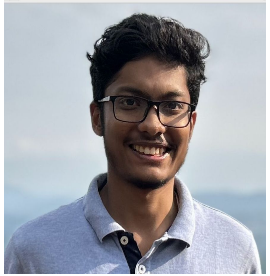
Understanding Hallucinations in Diffusion Models through Mode Interpolation


Sumukh Aithal


Pratyush Maini


Zachary Lipton


Zico Kolter

Carnegie Mellon University

NeurIPS 2024

Diffusion Models generate strange artifacts

Hands with extra (or missing) fingers are commonly seen in generated images.

Toy Experiment

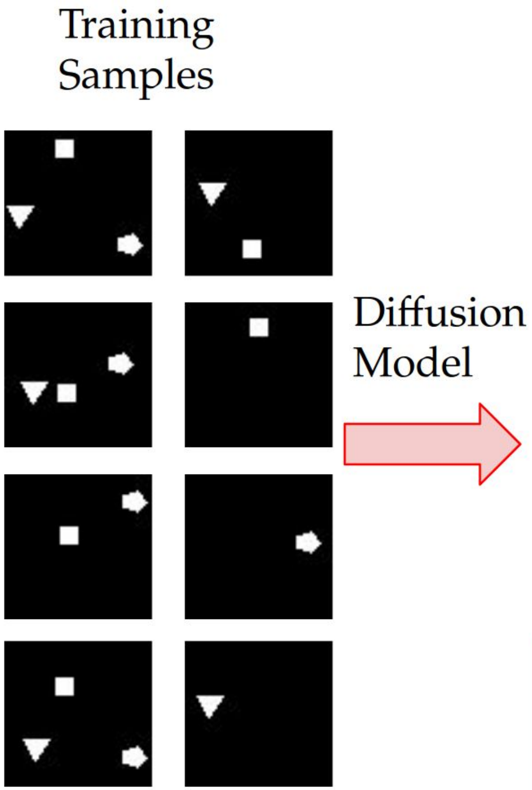
Let's start with a simple toy experiment

Dataset of 3 shapes:

1. Triangle
2. Square
3. Pentagon

All 64x64 grayscale images

Atmost one occurrence of each shape.



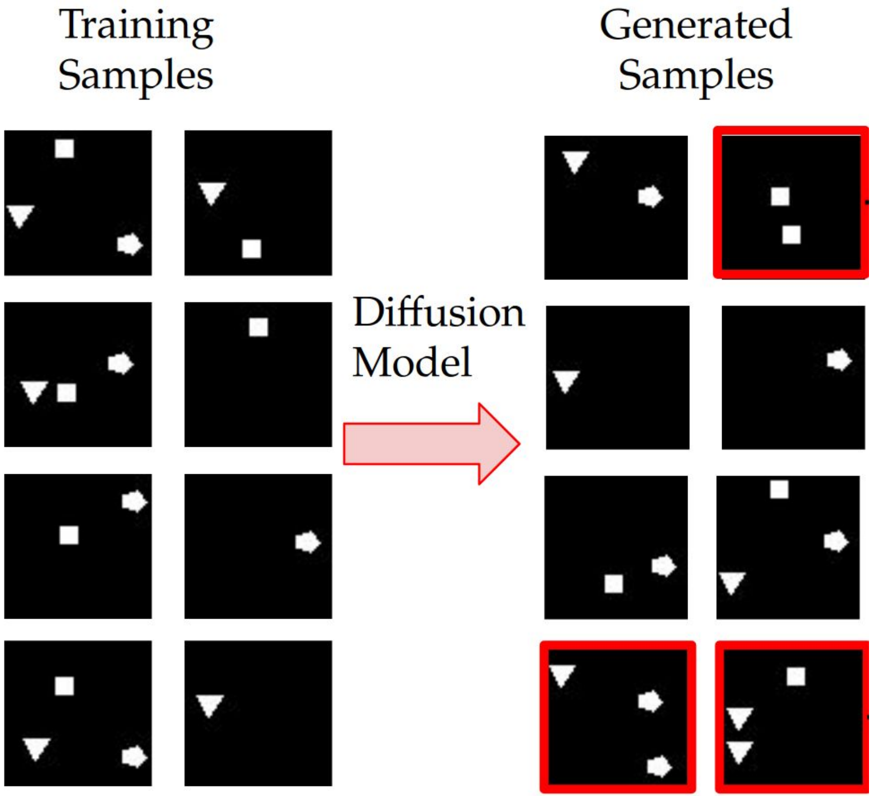
Let’s start with a simple toy experiment

Dataset of 3 shapes:

1. Triangle
2. Square
3. Pentagon

All 64x64 grayscale images

Atmost one occurrence of each shape.



Let’s start with a simple toy experiment

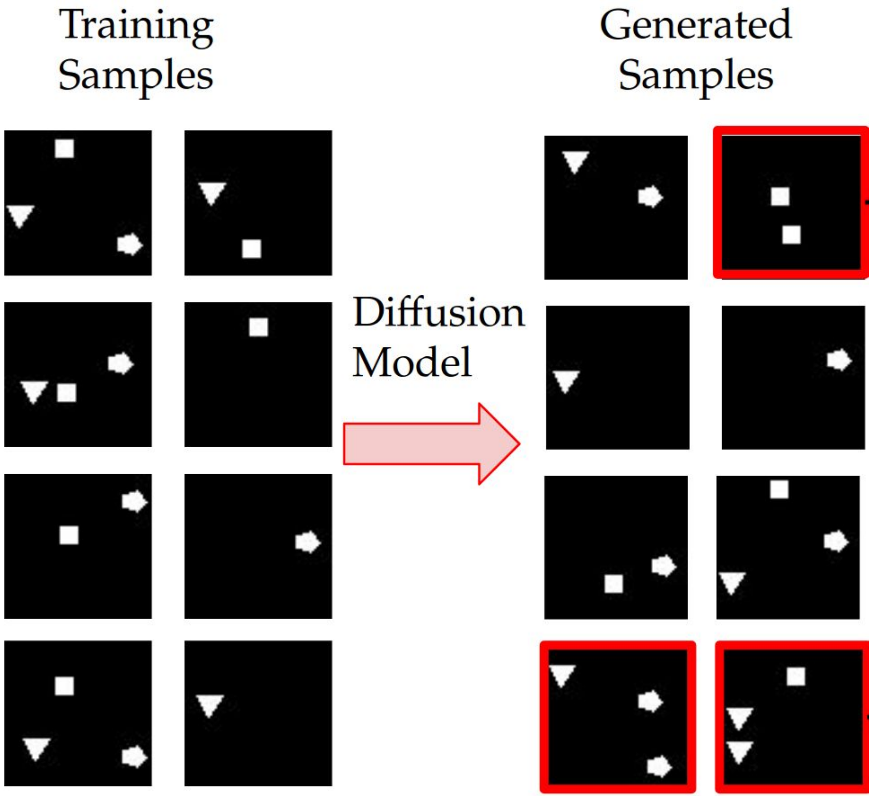
Dataset of 3 shapes:

1. Triangle
2. Square
3. Pentagon

All 64x64 grayscale images

Atmost one occurrence of each shape.

Surprising?



What are Diffusion Models?

What are Diffusion Models?

Forward Process (Data to Noise): Perturbing an image with multiple scales of Gaussian noise.

$\mathbf{x}_0$ —

What are Diffusion Models?

Forward Process (Data to Noise): Perturbing an image with multiple scales of Gaussian noise.

$\mathbf{x}_0$

What are Diffusion Models?

Forward Process (Data to Noise): Perturbing an image with multiple scales of Gaussian noise.

$\mathbf{x}_0$ $\longrightarrow$ $\mathbf{x}_T$

What are Diffusion Models?

Forward Process (Data to Noise): Perturbing an image with multiple scales of Gaussian noise.

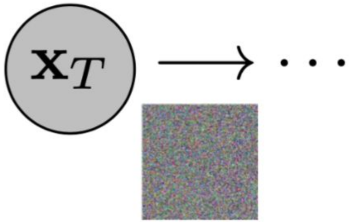
$\mathbf{x}_0 \longrightarrow \mathbf{x}_T$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

What are Diffusion Models?

