# Diffusion Models with Learned Adaptive Noise

Subham Sekhar Sahoo, Aaron Gokaslan, Chris De Sa, Volodymyr Kuleshov
Cornell Tech, NYC, USA

Spotlight — NEURAL INFORMATION PROCESSING SYSTEMS

Video Tutorial, Code, and more …

## MOTIVATION

We learn **forward** and **reverse** diffusion processes **jointly** to achieve **SOTA** likelihoods with diffusion models.

### Goal

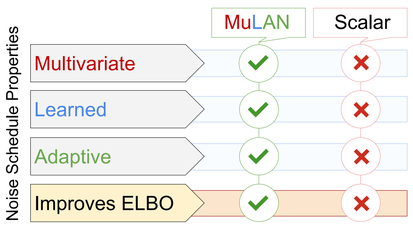
- Jointly optimize the "forward / noising" process and the "reverse / denoising" process for a **tighter** ELBO.
- Current diffusion models **ONLY** optimize the reverse process.

### Challenge

ELBO is **invariant** to (scalar) noise schedules [1].

### Solution

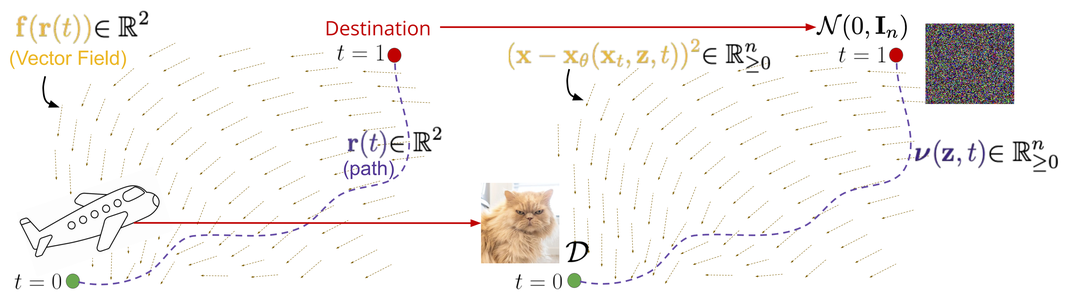
- **Key Discovery:** Multivariate schedules can alter ELBO.
- **Mu**ltivariate **L**earned Adaptive **N**oise schedule (**MuLAN**) **tightens** the ELBO.




## INTUITION

Objects in a vector field follow the path of least resistance, **NOT** the shortest Euclidean (straight-line) path.

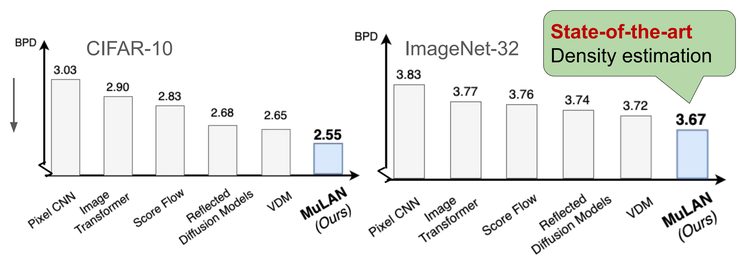
(a) Work Done = $\int_{t=0}^{t=1} \mathbf{f}(\mathbf{r}(t)) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}(t)\mathrm{d}t$ → (b) $\mathcal{L}_{\text{diff}} = -\frac{1}{2}\int_{t=0}^{t=1} (\mathbf{x} - \mathbf{x}_\theta(\mathbf{x}_t, \mathbf{z}, t))^2 \cdot \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\nu}(\mathbf{z}, t)\mathrm{d}t$




- The same intuition maps to diffusion processes:
  - **Objects**: Images
  - **Vector field**: per-pixel reconstruction error
  - **Path**: noise schedule
  - **Work Done**: Diffusion loss $\mathcal{L}_{\text{diff}}$
- **Scalar** noise schedule: **shortest Euclidean path** (i.e., straight line) from the data distribution to prior.
- Since **MuLAN** is multivariate (and adaptive), it learns a path of least resistance i.e., a noise schedule corresponding to the optimal ELBO.

