

# A Geometric Perspective on Diffusion Models

Yuanzhi Zhu

# Content

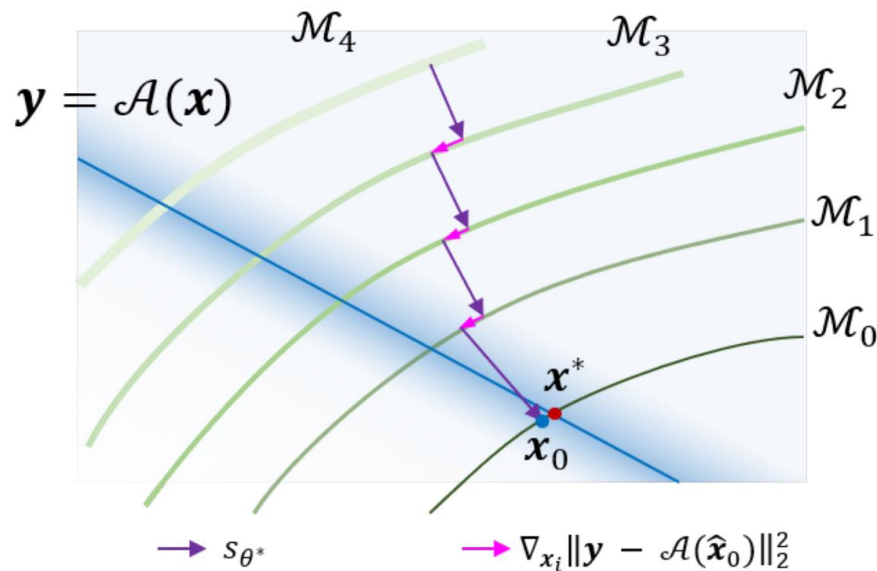
---

- Motivation
- Geometric Perspective
- Conclusion

# A Geometric Perspective on Diffusion Models\*

**(my) Motivation:** Not related to non-Euclidean geometry

1. How to understand this graph?



# A Geometric Perspective on Diffusion Models

**(my) Motivation:** Not related to MOLECULAR generation

2. What's the difference between the following two algorithm?

---

## Algorithm 1 DiffPIR

---

**Require:**  $s_\theta, T, \mathbf{y}, \sigma_n, \{\bar{\sigma}_t\}_{t=1}^T, \zeta, \lambda$

- 1: Initialize  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , pre-calculate  $\rho_t \triangleq \lambda \sigma_n^2 / \bar{\sigma}_t^2$ .
  - 2: **for**  $t = T$  **to** 1 **do**
  - 3:    $\mathbf{x}_0^{(t)} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1 - \bar{\alpha}_t)s_\theta(\mathbf{x}_t, t))$  // Predict  $\hat{\mathbf{z}}_0$  with score model as denoisor
  - 4:    $\hat{\mathbf{x}}_0^{(t)} = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathcal{H}(\mathbf{x})\|^2 + \rho_t \|\mathbf{x} - \mathbf{x}_0^{(t)}\|^2$  // Solving data proximal subproblem
  - 5:    $\hat{\epsilon} = \frac{1}{\sqrt{1 - \bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \hat{\mathbf{x}}_0^{(t)})$  // Calculate effective  $\hat{\epsilon}(\mathbf{x}_t, \mathbf{y})$
  - 6:    $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 7:    $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}_0^{(t)} + \sqrt{1 - \bar{\alpha}_{t-1}}(\sqrt{1 - \zeta} \hat{\epsilon} + \sqrt{\zeta} \epsilon_t)$  // Finish one step reverse diffusion sampling
  - 8: **end for**
  - 9: **return**  $\mathbf{x}_0$
- 

---

## Algorithm 2 Extended Sampling I: DPS $y_t$

---

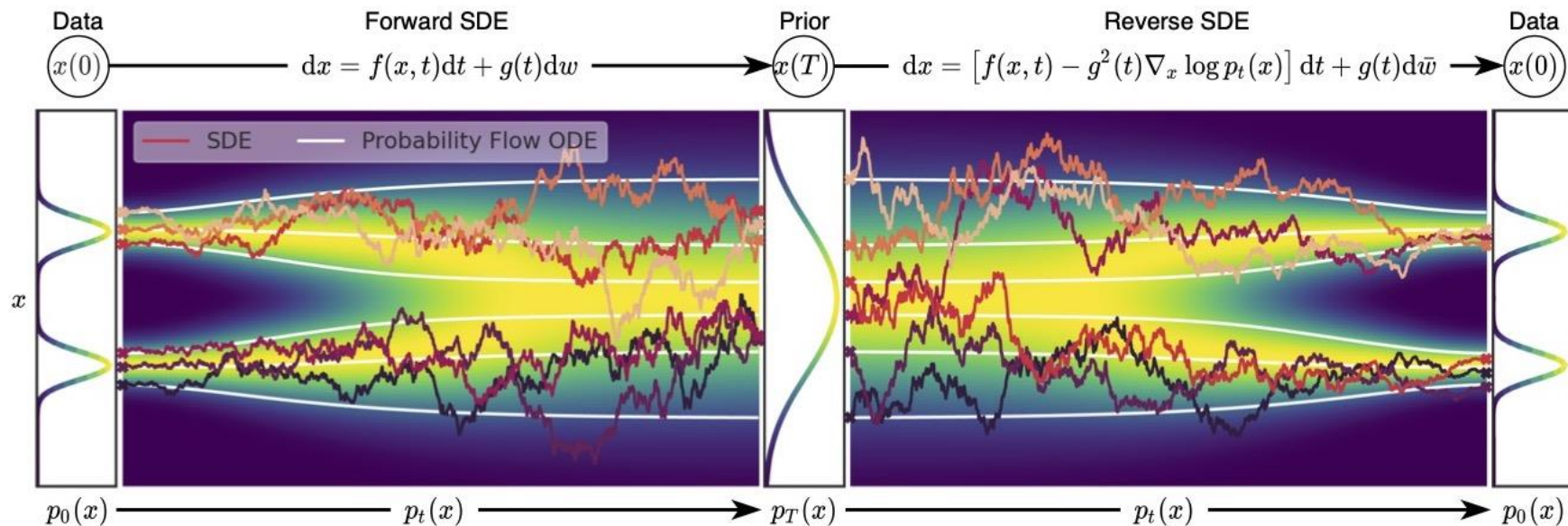
**Require:**  $s_\theta, T, \mathbf{y}, \sigma_n, \{\sigma_t\}_{t=1}^T, \lambda$