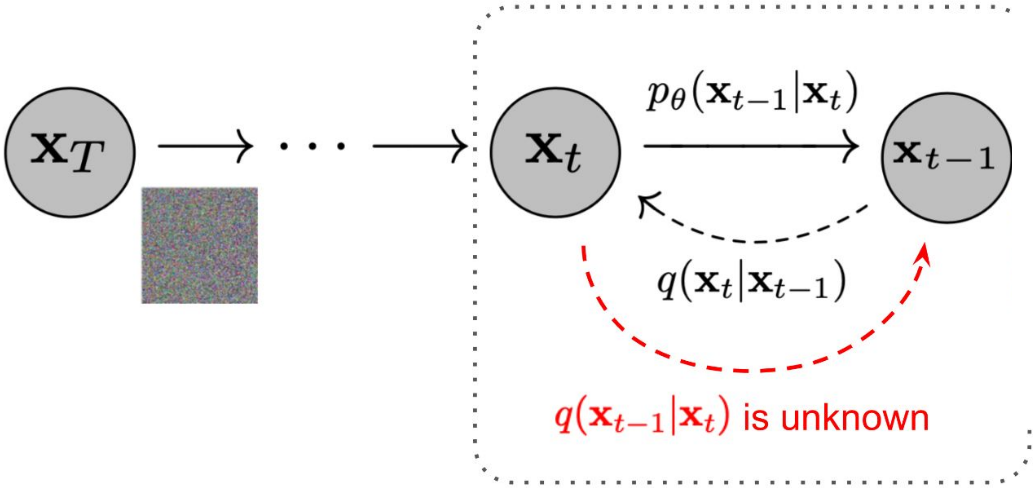
Reverse Process (Noise to Data): Predict the noise added in the previous timestep



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

What are Diffusion Models?

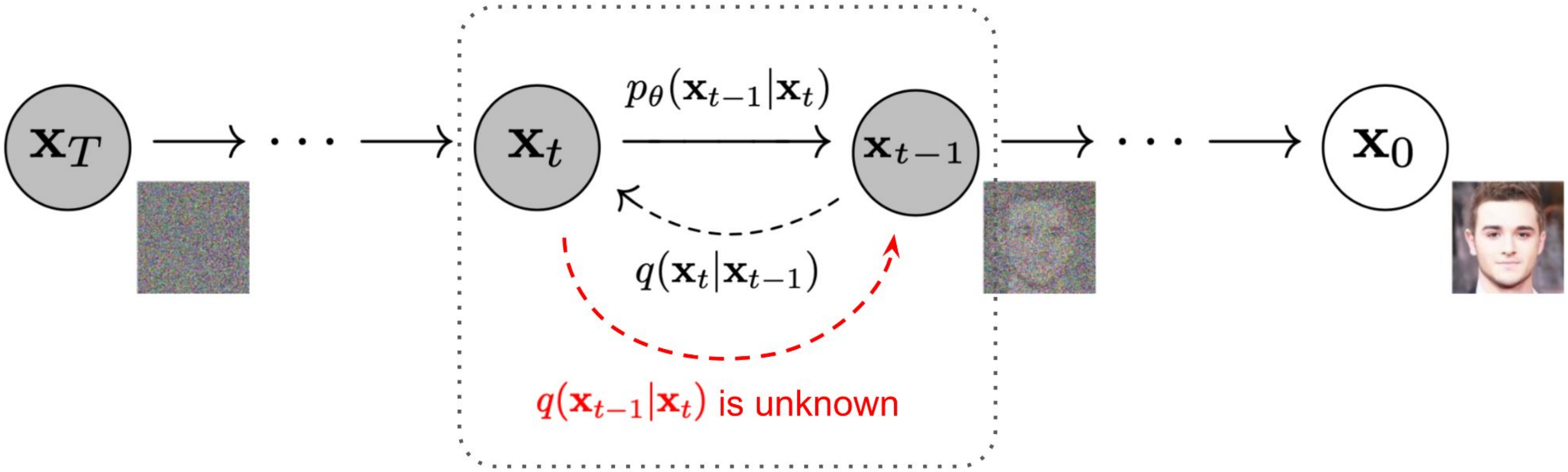
Reverse Process (Noise to Data): Predict the noise added in the previous timestep



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

What are Diffusion Models?

Reverse Process (Noise to Data): Predict the noise added in the previous timestep

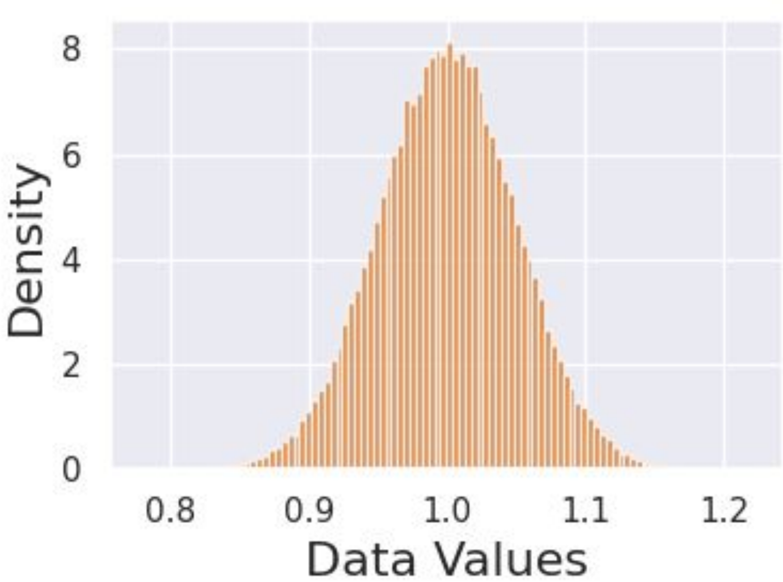


https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Mode Interpolation

Mode Interpolation: 1D Gaussian

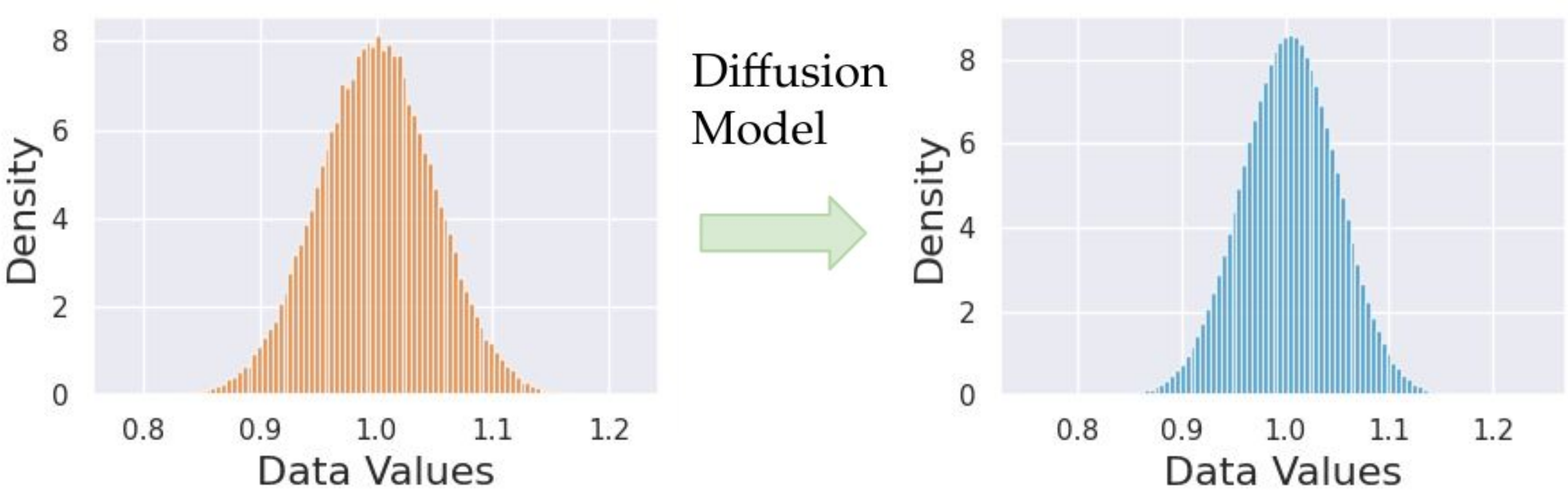
Let's start with a simple 1D Gaussian with mean 1.