## RESULTS

### Likelihood estimation on CIFAR-10 and ImageNet-32



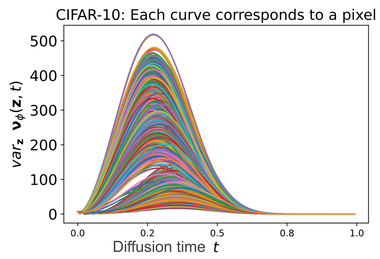

**State-of-the-art** Density estimation

### Faster Training

MuLAN augmented VDM achieves a BPD of
- 2.65 in **2M** steps (vs. **10M**) on CIFAR-10.
- 3.72 in **1M** steps (vs. **2M**) on ImageNet-32.

Upto **5x** faster training!

### Noise Schedule variations across different images




MuLAN generated samples

## NOTATION

$\mathbf{x}_0 \sim \mathcal{D}$ : Sample from the Data distribution
$\mathbf{x}_{(.)} \in \mathbb{R}^n$
$\mathbf{z} \in \{0, 1\}^m$
$m < n$
$0 \leq s < t \leq 1$
$\boldsymbol{\alpha}_s(\mathbf{z}), \boldsymbol{\alpha}_t(\mathbf{z}) \in [0, 1]^n$ $\quad$ $\boldsymbol{\sigma}_s(\mathbf{z}), \boldsymbol{\sigma}_t(\mathbf{z}) \in [0, 1]^n$
$\boldsymbol{\alpha}_{t|s}(\mathbf{z}) = \boldsymbol{\alpha}_t(\mathbf{z})/\boldsymbol{\alpha}_s(\mathbf{z})$ $\quad$ $\boldsymbol{\sigma}_{t|s}(\mathbf{z}) = \boldsymbol{\sigma}_t(\mathbf{z})/\boldsymbol{\sigma}_s(\mathbf{z})$
$\boldsymbol{\nu}_\phi(\mathbf{z}, t) = \boldsymbol{\alpha}_t^2(\mathbf{z})/\boldsymbol{\sigma}_t^2(\mathbf{z})$
$\cdot$ denotes dot product $\quad$ $\odot$ : element wise mult.
$\mathbf{x}_\theta(\mathbf{x}_t, \mathbf{z}, t)$ : Denoising Model with parameters $\theta$
$\phi$ : Parameters of the Encoder network

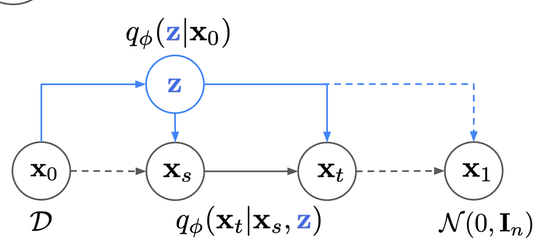
## REFERENCES

[1] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. Advances in neural information processing systems, 34:21696–21707, 2021.

[2] Subham S. Sahoo, Anselm Paulus, Marin Vlastelica, Vit Musil, Volodymyr Kuleshov, Georg Martius. Backpropagation through Combinatorial Algorithms: Identity with Projection Works. International Conference on Learning Representations (ICLR - 2023), 2023.

## OUR METHOD: Multivariate Learned Adaptive Noise (MuLAN)

### 1 Forward Process



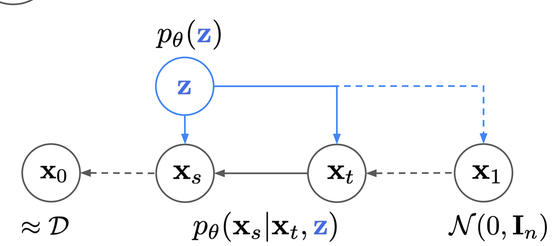

Step 1: $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_0)$: captures high-level information of $\mathbf{x}_0$
Step 2: Forward corruption from $\mathcal{D} \to \mathcal{N}(0, \mathbf{I_n})$ on $\mathbf{z}$:

$$q_\phi(\mathbf{x}_t|\mathbf{x}_s, \mathbf{z}) = \mathcal{N}(\boldsymbol{\alpha}_{t|s}(\mathbf{z}) \odot \mathbf{x}_0, \mathrm{diag}(\boldsymbol{\sigma}_{t|s}(\mathbf{z})))$$