- 1: Initialize  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 2: **for**  $t = T$  **to** 1 **do**
  - 3:    $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 4:    $\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sqrt{\beta_t} \epsilon_t$  // one step reverse diffusion sampling
  - 5:    $\mathbf{x}_{t-1} = \mathbf{z}_{t-1} - \frac{\sigma_t^2}{2\lambda\sigma_n^2} \nabla_{\mathbf{z}_{t-1}} \|\mathbf{y}_{t-1} - \mathcal{H}(\mathbf{z}_{t-1})\|^2$  // Solving data proximal subproblem
  - 6: **end for**
  - 7: **return**  $\mathbf{x}_0$
-

# A Geometric Perspective on Diffusion Models

**(my) Motivation:** Use only VE-ODE for example

3. How to understand the diffusion trajectory better?

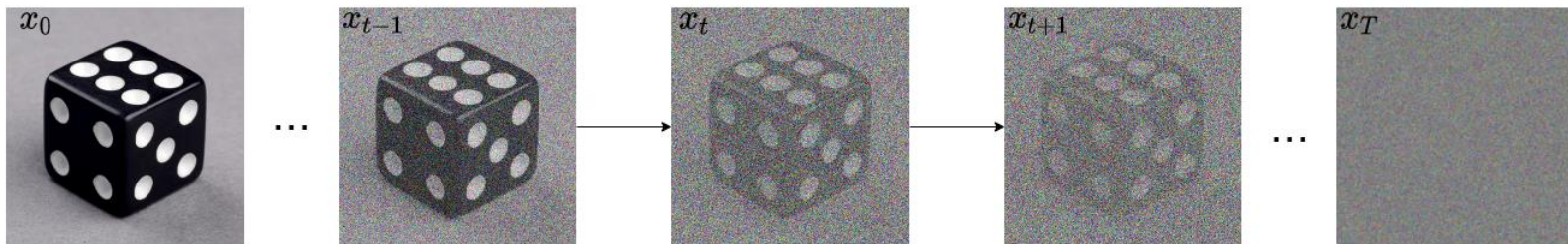


# A Geometric Perspective on Diffusion Models

**(my) Motivation:** Understanding through experiment observation

4. Where does the generative power come from?

$$x_0 \sim q(x_0) \quad \longrightarrow \quad q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad \longrightarrow \quad q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1})$$



$$p_\theta(x_0) = \int p(x_T) \prod_{i=1}^T p_\theta(x_{t-1}|x_t) dx_{1:T} \quad \longleftarrow \quad p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad \longleftarrow \quad x_T \sim \mathcal{N}(0, I)$$

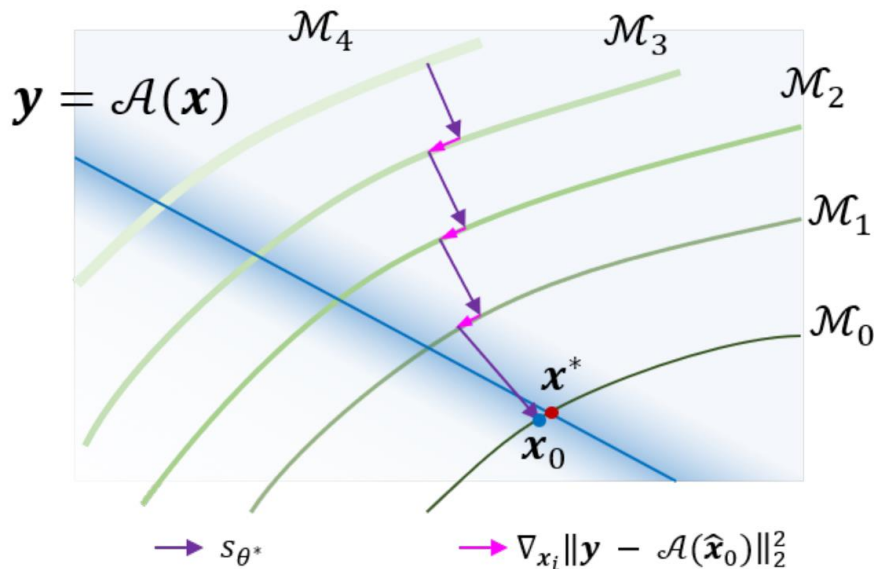
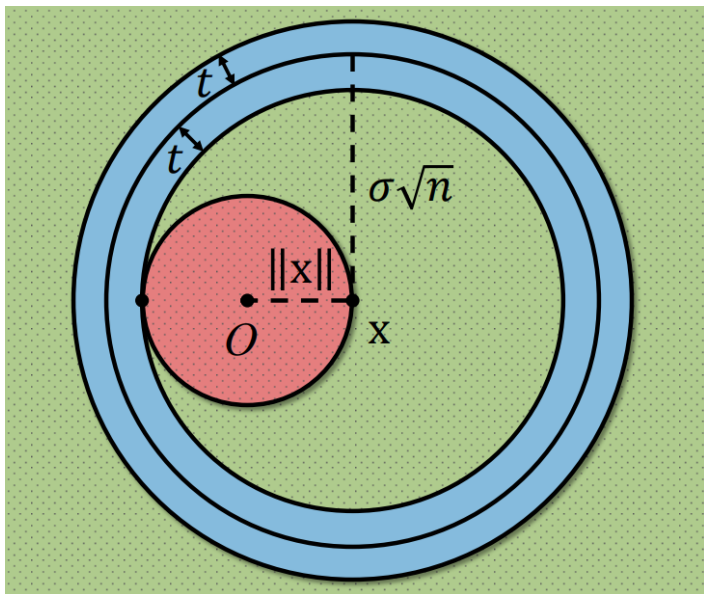
# Content

