Diffusion Models as Cartoonists: A Curious Case of High-Density Regions

Rafał Karczewski [1] Markus Heinonen [1] Vikas Garg [1, 2]

[1]Aalto University [2]YaiYai Ltd.

Overview

- We extend $\log p_t(\mathbf{x}_t)$ estimation in diffusion models to stochastic sampling;
- High-density regions are easily sampled from;
- Highest-density regions contain blurry images and cartoons - even though **there are no such examples in the training data!**

Density estimation in stochastic diffusion models

For a diffusion model

$$d\mathbf{x}_t = f(t)\mathbf{x}_t dt + g(t)dW_t$$

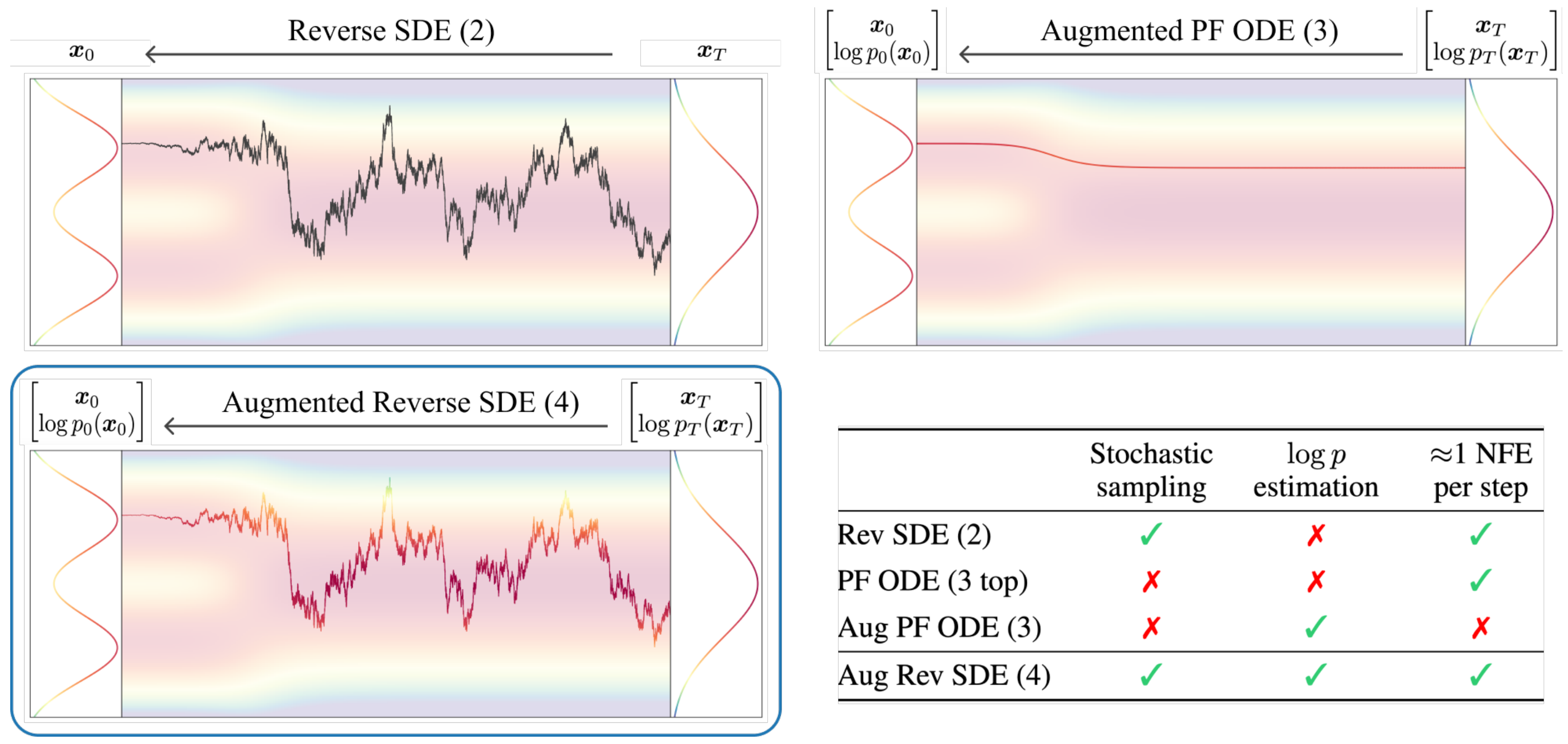
it is known that [1]:

