+1}^2} \mathbf{x}_t$ , assume a sequence of conditional EBMs

$$p_\theta(\mathbf{y}_t | \mathbf{x}_{t+1}) = \frac{1}{\tilde{Z}_{\theta,t}(\mathbf{x}_{t+1})} \exp \left( f_\theta(\mathbf{y}_t, t) - \frac{1}{2\sigma_{t+1}^2} \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 \right).$$

# Diffusion recovery likelihood

We propose to learn a sequence of recovery likelihoods on diffusion data

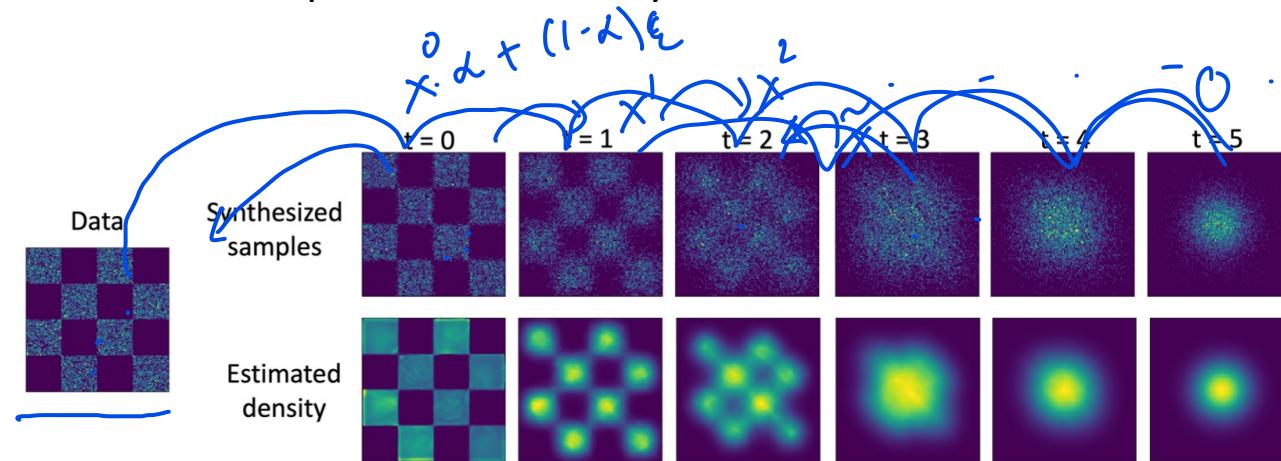
$$\underline{\mathbf{x}_0} \sim p_{\text{data}}(\mathbf{x}); \quad \underline{\mathbf{x}_{t+1}} = \sqrt{1 - \sigma_{t+1}^2} \mathbf{x}_t + \sigma_{t+1} \boldsymbol{\epsilon}_{t+1}, \quad t = 0, 1, \dots, T-1.$$

Let  $\mathbf{y}_t = \sqrt{1 - \sigma_{t+1}^2} \mathbf{x}_t$ , assume a sequence of conditional EBM

$$p_\theta(\mathbf{y}_t | \mathbf{x}_{t+1}) = \frac{1}{\tilde{Z}_{\theta,t}(\mathbf{x}_{t+1})} \exp \left( \boxed{f_\theta(\mathbf{y}_t, t)} - \frac{1}{2\sigma_{t+1}^2} \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 \right).$$

# Diffusion recovery likelihood

We propose to learn a sequence of recovery likelihoods on diffusion data.



Learning from **conditional distributions** gives accurate **marginal density** estimations!

# Results

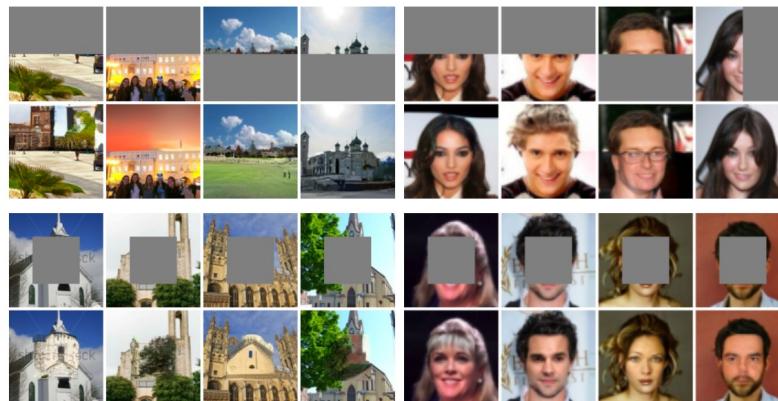
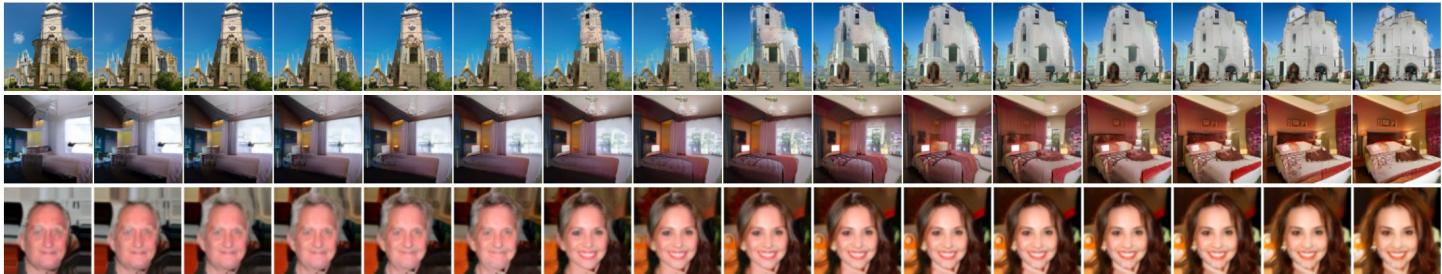
# High fidelity image generation



Table 1: FID and inception scores on CIFAR-10.