---

- Motivation
- Geometric Perspective
- Conclusion

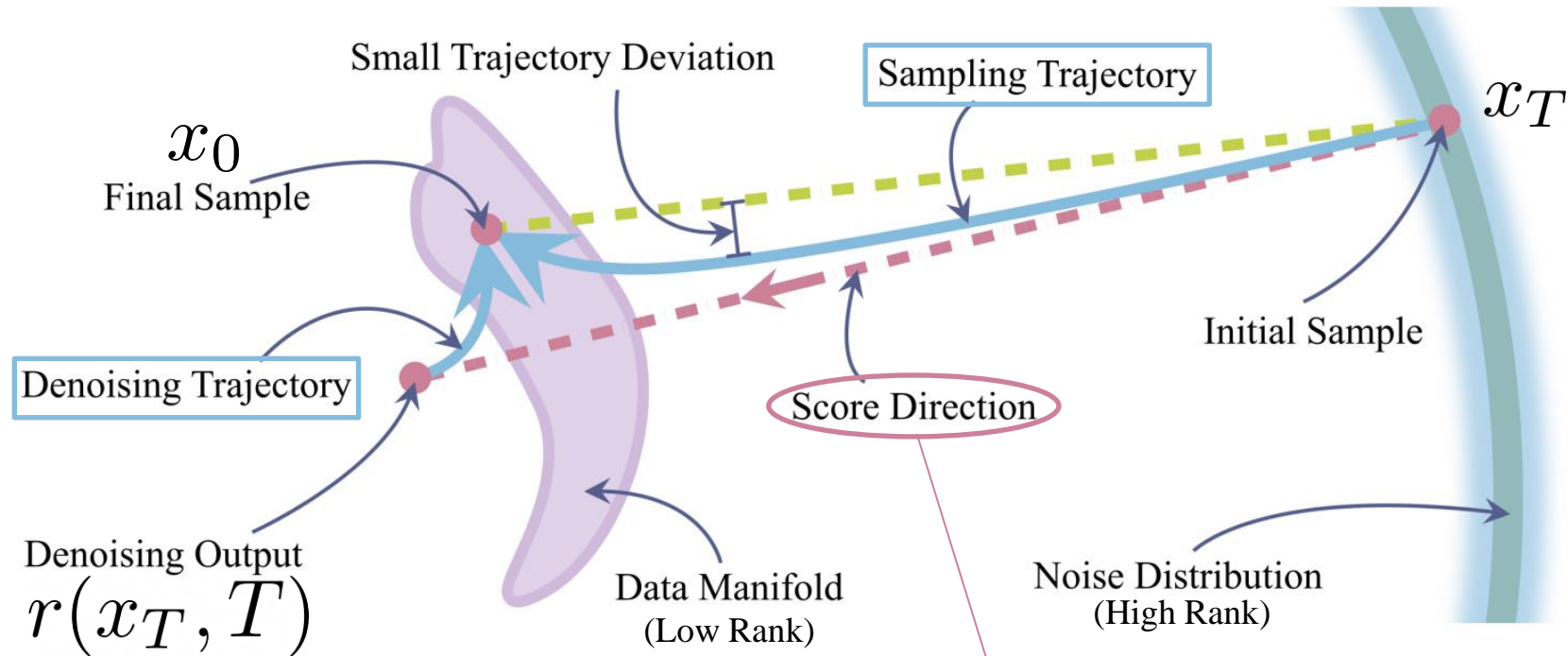
# Visualization of High Dimensional Trajectory

**Proposition 1.** Given a high-dimensional vector  $\mathbf{x} \in \mathbb{R}^d$  and an isotropic Gaussian noise  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \sigma^2 \mathbf{I}_d)$ ,  $\sigma > 0$ , we have  $\mathbb{E} \|\mathbf{z}\|^2 = \sigma^2 d$ , and with high probability,  $\mathbf{z}$  stays within a “thin shell”:  
 $\|\mathbf{z}\| = \sigma\sqrt{d} \pm O(1)$ . Additionally,  $\mathbb{E} [\underbrace{\|\mathbf{x} + \mathbf{z}\|^2 - \|\mathbf{x}\|^2}_{\text{Perpendicular}}] = \sigma^2 d$ ,  $\lim_{d \rightarrow \infty} \mathbb{P}(\|\mathbf{x} + \mathbf{z}\| > \|\mathbf{x}\|) = 1$ .