$$d\begin{bmatrix}\mathbf{x}_t \\ \log p_t(\mathbf{x}_t)\end{bmatrix} = \begin{bmatrix} f(t)\mathbf{x}_t - \frac{1}{2}g^2(t)\nabla \log p_t(\mathbf{x}_t) \\ -f(t)D + \frac{1}{2}g^2(t)\,\mathrm{div}\,\nabla \log p_t(\mathbf{x}_t)\end{bmatrix} dt, \tag{1}$$

Our novel result:

$$d\begin{bmatrix}\mathbf{x}_t \\ \log p_t(\mathbf{x}_t)\end{bmatrix} = \begin{bmatrix} f(t)\mathbf{x}_t - g^2(t)\nabla \log p_t(\mathbf{x}_t) \\ -f(t)D - \frac{1}{2}g^2(t)\|\nabla \log p_t(\mathbf{x}_t)\|^2\end{bmatrix} dt + g(t)\begin{bmatrix}\boldsymbol{I}_D \\ \nabla \log p_t(\mathbf{x}_t)^T\end{bmatrix} d\overline{W}_t, \tag{2}$$

which only requires a single evaluation of the score and no higher order derivatives!



	Stochastic sampling	$\log p$ estimation	≈1 NFE per step
Rev SDE (2)	✓	✗	✓
PF ODE (3 top)	✗	✗	✓
Aug PF ODE (3)	✗	✓	✗
Aug Rev SDE (4)	✓	✓	✓

Figure 1. Our novel augmented SDE allows for density tracking for no extra cost.

Watch out for the bias!

When we approximate $\mathbf{s}_\theta(t,\mathbf{x}) \approx \nabla \log p_t(\mathbf{x})$, we have $p_t^{\mathrm{SDE}} \neq p_t^{\mathrm{ODE}}$, and:

- Equation 1 correctly estimates $\log p_0^{\mathrm{ODE}}(\mathbf{x}_0)$;
- Equation 2 provides a **biased estimate** $r_0 = \log p_0^{\mathrm{SDE}}(\mathbf{x}_0) + X$, where

$$\mathbb{E}X = \frac{T}{2}\mathbb{E}_{t\sim\mathcal{U}(0,T),\mathbf{x}_t\sim p_t^{\mathrm{SDE}}(\mathbf{x}_t)}\left[g^2(t)\underbrace{\|\mathbf{s}_\theta(t,\mathbf{x}_t) - \nabla \log p_t^{\mathrm{SDE}}(\mathbf{x}_t)\|^2}_{\text{score approximation error}}\right] \geq 0 \tag{3}$$

Finding highest-density regions

How to estimate the *denoising mode*?

$$\arg\max_{\mathbf{x}_0} p_{0|t}(\mathbf{x}_0|\mathbf{x}_t) = ? \tag{4}$$

We approach this by finding the *mode-tracking curve* starting at $(t, \mathbf{x}_t)$

$$\mathbf{y}_s := \arg\max_{\mathbf{x}_s} p_{s|t}(\mathbf{x}_s|\mathbf{x}_t). \tag{5}$$