Diffusion Model

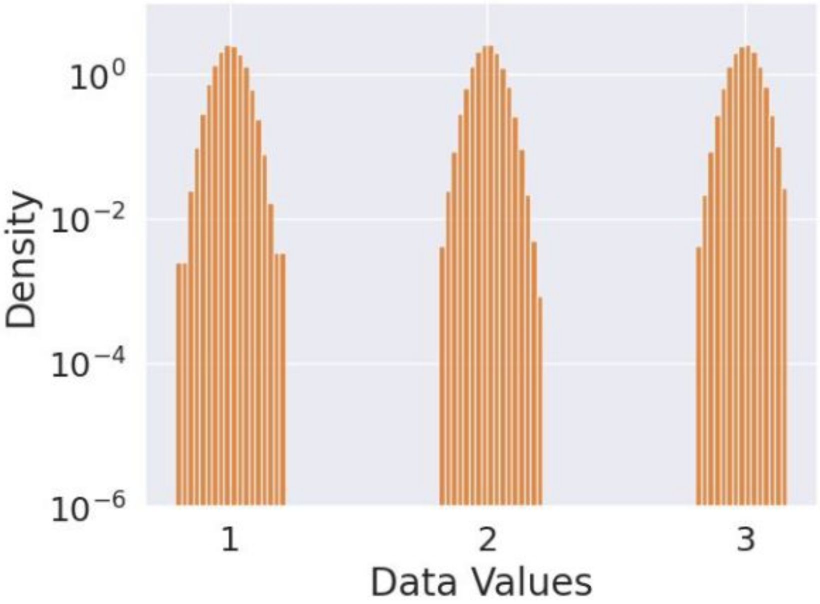
Mode Interpolation: 1D Gaussian

Let's start with a simple 1D Gaussian with mean 1.



Mode Interpolation: 1D Gaussian

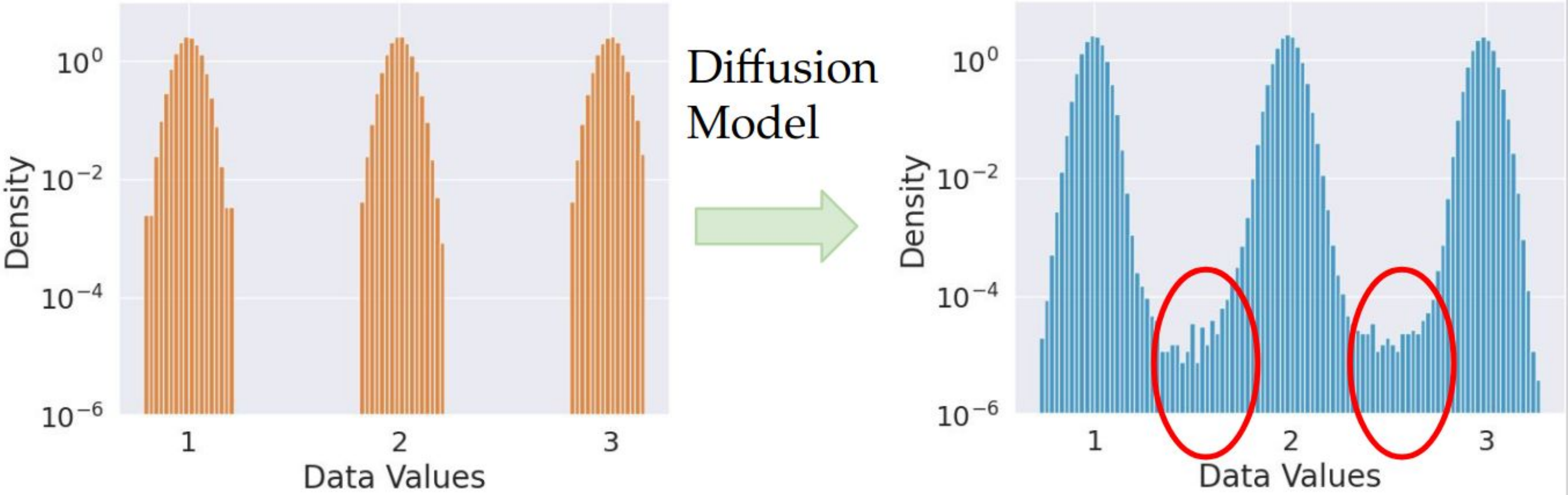
Consider a simple mixture of 1D Gaussians: $p(x) = \frac{1}{3}\mathcal{N}(\mu_1, \sigma^2) + \frac{1}{3}\mathcal{N}(\mu_2, \sigma^2) + \frac{1}{3}\mathcal{N}(\mu_3, \sigma^2)$



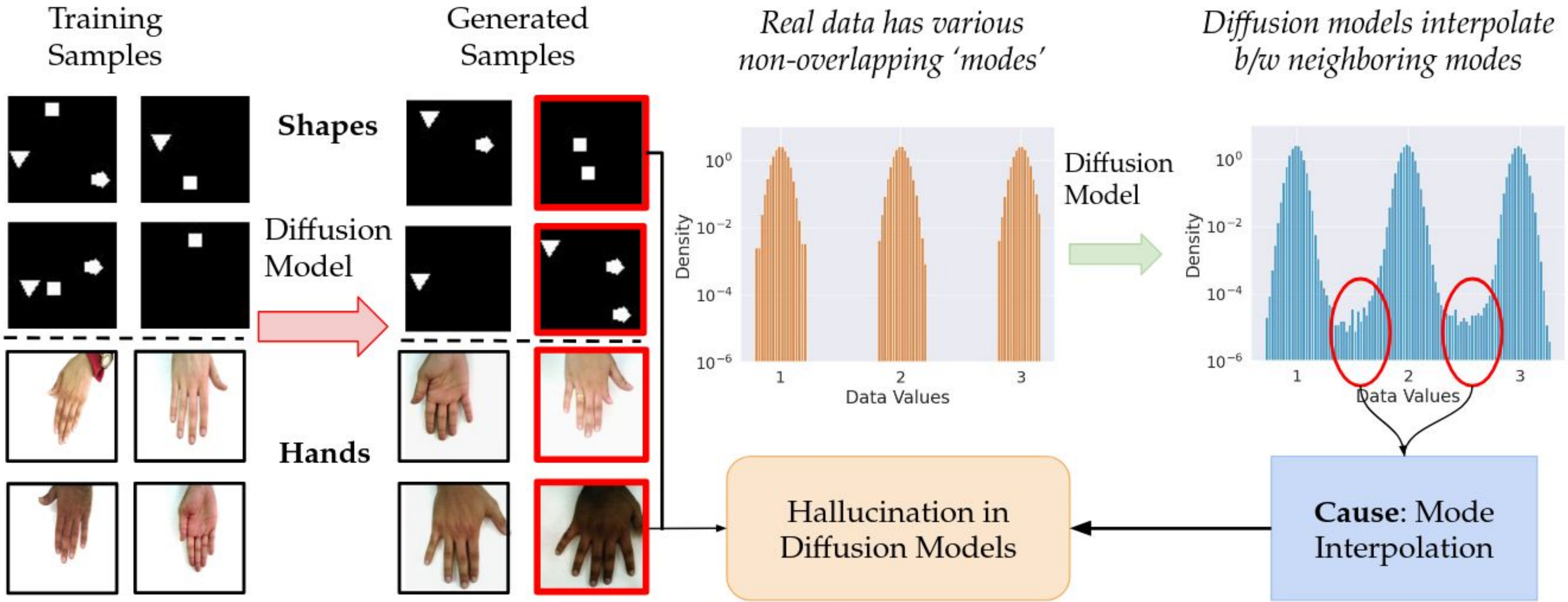
Diffusion Model

Mode Interpolation: 1D Gaussian

Consider a simple mixture of 1D Gaussians: $p(x) = \frac{1}{3}\mathcal{N}(\mu_1, \sigma^2) + \frac{1}{3}\mathcal{N}(\mu_2, \sigma^2) + \frac{1}{3}\mathcal{N}(\mu_3, \sigma^2)$



What is Mode Interpolation?

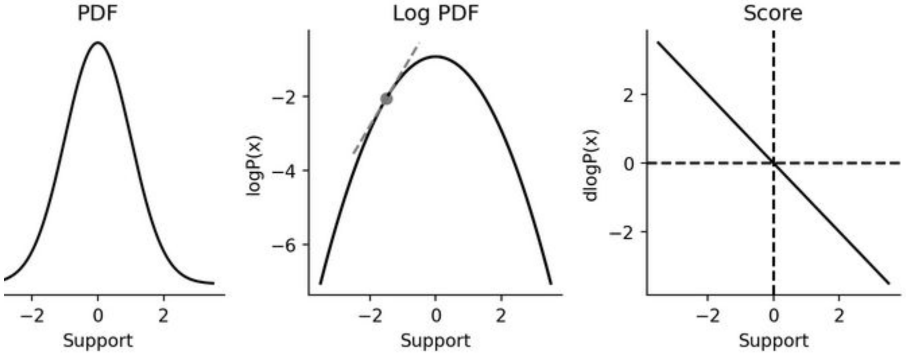


What causes Mode Interpolation?

What causes mode interpolation?