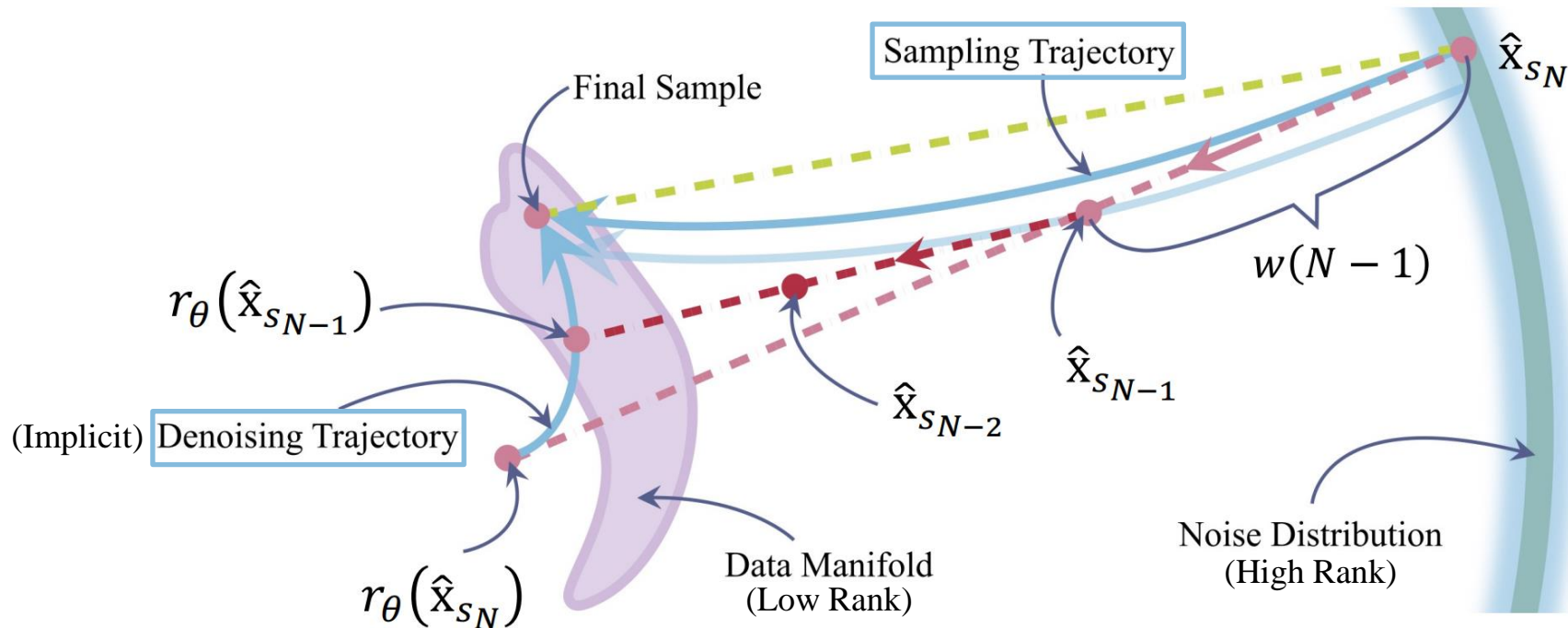
# Visualization of High Dimensional Trajectory



$$\frac{dx}{dt} = -\dot{\sigma}(t)\sigma(t)\nabla \log p(x, \sigma(t)) = -\dot{\sigma}(t) \frac{r(x, \sigma(t)) - x}{\sigma(t)} \xrightarrow{\sigma(t)=t} -\frac{r(x, t) - x}{t}$$

# Visualization of High Dimensional Trajectory

1. Straightness of the trajectories
2. Properties of denoising trajectory



# Experiments on High Dimensional Trajectory

## Notations:

*sampling trajectory* sequence  $\{\hat{\mathbf{x}}_s\}_{s_N}^{s_0}$  (reverse diffusion with trained model)

*optimal sampling* sequence  $\{\hat{\mathbf{x}}_s^*\}_{s_N}^{s_0}$  (trajectory of image from dataset)

$\ell_2$  distance  $d(\cdot, \cdot)$

*trajectory deviation*  $d(\hat{\mathbf{x}}_s, [\hat{\mathbf{x}}_{s_0}, \hat{\mathbf{x}}_{s_N}])$  (straightness)

*denoising trajectory* sequence  $\{r_\theta(\hat{\mathbf{x}}_s, s)\}_{s_N}^{s_1}$

