DTU Compute
Department of Applied Mathematics and Computer Science

Danish Data Science Academy

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

DTU

Denoising Uncertainty via Gaussian Distributional Diffusion Models

Noa Margeta[1], Jes Frellsen[1,2], Ignacio Peis[1,2]

[1] Technical University of Denmark, Copenhagen, Denmark
[2] Pioneer Centre for Artificial Intelligence, Copenhagen, Denmark

Code Paper

1 Why model denoising uncertainty?

Standard diffusion models predict a *single* clean signal $\hat{\boldsymbol{x}}_0$, collapsing all explanations of $\boldsymbol{x}_t$ to a point:

$$p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) \longrightarrow \delta(\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0),$$

where $\hat{\boldsymbol{x}}_0 = f_\theta(\boldsymbol{x}_t)$. Yet the exact reverse transition averages over all clean-data hypotheses:

$$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \int q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)\, q(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)\, \mathrm{d}\boldsymbol{x}_0.$$

The reverse posterior has tractable moments:

$$\mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0] = A_t \boldsymbol{x}_0 + B_t \boldsymbol{x}_t,$$
$$\mathrm{Cov}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0] = \tilde{\beta}_t \boldsymbol{I},$$

where

$$A_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}, \qquad B_t = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t},$$
$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.$$

The dependence on the clean signal shrinks at noisier timesteps. Replacing $q(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)$ by a point mass cannot express denoising uncertainty at finite reverse steps.

2 Gaussian Distributional Diffusion

We learn a data-dependent diagonal Gaussian posterior:

$$p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\theta^{(\boldsymbol{x}_0)}(\boldsymbol{x}_t, t),\, \boldsymbol{\Sigma}_\theta^{(\boldsymbol{x}_0)}(\boldsymbol{x}_t, t)\big).$$

Marginalizing $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ over it remains Gaussian:

$$\boxed{\begin{aligned}\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) &= A_t \boldsymbol{\mu}_\theta^{(\boldsymbol{x}_0)} + B_t \boldsymbol{x}_t,\\ \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) &= \tilde{\beta}_t \boldsymbol{I} + A_t^2 \boldsymbol{\Sigma}_\theta^{(\boldsymbol{x}_0)}.\end{aligned}}$$

The learned uncertainty in $\boldsymbol{x}_0 \mid \boldsymbol{x}_t$ is propagated into each reverse step.

3 Equivalent noise parameterization

The forward relation is affine:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}.$$

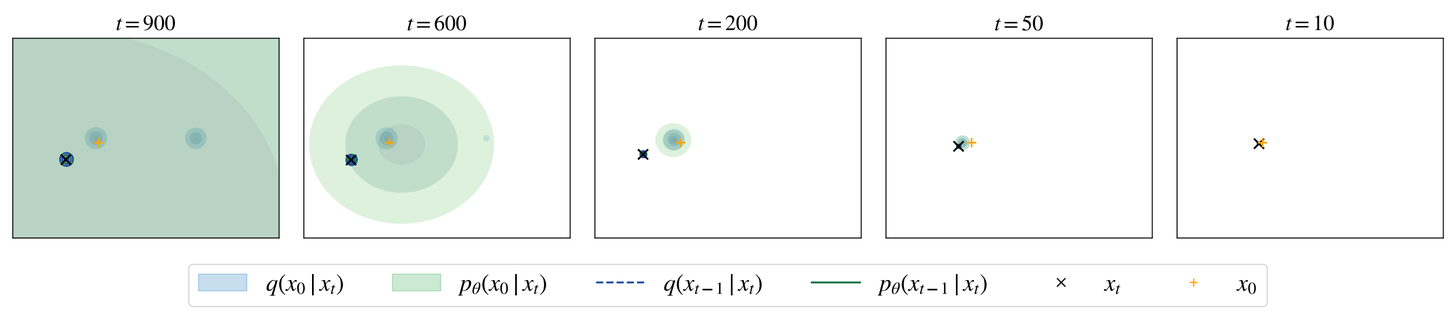
For fixed $\boldsymbol{x}_t$, a Gaussian over $\boldsymbol{x}_0$ induces an equivalent Gaussian over $\boldsymbol{\varepsilon}$:

$$p_\theta(\boldsymbol{\varepsilon} \mid \boldsymbol{x}_t) = \mathcal{N}\big(\boldsymbol{\mu}_\theta^{(\varepsilon)}(\boldsymbol{x}_t, t),\, \boldsymbol{\Sigma}_\theta^{(\varepsilon)}(\boldsymbol{x}_t, t)\big).$$

$$\boxed{\begin{aligned}\boldsymbol{\mu}_\theta^{(\varepsilon)} &= \frac{\boldsymbol{x}_t}{\sqrt{1-\bar{\alpha}_t}} - \sqrt{\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}}\,\boldsymbol{\mu}_\theta^{(\boldsymbol{x}_0)},\\ \boldsymbol{\Sigma}_\theta^{(\varepsilon)} &= \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\boldsymbol{\Sigma}_\theta^{(\boldsymbol{x}_0)}.\end{aligned}}$$

We predict $(\boldsymbol{\mu}_\theta^{(\varepsilon)}, \boldsymbol{\Sigma}_\theta^{(\varepsilon)})$ directly, retaining the DDPM interface.