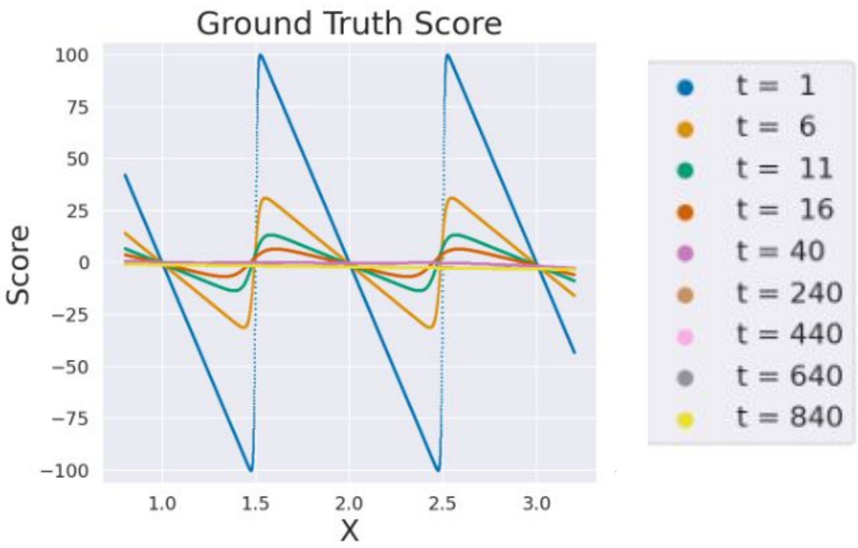
Diffusion models are score-based generative models

Score Function = $\nabla_{\mathbf{x}} \log q(\mathbf{x})$



What causes mode interpolation?

Diffusion models are score-based generative models



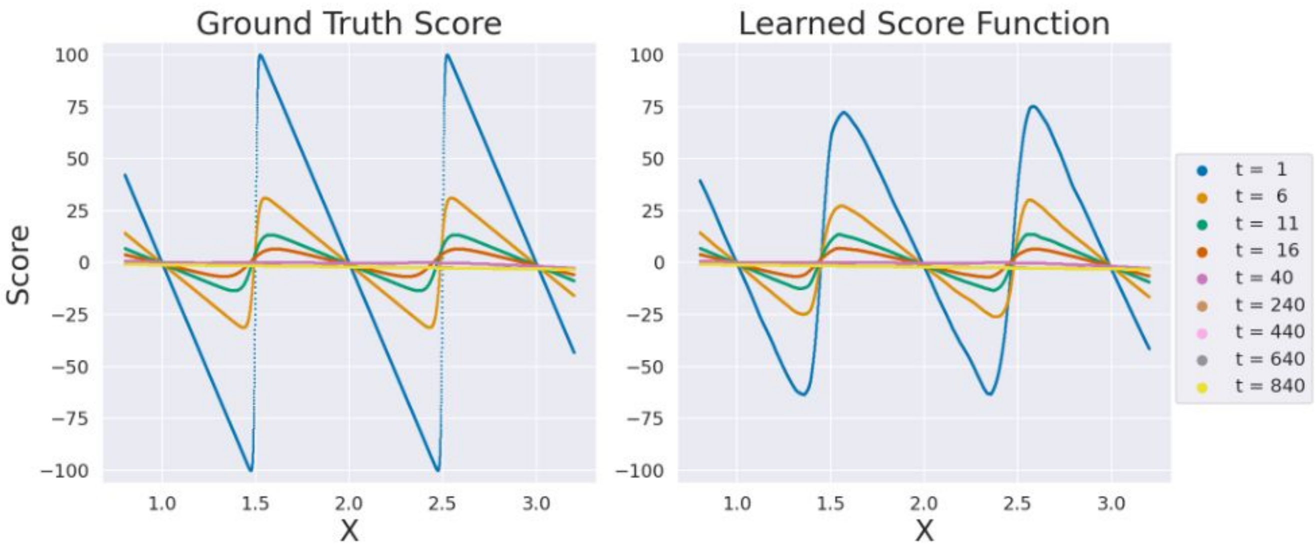
Score Function = $\nabla_{\mathbf{x}} \log q(\mathbf{x})$

Ground truth score for 1D Mixture of Gaussians at various timesteps

At t=T (1000), the ground truth score would be same as the score of a isotropic Gaussian

What causes mode interpolation?

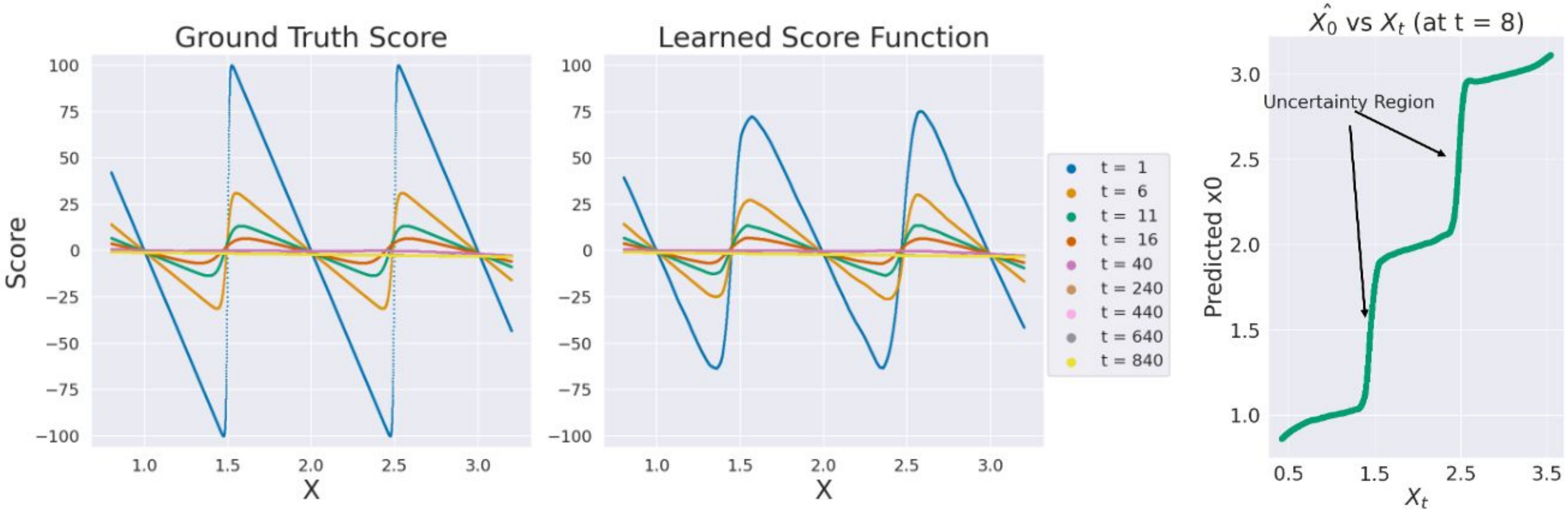
Diffusion models smoothly approximates the true score function



Smooth approximation of the true score function, particularly around the regions between disjoint modes of the distribution.

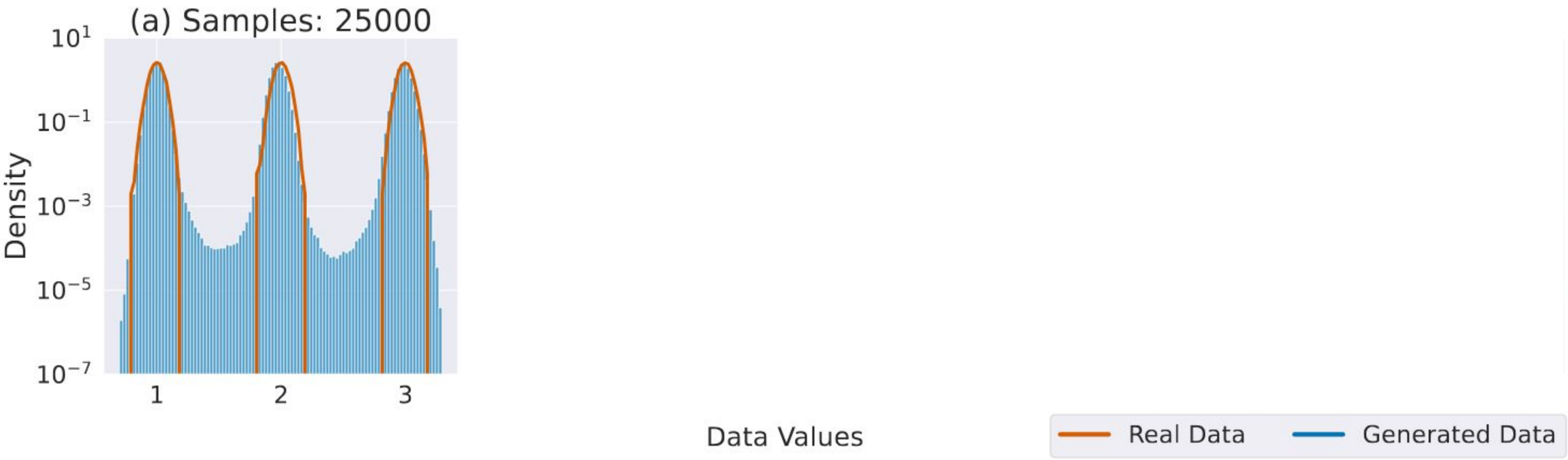
What causes mode interpolation?