*optimal denoiser*

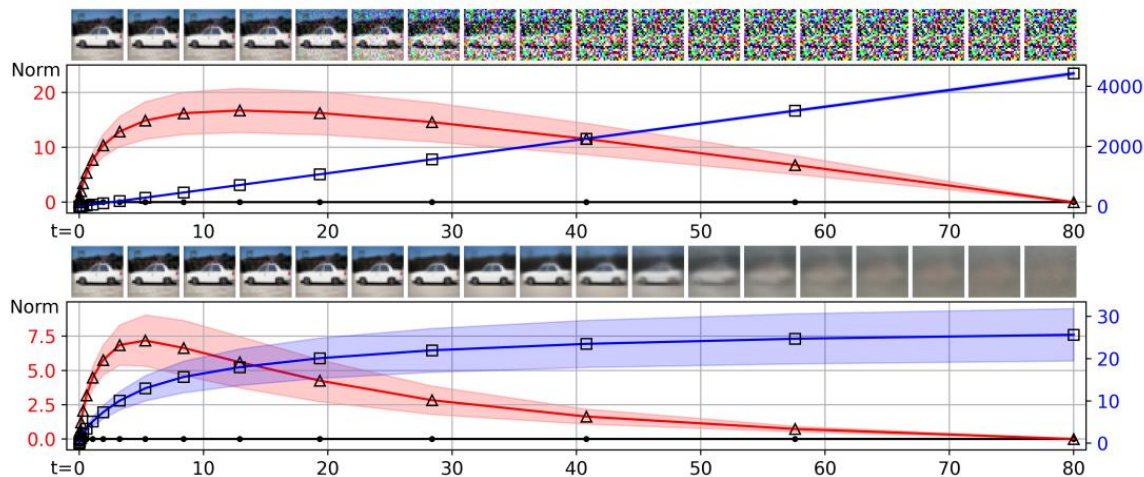
$$r_{\theta}^*(\hat{\mathbf{x}}; \sigma_t) = \sum_i u_i \mathbf{x}_i = \sum_i \frac{\exp(-\|\hat{\mathbf{x}} - \mathbf{x}_i\|^2 / 2\sigma_t^2)}{\sum_j \exp(-\|\hat{\mathbf{x}} - \mathbf{x}_j\|^2 / 2\sigma_t^2)} \mathbf{x}_i, \quad \sum_i u_i = 1.$$

$(\cdot)^*$  optimal/theoretical variable

# Experiments on High Dimensional Trajectory

**Observation 1.** The sampling trajectory is almost straight while the denoising trajectory is bent.

**Observation 2.** The generated samples on the sampling trajectory and denoising trajectory both move monotonically from the initial points toward their converged points in expectation, i.e.,  $\{\mathbb{E}[d(\hat{\mathbf{x}}_s, \hat{\mathbf{x}}_{s_0})]\}_{s_N}^{s_0}$  and  $\{\mathbb{E}[d(r_\theta(\hat{\mathbf{x}}_s), r_\theta(\hat{\mathbf{x}}_{s_1}))]\}_{s_N}^{s_1}$  are monotone decreasing sequences.



curvature of sampling trajectory  
 $16/4428 \approx 0.0036$

curvature of denoising trajectory  
 $7/26 \approx 0.27$

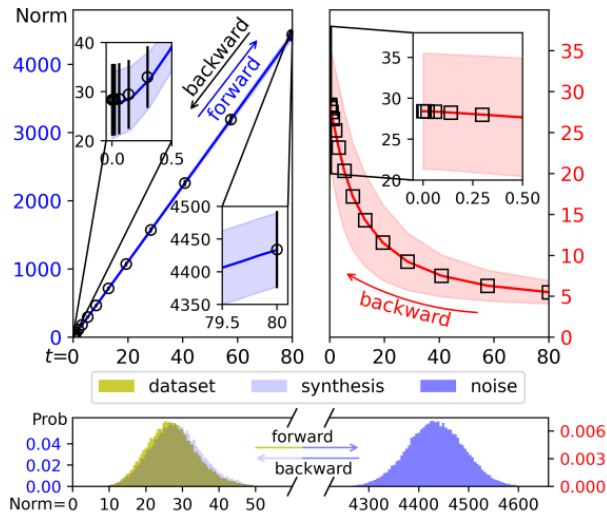
(b) Deviation in the sampling (top)/denoising (bottom) trajectories.

trajectory deviation  $\ell_2$  distance

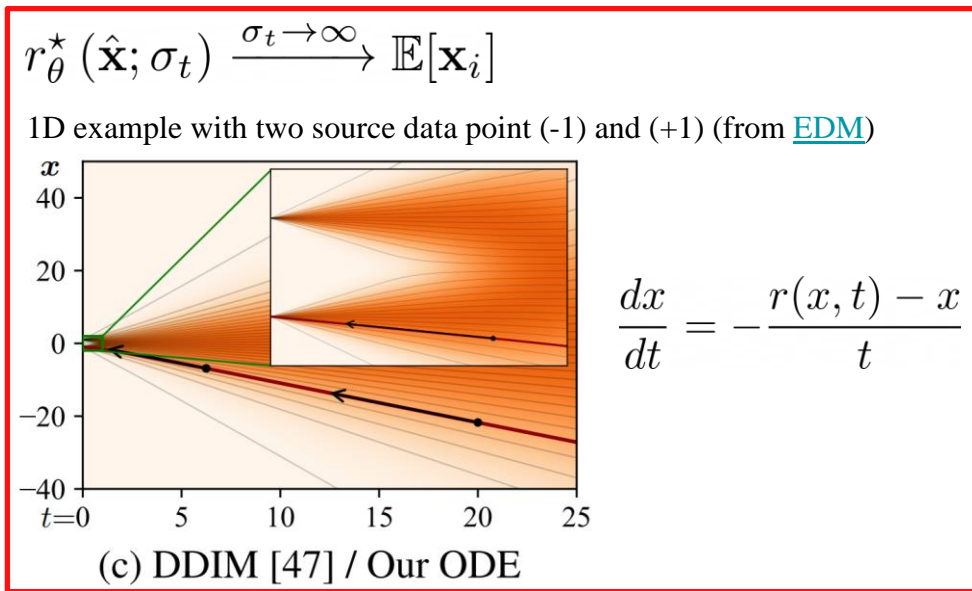
# Experiments on High Dimensional Trajectory