The **mode-tracking curve is analytically obtainable!** We show that

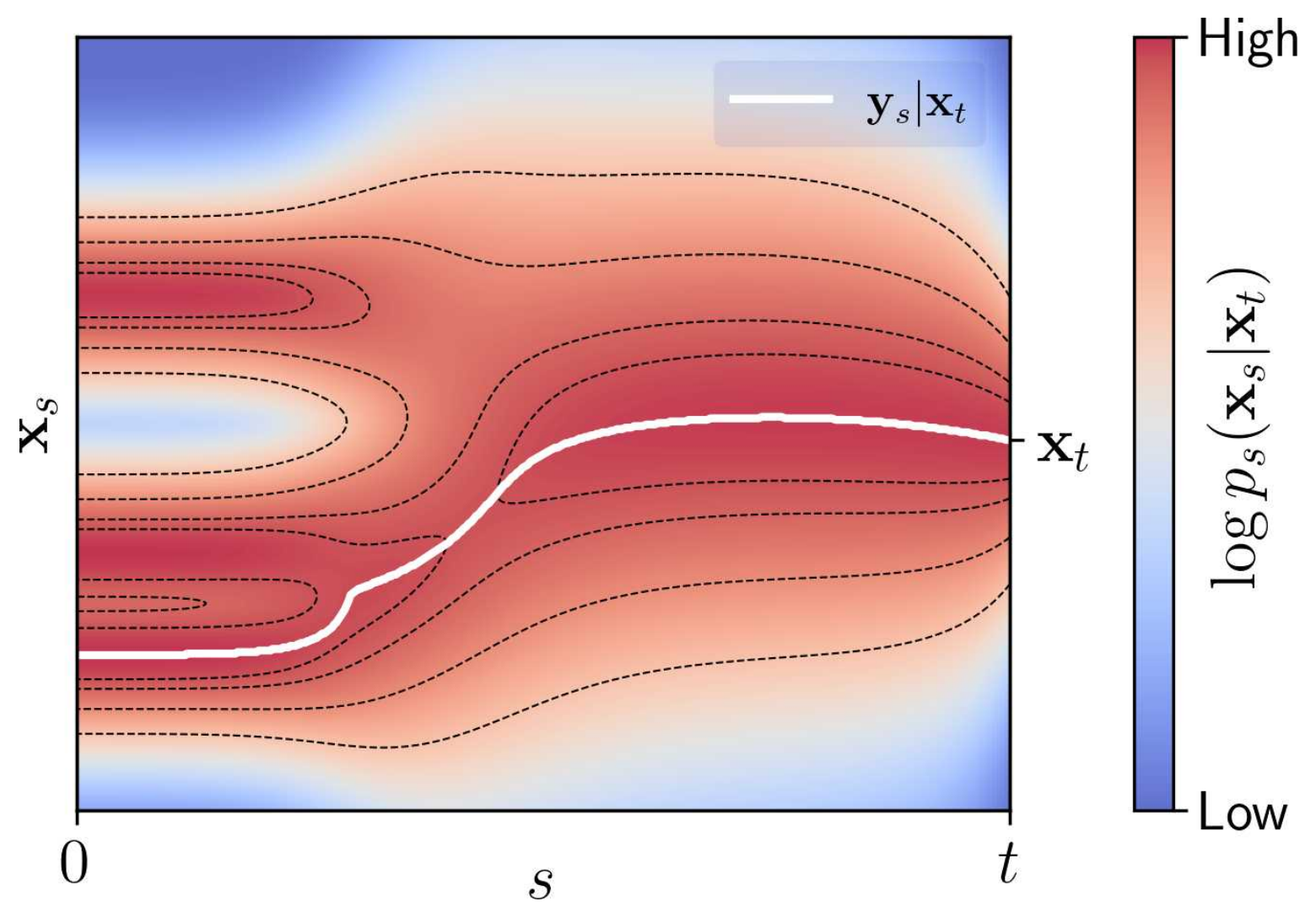


Figure 2. Mode-tracking curve.

$$\frac{d}{ds}\mathbf{y}_s = f(s)\mathbf{y}_s - g^2(s)\nabla \log p_s(\mathbf{y}_s) + \underbrace{H(s,\mathbf{y}_s)}_{\text{expensive}}. \tag{6}$$

We use Eq (2) to show that $H(s,\mathbf{y}) = 0$ yields extremely high-density samples.