Diffusion models smoothly approximates the true score function



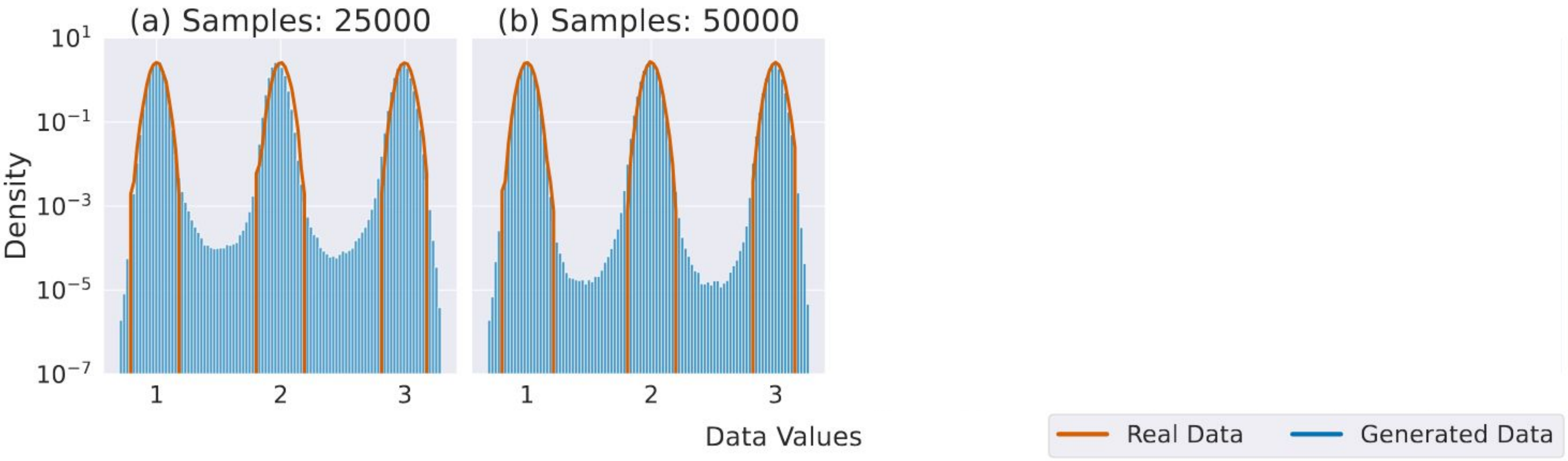
Mode Interpolation: 1D Gaussian

Rate of mode interpolation decreases as the number of training samples increases



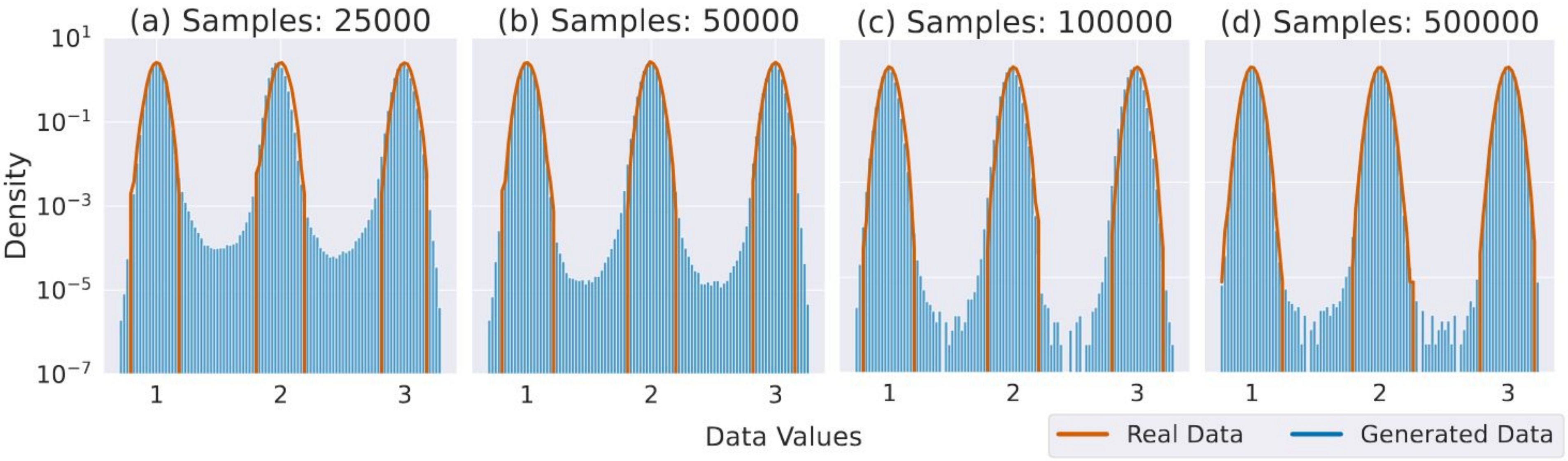
Mode Interpolation: 1D Gaussian

Rate of mode interpolation decreases as the number of training samples increases

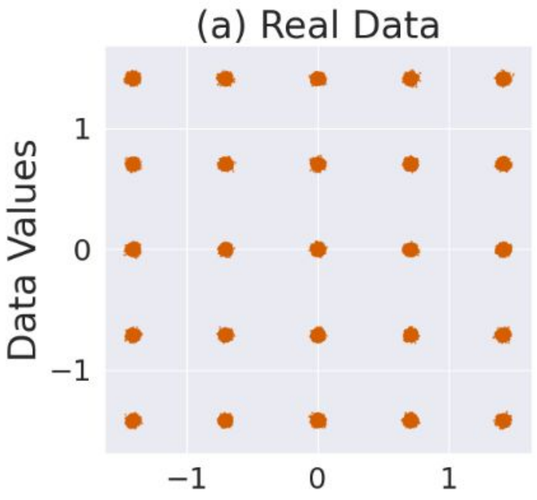


Mode Interpolation: 1D Gaussian

Rate of mode interpolation decreases as the number of training samples increases