**Observation 3.** The sampling trajectory converges to the data distribution in a monotone magnitude shrinking way. Conversely, the denoising trajectory converges to the data distribution in a monotone magnitude expanding way. Formally, we have  $\{\mathbb{E}\|\hat{\mathbf{x}}_s\|\}_{s_N}^{s_0} \downarrow$  and  $\{\mathbb{E}\|r_\theta(\hat{\mathbf{x}}_s)\|\}_{s_N}^{s_1} \uparrow$

optimal denoiser:  $r_\theta^*(\hat{\mathbf{x}}; \sigma_t) = \sum_i u_i \mathbf{x}_i = \sum_i \frac{\exp(-\|\hat{\mathbf{x}} - \mathbf{x}_i\|^2 / 2\sigma_t^2)}{\sum_j \exp(-\|\hat{\mathbf{x}} - \mathbf{x}_j\|^2 / 2\sigma_t^2)} \mathbf{x}_i, \quad \sum_i u_i = 1.$



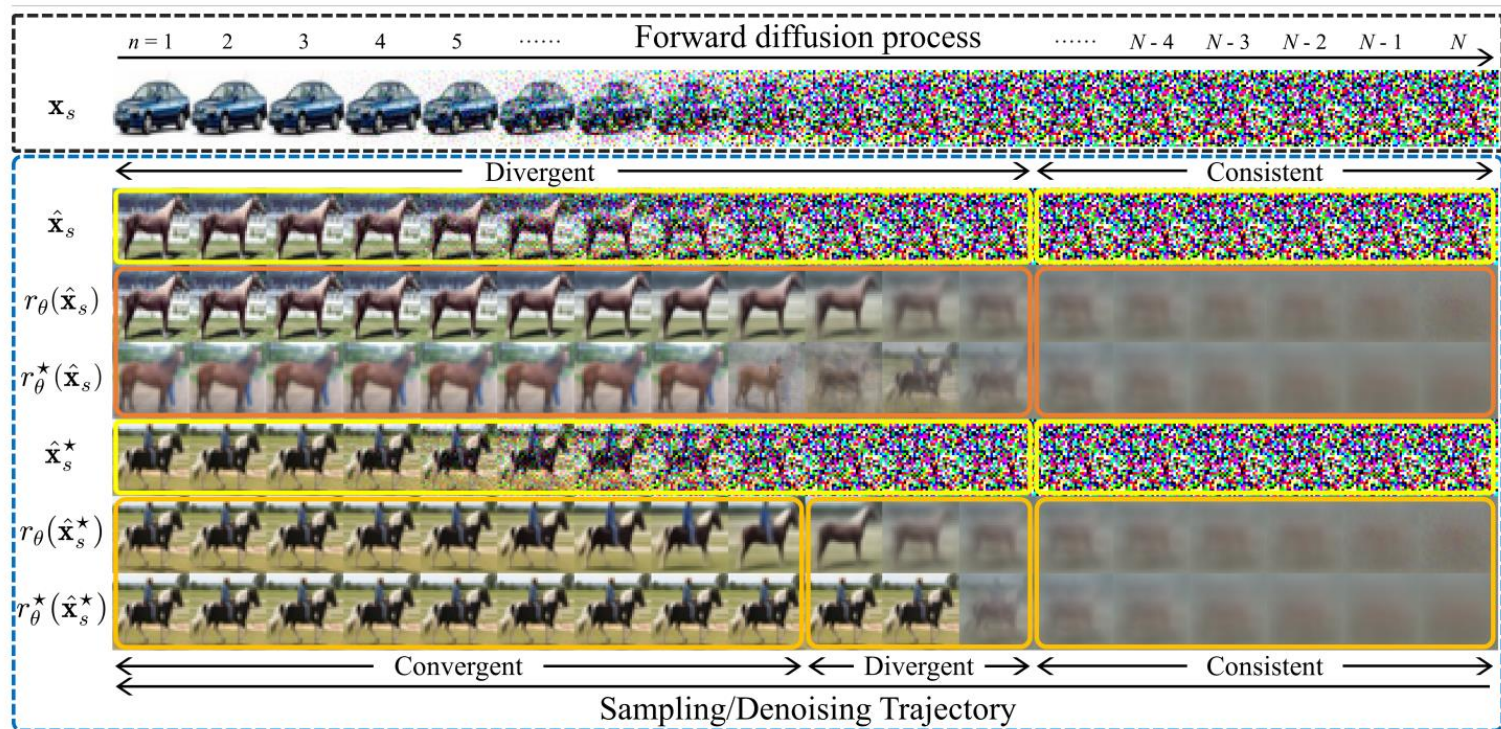
(a) The statistics of magnitude.  $\|\mathbf{x}\|$