### 2 Noise Schedule

Properties of $\boldsymbol{\alpha}_t(\mathbf{z})$:
- Parameterized using a Neural Network: $\{0, 1\}^m \to [0, 1]^n$
- monotonically increasing (↑).
- $\boldsymbol{\alpha}_t^i(\mathbf{z})$ is a degree 5 polynomial in t for every pixel $i$.
- $\boldsymbol{\alpha}_{t=0}(\mathbf{z}) \approx \{0\}^n$
- $\boldsymbol{\alpha}_{t=1}(\mathbf{z}) \approx \{1\}^n$

### 3 Reverse Process




Step 1: Sample $\mathbf{z} \sim p_\theta(\mathbf{z})$.
Step 2: Denoise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I_n})$ iteratively conditioned on $\mathbf{z}$.

- True reverse posterior:

$$q(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0, \mathbf{z}) = \mathcal{N}\left(\frac{\boldsymbol{\alpha}_{t|s}(\mathbf{z})\boldsymbol{\sigma}_s^2(\mathbf{z})}{\boldsymbol{\sigma}_t^2(\mathbf{z})} \odot \mathbf{x}_t + \frac{\boldsymbol{\sigma}_{t|s}^2(\mathbf{z})\boldsymbol{\alpha}_s(\mathbf{z})}{\boldsymbol{\sigma}_t^2(\mathbf{z})} \odot \mathbf{x}_0, \mathrm{diag}\left(\frac{\boldsymbol{\sigma}_s^2(\mathbf{z})\boldsymbol{\sigma}_{t|s}^2(\mathbf{z})}{\boldsymbol{\sigma}_t^2(\mathbf{z})}\right)\right)$$

- Approximate reverse posterior:

$$p_\theta(\mathbf{x}_s|\mathbf{x}_t, \mathbf{z}) = q_\phi(\mathbf{x}_s|\mathbf{x}_t, \mathbf{x}_0 = \mathbf{x}_\theta(\mathbf{x}_t, \mathbf{z}, t), \mathbf{z})$$

### 4 Auxiliary Latent

- $\mathbf{z}$ : $k$-hot vectors of length $m$.
- More stable training with discrete latents vs Gaussian latents.
- Used Identity [2] to differentiate through the sampling process:

$$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_0)$$

- $p_\theta(\mathbf{z})$ Uniform distribution over all possible $k$-hot vectors.
- For our experiments, we chose $k = 15$ and $m = 50$.

### 5 Training Objective

Negative Evidence Lower Bound (NELBO):

$$\mathbb{E}_{q,t}\Bigg[\underbrace{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_\theta(\mathbf{x}_t, \mathbf{z}, t))^2 \cdot \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\nu}(\mathbf{z}, t)}_{\mathcal{L}_{\text{diff}}} + \underbrace{\mathrm{D}_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p_\theta(\mathbf{z}))}_{\mathcal{L}_{\text{aux. latent prior}}} + \underbrace{\mathrm{D}_{\mathrm{KL}}(q_\phi(\mathbf{x}_1|\mathbf{x}_0)\|p_\theta(\mathbf{x}_1))}_{\mathcal{L}_{\text{Gaussian prior}}}\Bigg]$$

- Features an extra loss term $\mathcal{L}_{\text{aux. latent prior}}$.
- $\mathcal{L}_{\text{diff}}$ is lowered due to the introduction of $\mathbf{z}$.

**Net Impact: Lower overall NELBO**