Visualizing highest-density regions

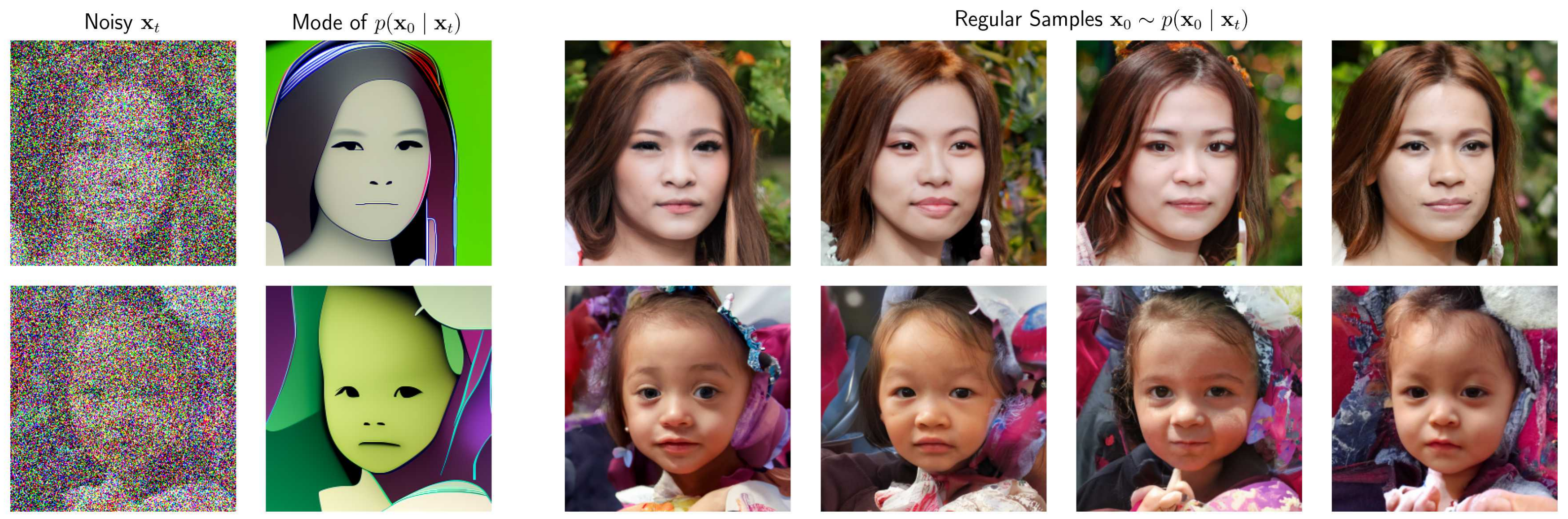

Noisy $\mathbf{x}_t$
Mode of $p(\mathbf{x}_0 \mid \mathbf{x}_t)$
Regular Samples $\mathbf{x}_0 \sim p(\mathbf{x}_0 \mid \mathbf{x}_t)$

Figure 3. Highest density images resemble cartoon drawings.

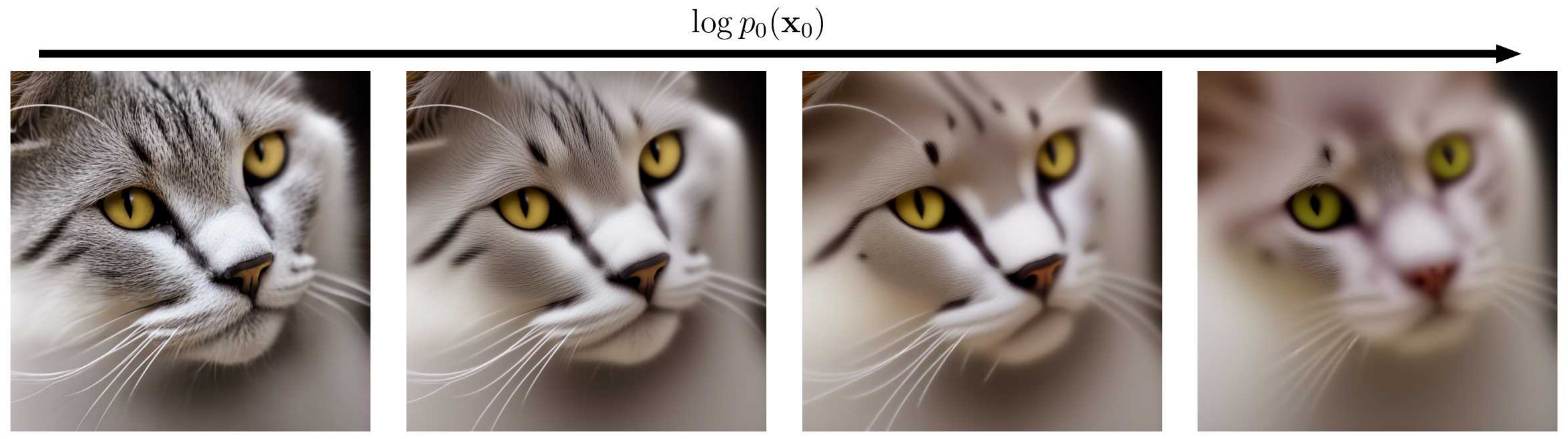


Figure 4. $\log p_0(\mathbf{x}_0)$ estimates correlates with amount of detail.