(c) DDIM [47] / Our ODE



# Experiments on High Dimensional Trajectory



# Experiments on High Dimensional Trajectory

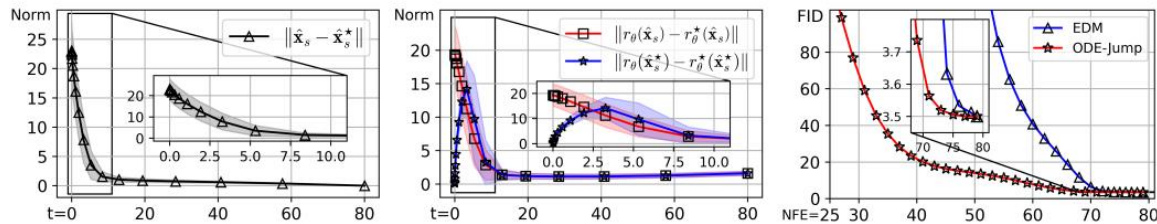
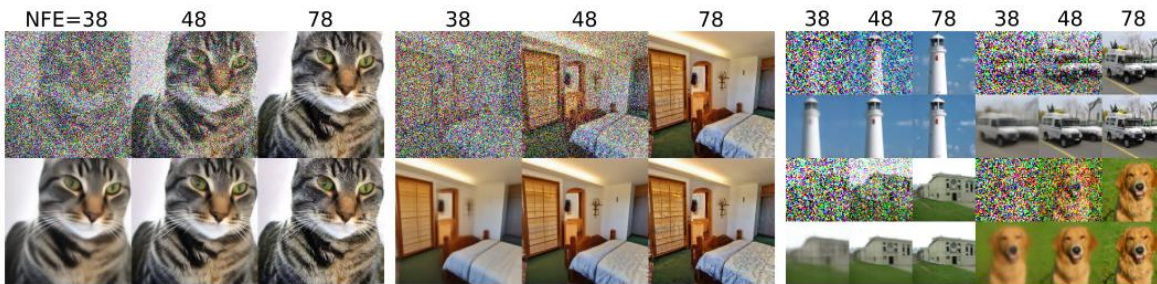


Figure 4: The score deviation in expectation (left and middle) and FID with different NFEs (right).

**Observation 4.** *The learned score is well-matched to the optimal score in the large-noise region (from 80 to around 10), otherwise they may diverge or almost coincide depending on different regions*



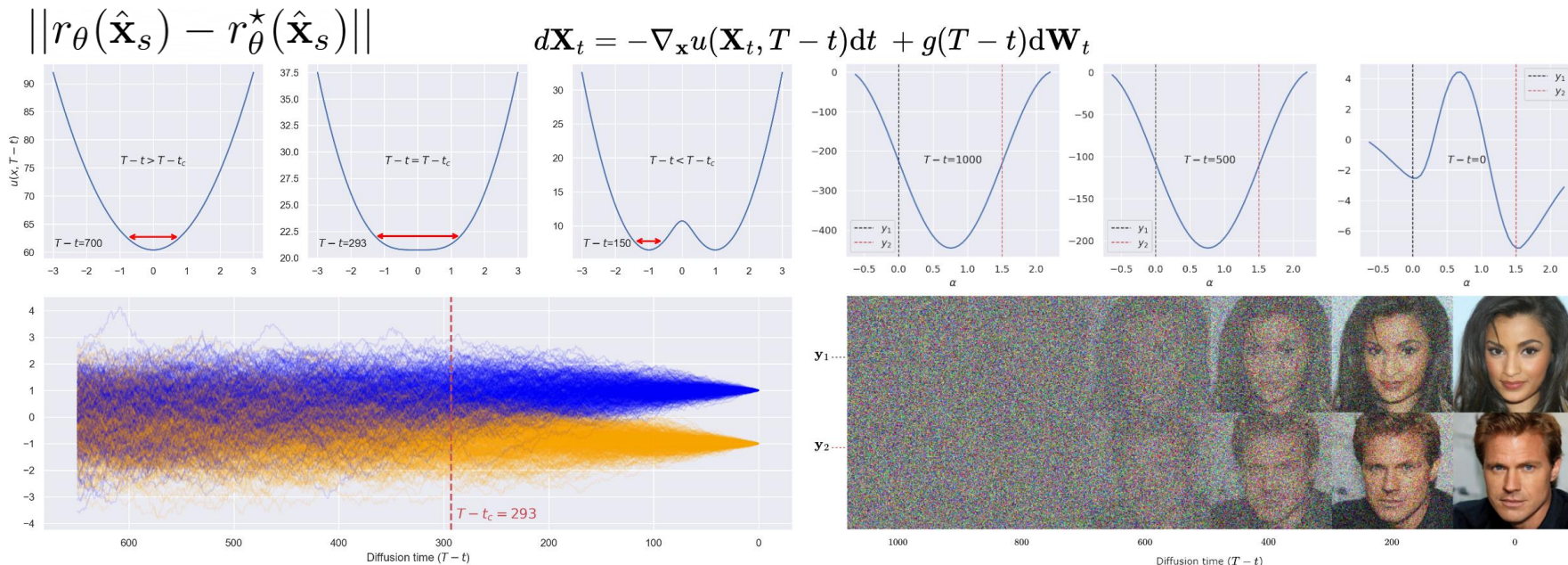
**Observation 5.** *The (optimal) denoising trajectory converges faster than the (optimal) sampling trajectory in terms of visual quality.*

Figure 5: The synthesized images of our proposed ODE-Jump sampling (bottom) converge much faster than that of EDMs [KAAL22] (top) in terms of visual quality.

“In fact, our learned score has to moderately diverge from the optimum to guarantee the generative ability.”