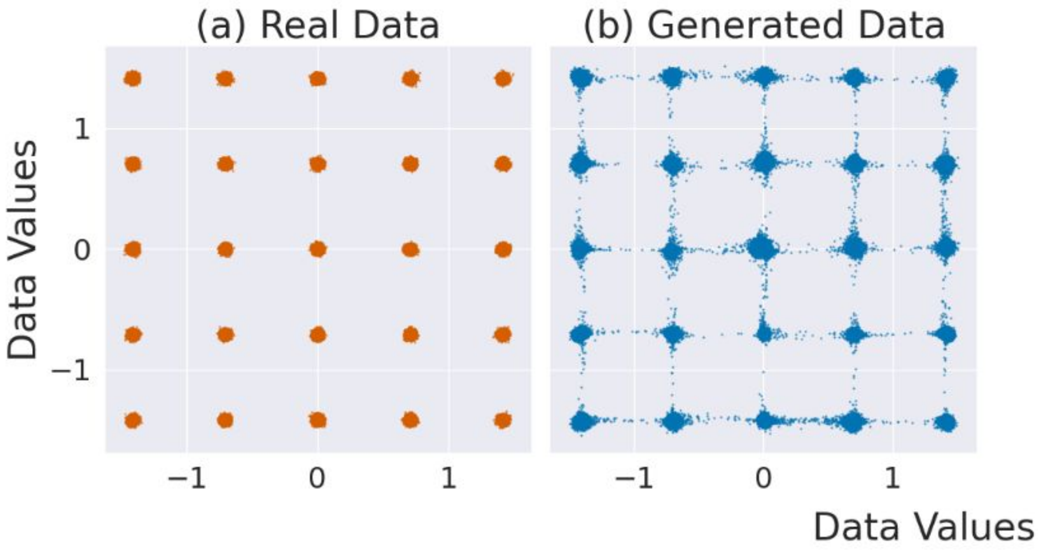


Mode Interpolation: 2D Gaussian

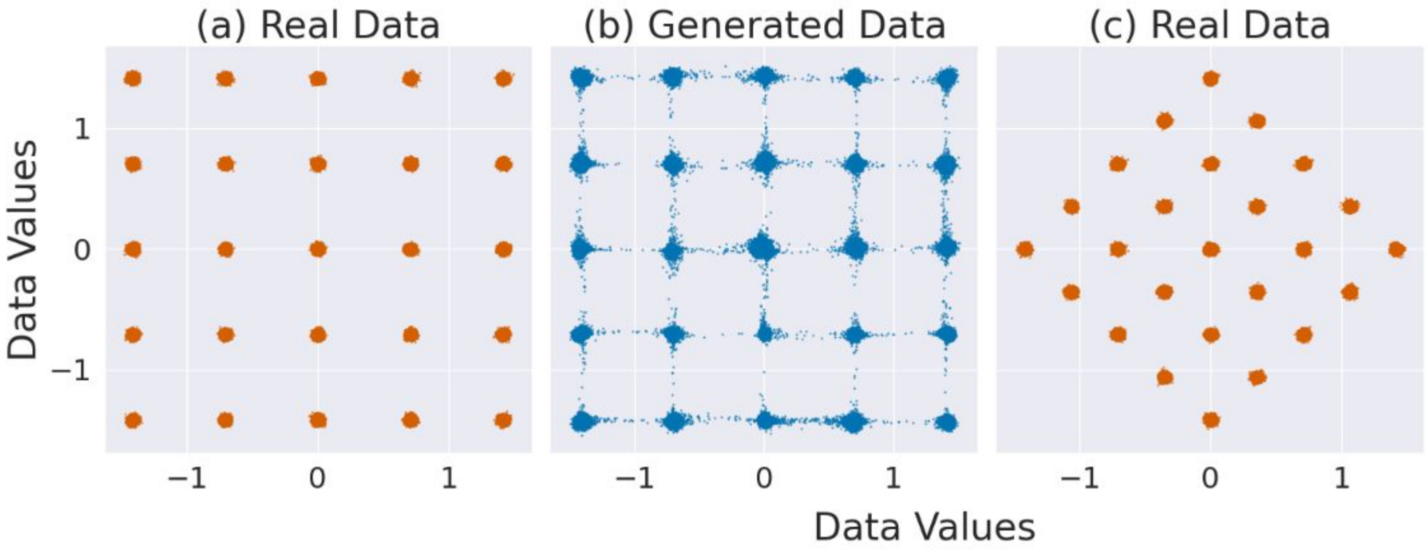


Data Values

Mode Interpolation: 2D Gaussian

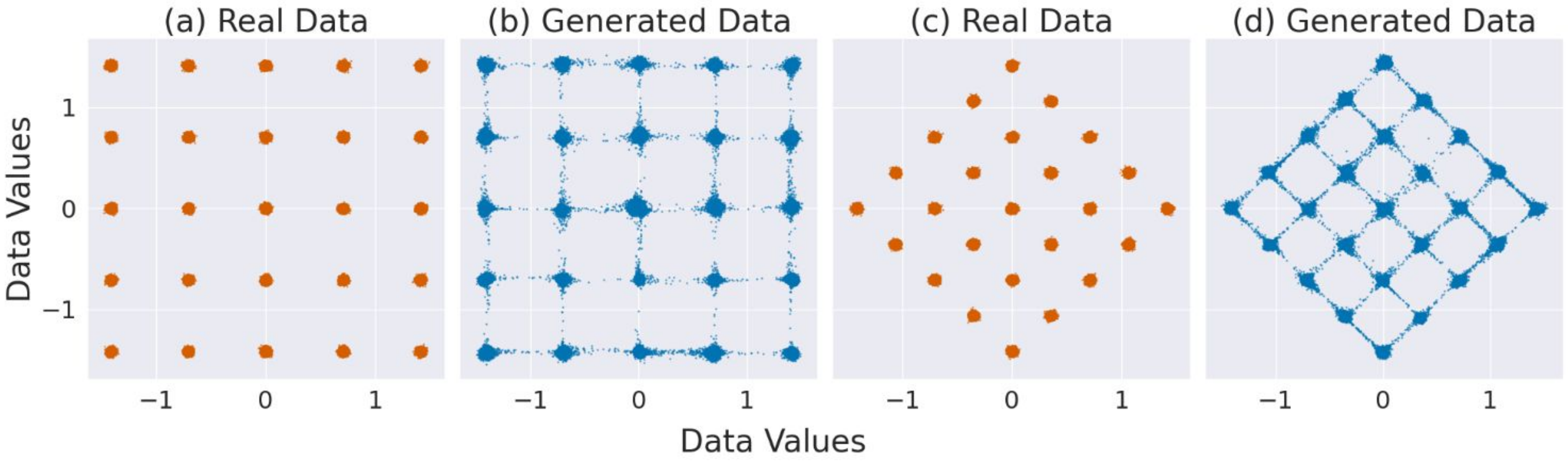


Mode Interpolation: 2D Gaussian



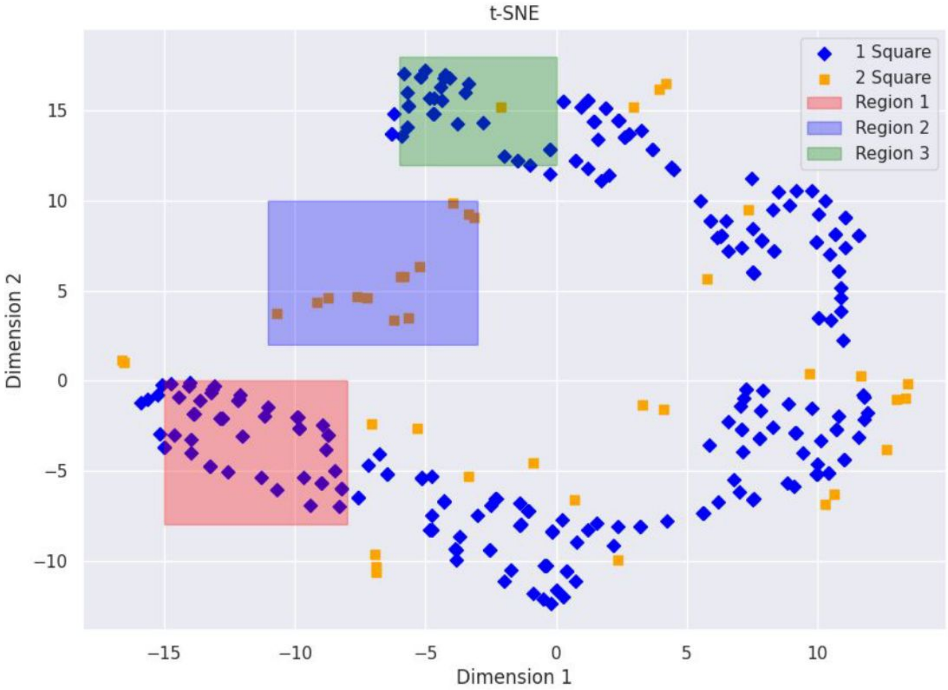
Mode Interpolation: 2D Gaussian

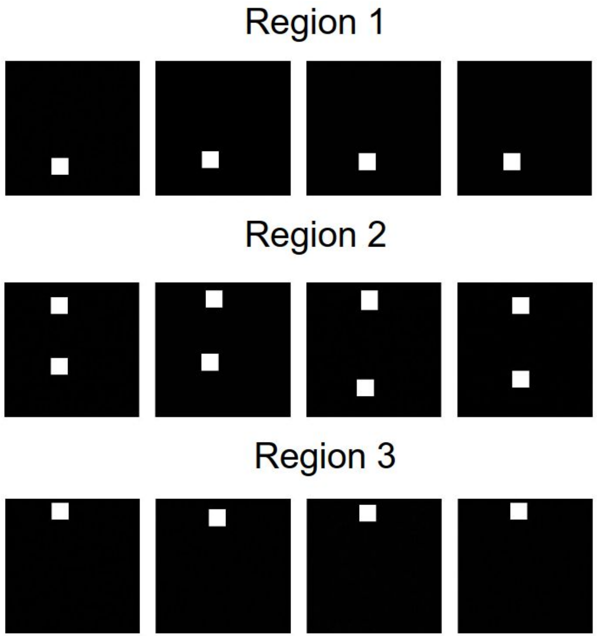
Diffusion models choose to interpolate between nearest modes



What is happening in the case of shapes?