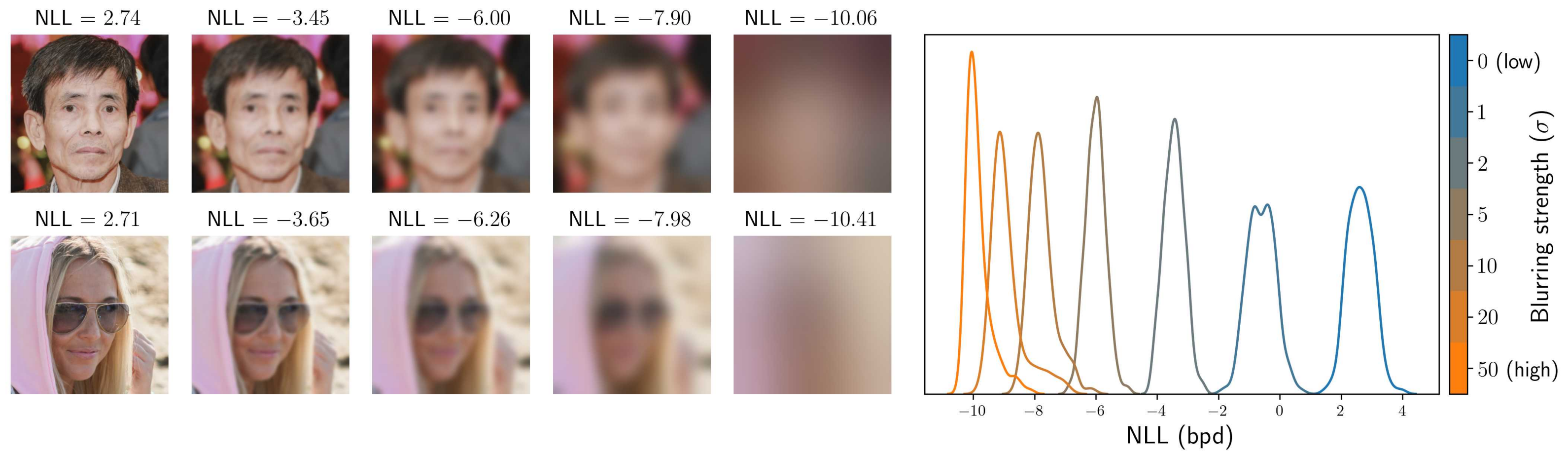


Figure 5. Adding blur to an image monotonically increases its estimated density.

What does log-density measure?

Why are highest-density regions occupied by unnatural images? It is known that generative models can assign higher densities to OOD than training data [2], so it does not measure *in-distributionness*.

Model's OOD freedom

The model only sees in-distribution examples during training. It is unconstrained in how it extrapolates to irregular examples.