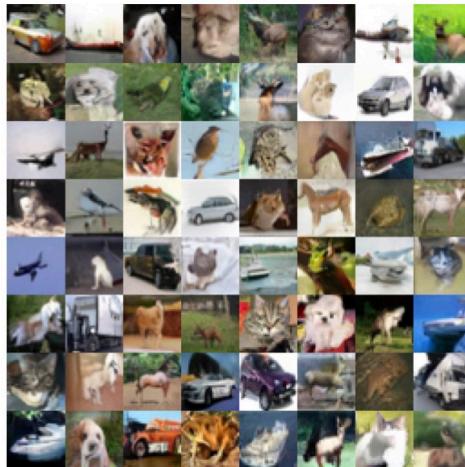
Model	FID↓	Inception↑
<b>GAN-based</b>		
WGAN-GP (Gulrajani et al., 2017)	36.4	7.86 ± .07
SNGAN (Miyato et al., 2018)	21.7	8.22 ± .05
SNGAN-DDLS (Che et al., 2020)	15.42	9.09 ± .10
StyleGAN2-ADA (Karras et al., 2020)	3.26	<b>9.74 ± .05</b>
<b>Score-based</b>		
NCSN (Song & Ermon, 2019)	25.32	8.87 ± .12
NCSN-v2 (Song & Ermon, 2020)	10.87	8.40 ± .07
DDPM (Ho et al., 2020)	<b>3.17</b>	9.46 ± .11
<b>Explicit EBM-conditional</b>		
CoopNets (Xie et al., 2019)	-	7.30
EBM-IG (Du & Mordatch, 2019)	37.9	8.30
JEM (Grathwohl et al., 2019)	38.4	8.76
<b>Explicit EBM</b>		
Muli-grid (Gao et al., 2018)	40.01	6.56
CoopNets (Xie et al., 2016a)	33.61	6.55
EBM-SR (Nijkamp et al., 2019b)	-	6.21
EBM-IG (Du & Mordatch, 2019)	38.2	6.78
<b>Ours (T6)</b>	<b>9.58</b>	<b>8.30 ± .11</b>

# Image interpolation & inpainting

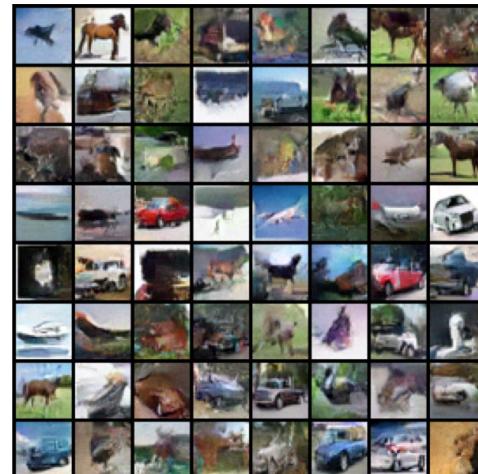


# Steady long-run MCMC chains

Short-run samples look fine for both methods. (100 steps)



Ours



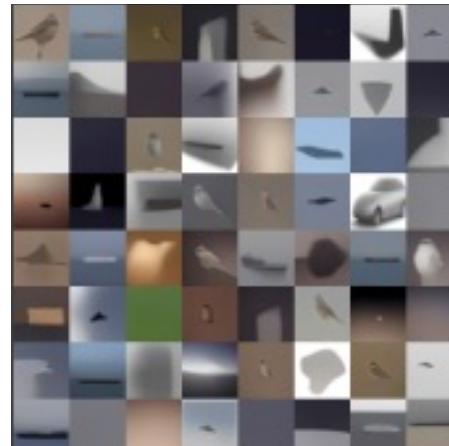
Learning marginal  $p(x)$  by short-run MCMC

# Steady long-run MCMC chains

Our samples remain realistic even after 100,000 steps.



Ours



Learning marginal  $p(x)$  by short-run MCMC

# Estimation of normalized density

Estimate  $Z_\theta$  by annealing importance sampling (AIS).

First to get competitive likelihood estimation using EBM on image datasets.

Table 4: Test bits per dimension on CIFAR-10.

Model	BPD $\downarrow$
DDPM (Ho et al., 2020)	3.70
Glow (Kingma & Dhariwal, 2018)	3.35
Flow++ (Ho et al., 2019)	3.08
GPixelCNN (Van den Oord et al., 2016)	3.03
Sparse Transformer (Child et al., 2019)	2.80
DistAug (Jun et al., 2020)	<b>2.56</b>
<b>Ours<math>^\dagger</math> (<math>T1k</math>)</b>	3.18

Thank you