# Spontaneous Symmetry Breaking in Generative Diffusion Models\*



(a) Symmetry breaking in 1D diffusion model

(b) Symmetry breaking in CelebA HQ 256x256

$$u(\mathbf{x}, s) = -g^2(s) \log p(\mathbf{x}, s) + \int_0^{\mathbf{x}} f(\mathbf{z}, s) \cdot d\mathbf{z}$$

$$u(x, t) = \beta(T-t) \left( -\frac{1}{4}x^2 - \log \left( e^{-\frac{(x-\theta_{T-t})^2}{2(1-\theta_{T-t}^2)}} + e^{-\frac{(x+\theta_{T-t})^2}{2(1-\theta_{T-t}^2)}} \right) \right)$$

$\theta_{T-t}$  is a monotonic function of  $t$  ranging from 0 to 1

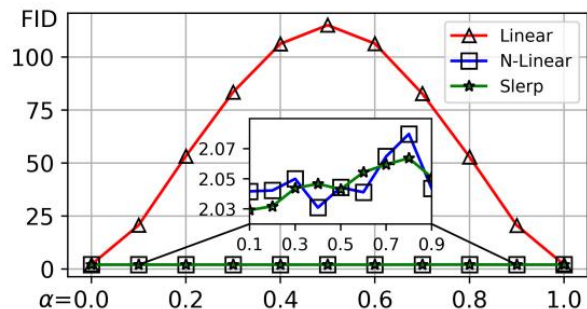
[\\*\[2305.19693\] Spontaneous symmetry breaking in generative diffusion models \(arxiv.org\)](#)

[Gabriel Raya | Spontaneous symmetry breaking in generative diffusion models](#)

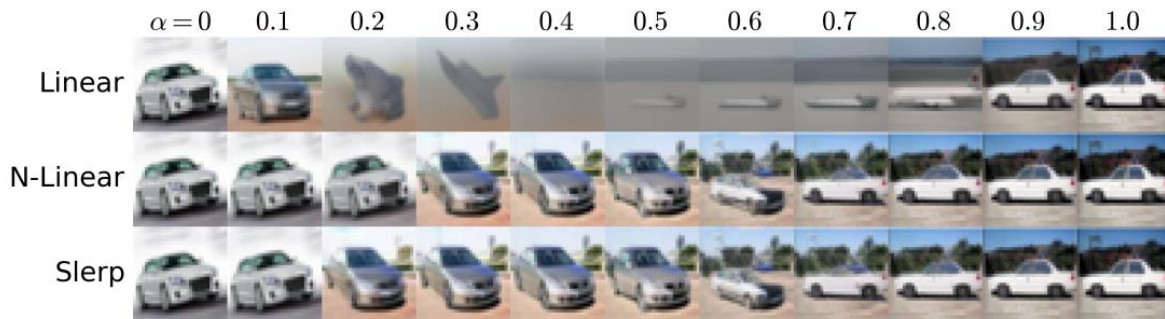


# In-Distribution Latent Interpolation

**Proposition 5.** *In high dimensions, linear interpolation [HJA20] shifts the latent distribution while spherical linear interpolation [SME21] asymptotically ( $d \rightarrow \infty$ ) maintains the latent distribution.*



(a) The comparison of FID.



(b) Visualization of latent interpolation with different strategies.

Figure 6: Linear latent interpolation results in blurry images, while a simple re-scaling trick greatly preserves the fine-grained image details and enables a smooth traversal among different modes.

# Rethinking Distillation-Based Fast Sampling Techniques

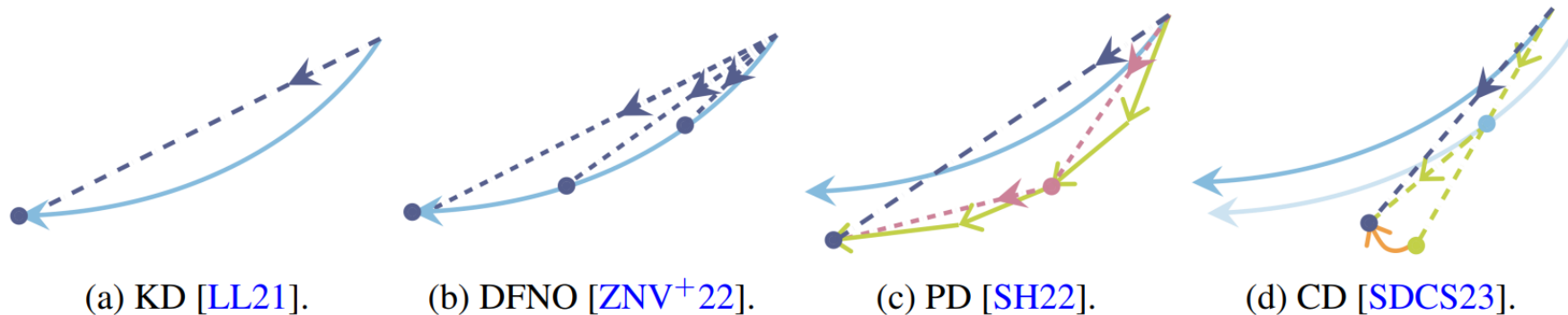


Figure 7: The comparison of distillation-based techniques. The *offline* techniques first simulate a long ODE trajectory with the teacher score and then make the student score points to the final point (KD [LL21]) or also include intermediate points on the trajectory (DFNO [ZNV<sup>+</sup>22]). The *online* techniques iteratively fine-tune the student prediction to align with the target simulated by a few-step teacher model along the sampling trajectory (PD [SH22]) or the denoising trajectory (CD [SDCS23]).

# Content

---

- Motivation
- Geometric Perspective
- Conclusion

# Conclusion

---

- ✓ Geometric perspective on (VE) diffusion models
- ✗ Theoretical results do not entirely substantiate the observations