Interpolation happens in representation space

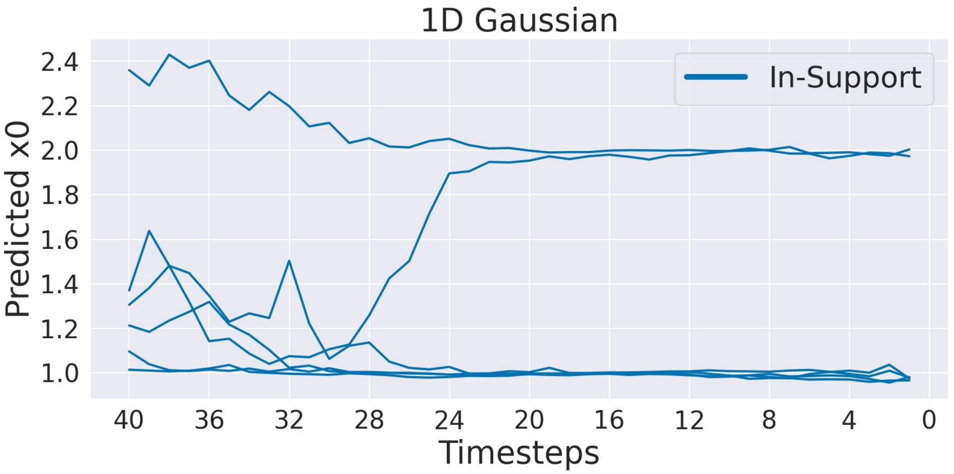




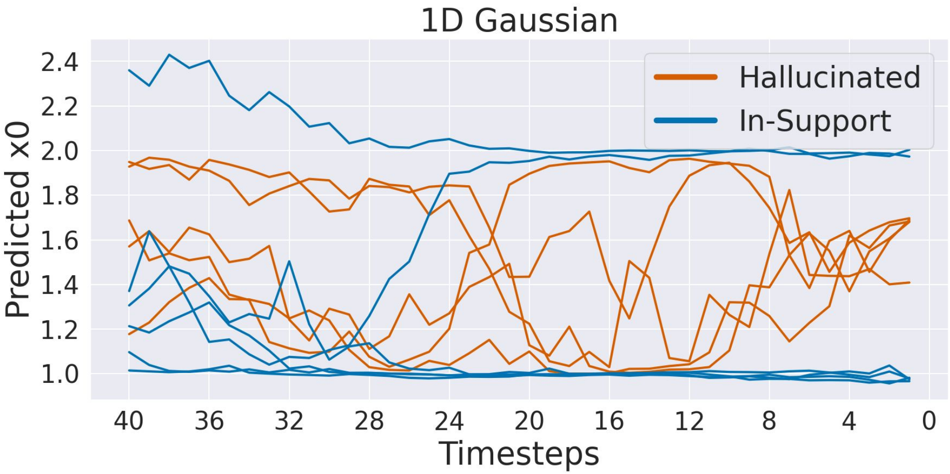
Diffusion Models know when they Hallucinate

Diffusion Models know when they Hallucinate

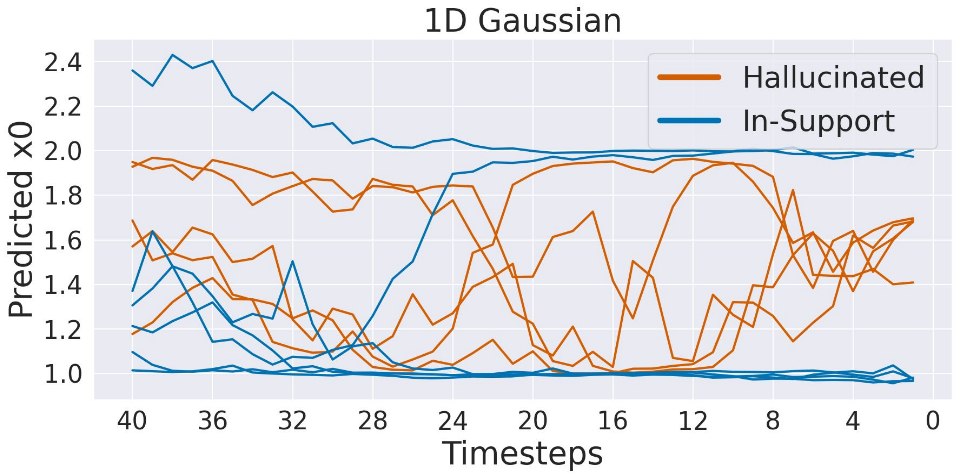
Diffusion Models know when they Hallucinate

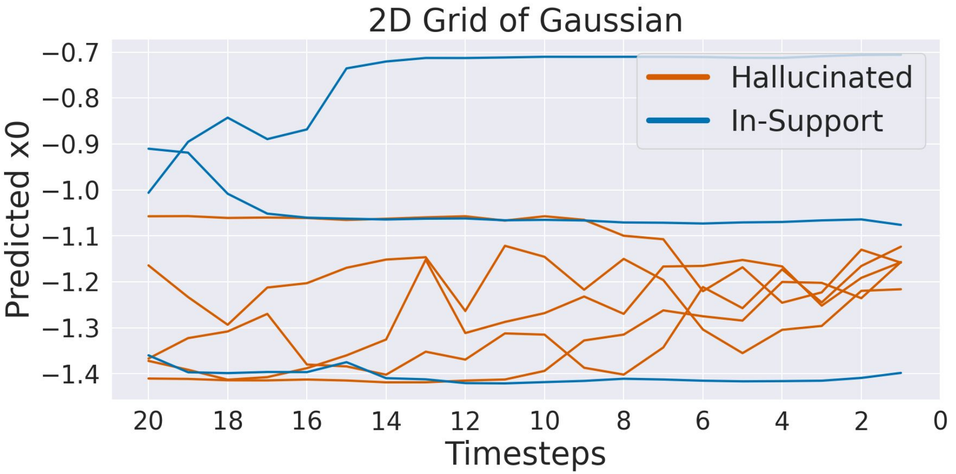


Diffusion Models know when they Hallucinate



Diffusion Models know when they Hallucinate



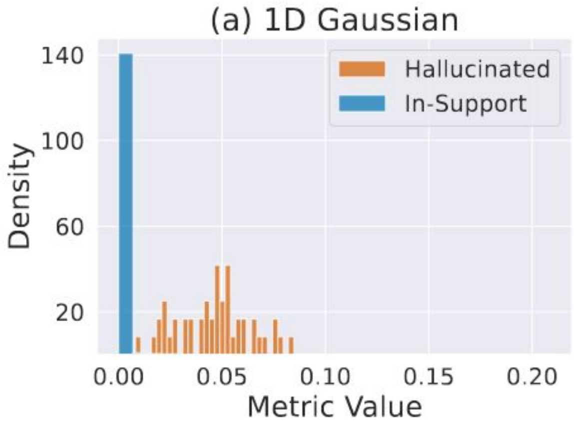


Diffusion Models know when they Hallucinate

$$\texttt{Hal}(x) = \frac{1}{|T_2 - T_1|} \sum_{i=T_1}^{T_2} \left(\hat{x_0}^{(i)} - \overline{\hat{x_0}^{(t)}} \right)^2$$

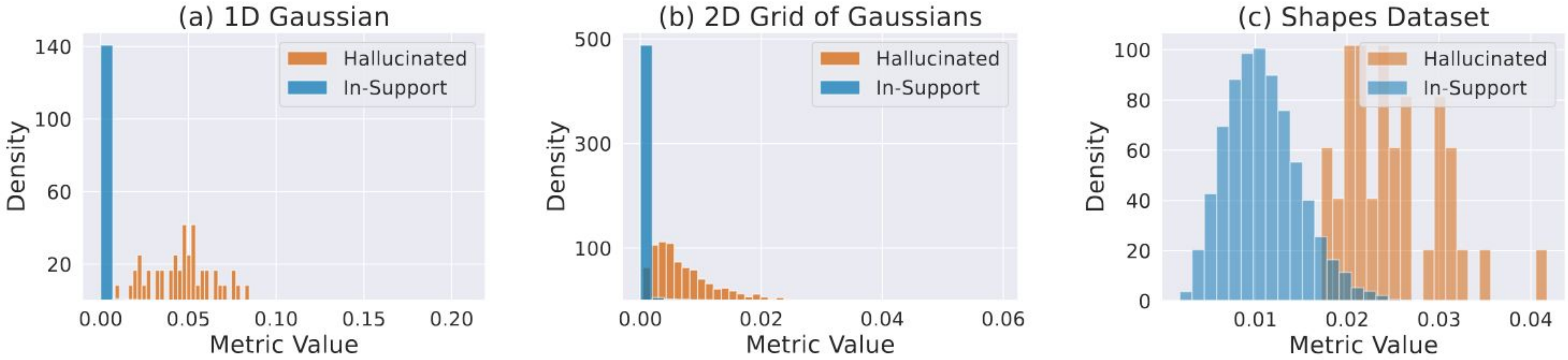
Diffusion Models know when they Hallucinate

$$\texttt{Hal}(x) = \frac{1}{|T_2 - T_1|} \sum_{i=T_1}^{T_2} \left(\hat{x_0}^{(i)} - \overline{\hat{x_0}^{(t)}} \right)^2$$