Same DDPM backbone, with one additional diagonal-variance prediction and exact conversion between $\boldsymbol{x}_0$ and noise space.



Posterior and reverse-transition recovery. On a two-component Gaussian mixture, the true posterior is available in closed form. Filled regions show the true and learned $\boldsymbol{x}_0$ posteriors; contours show the corresponding true and learned reverse transitions. Uncertainty contracts with denoising.

4 Closed-form training objective

Let

$$C_t = \frac{\beta_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}, \qquad V_{t,i} = \tilde{\beta}_t + C_t^2 \Sigma_{\theta,i}^{(\varepsilon)}.$$

The per-step reverse-process KL is

$$\boxed{\mathcal{L}_t = \frac{1}{2}\sum_{i=1}^{d}\left[\log\frac{V_{t,i}}{\tilde{\beta}_t} + \frac{C_t^2(\varepsilon_i - \mu_{\varepsilon,i})^2 + \tilde{\beta}_t}{V_{t,i}} - 1\right].}$$

This heteroscedastic Gaussian loss reduces to weighted DDPM MSE as $\boldsymbol{\Sigma}_\theta^{(\varepsilon)} \to 0$.

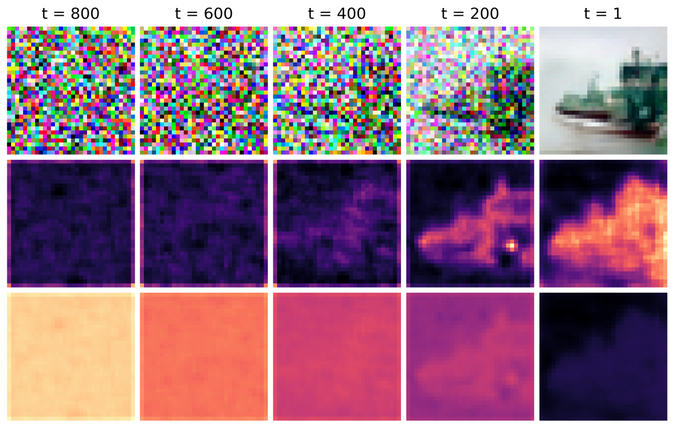
5 Better variational fidelity

CIFAR-10 likelihood and sample quality. GDDM obtains the tightest variational bound, while FID reflects the likelihood–perceptual-quality trade-off.

Model	BPD ↓	FID ↓
DDPM	4.534	5.53
Improved DDPM	4.272	**4.54**
GDDM	**4.244**	8.71

The tighter bound indicates a better reverse-posterior match; FID shows the likelihood–perceptual-quality trade-off.

6 Spatially structured uncertainty



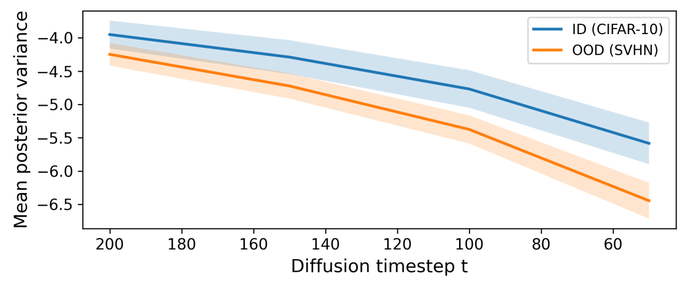
A CIFAR-10 denoising trajectory. Top: noisy input toward clean image; middle: normalized variance; bottom: global variance. Uncertainty contracts and localizes around boundaries and textures.

7 Uncertainty as an OOD signal

Average the predicted variance over pixels and a small set of early timesteps $\mathcal{T}$:

$$u_t(\boldsymbol{x}_t) = \frac{1}{d}\sum_{i=1}^{d}\Sigma_{\theta,i}^{(\boldsymbol{x}_0)}(\boldsymbol{x}_t, t), \qquad s(\boldsymbol{x}) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} u_t(\boldsymbol{x}_t).$$

One-step scoring. For $|\mathcal{T}| = 1$, diffuse the input directly to one selected early timestep and evaluate the variance head once. No iterative reverse trajectory or reconstruction is required.



ID/OOD separation. Mean predicted log-variance follows distinct trajectories for CIFAR-10 and SVHN. One variance evaluation achieves 97.2% AUROC, compared with 2812 evaluations for reconstruction-based scoring.

8 Take-home messages

- Model the second moment of $p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)$, not only its mean.
- Closed-form training and reverse transitions.
- Better CIFAR-10 BPD than both DDPM baselines.
- Competitive OOD detection with one model evaluation.
- Limitation: Gaussians cannot capture multimodality.

Code: `https://github.com/nmarg/gddm`

9 Acknowledgements

This research was supported by the Villum Foundation through the Synergy project number 50091, by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (MLLS, grant no. NNF20OC0062606), and by the Independent Research Fund Denmark (grant no. 5334-00122B and 5334-00076B). Jes Frellsen was further supported by funding from the Reinholdt W. Jorck og Hustrus Fond. Ignacio Peis acknowledges support by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429).