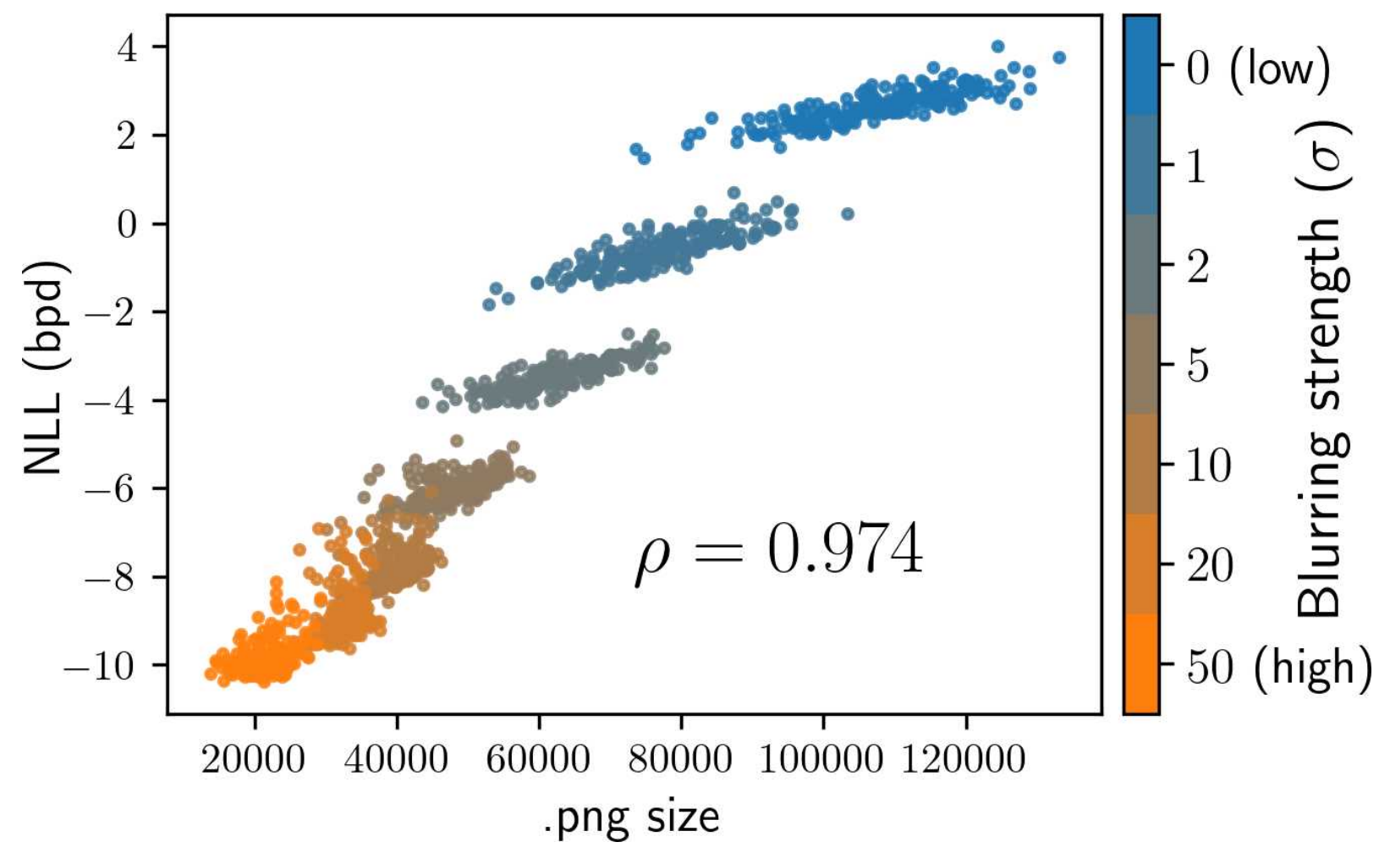


Figure 6. **Log-density correlates with information content** (.png size)

Information theory and compressibility

Information theory suggests that high-likelihood images should be more compressible, which translates to low-detail in the context of images. The highest-likelihood images are low-detail, simple, and lacking in complex textures, which makes them look like cartoons or appear blurry.

Density trade-off

The set of realistic images is vast. Some are high-detail, while others are low-detail, but the total number of high-detail images is significantly greater due to their higher number of degrees of freedom. Since probability density must integrate to 1, this forces the model to assign lower likelihood to high-detail images simply because there are so many of them.

High-dimensional high-density ”paradoxes”

The fact that the highest-density points look very different from regular samples is not unusual in high-dimensional probability distributions. In the standard D-dimensional Gaussian, the mode is at the origin, but as dimensionality increases, almost all samples are concentrated on a thin spherical shell at radius $\sqrt{D}$. Similarly, in diffusion models, the most frequently sampled (realistic) images form a high-dimensional structure away from the peak density points.

More on our blog

- More images and animations
- Follow-up work: *Density Guidance*
  - Log-density estimation to **arbitrary sampling paths**;
  - **Explicit control of** $\log p_0(\mathbf{x}_0)$ **even for stochastic sampling!**
  - Theoretical analysis of temperature scaling through the lens of log-density

SCAN ME

References

[1] Ricky Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[2] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv*, 2018.