Diffusion Models know when they Hallucinate

$$\texttt{Hal}(x) = \frac{1}{|T_2 - T_1|} \sum_{i=T_1}^{T_2} \left(\hat{x_0}^{(i)} - \overline{\hat{x_0}^{(t)}} \right)^2$$



Realistic Settings

Let’s move on to realistic settings

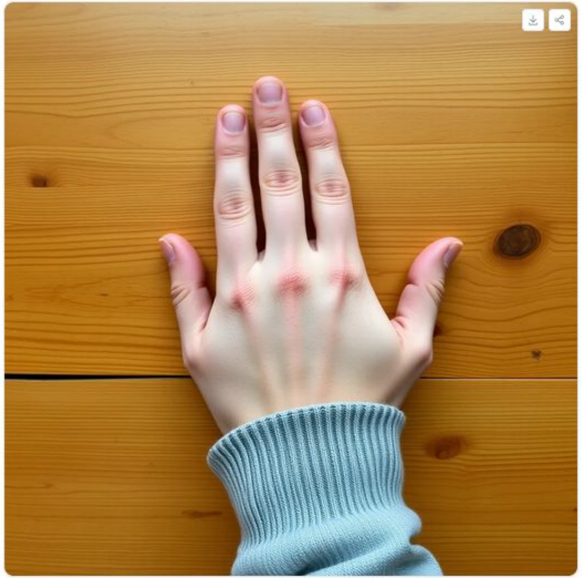
Let’s move on to realistic settings

FLUX.1 [schnell]

12B param rectified flow transformer distilled from FLUX.1 [pro] for 4 step generation
[blog] [model]

image of a left hand placed on a wooden table. top view

Run

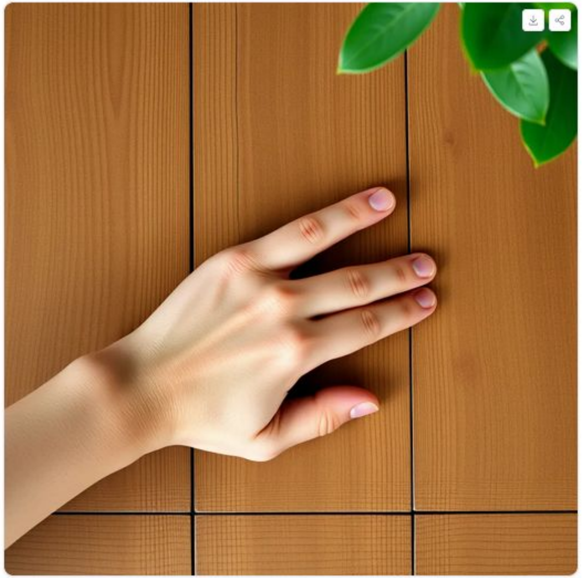


FLUX.1 [schnell]

12B param rectified flow transformer distilled from FLUX.1 [pro] for 4 step generation
[blog] [model]

image of a right hand placed on a wooden table. top vie

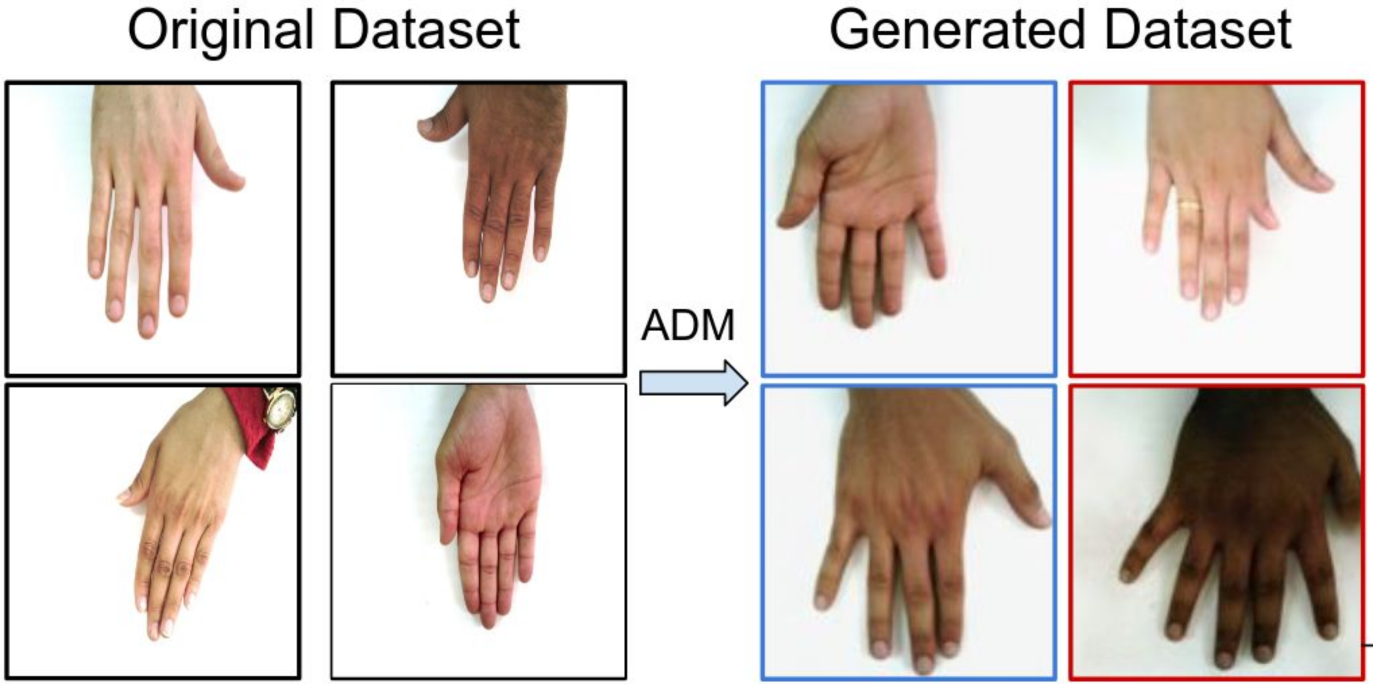
Run



Hands Dataset

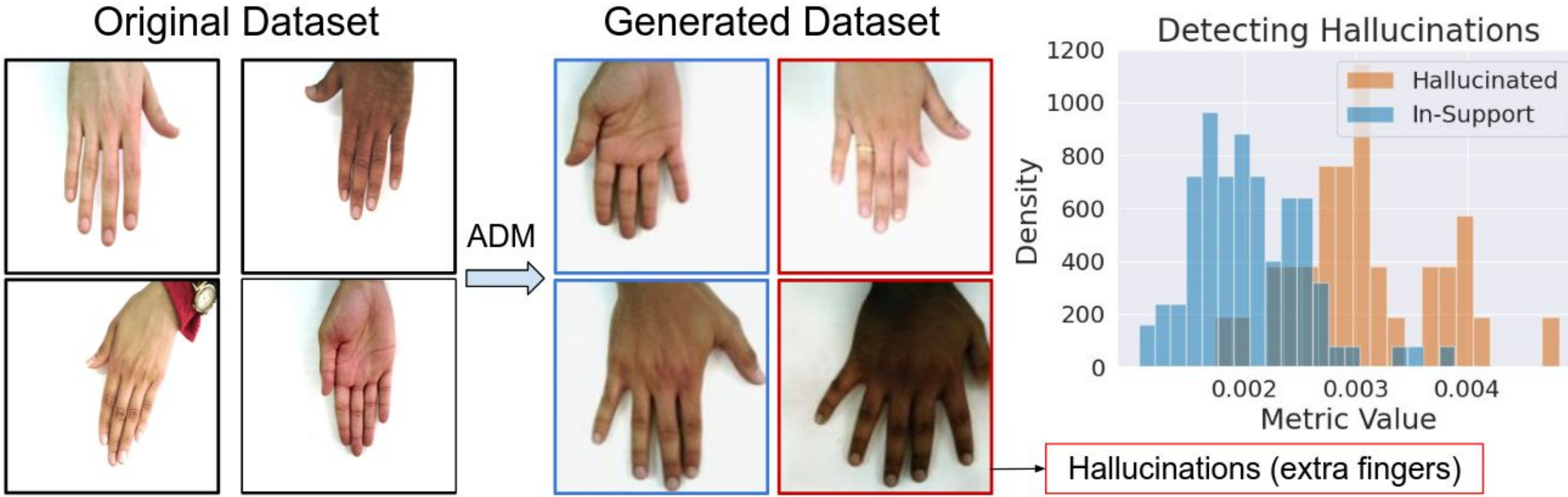
Afifi, Mahmoud. "11K Hands: Gender recognition and biometric identification using a large dataset of hand images." Multimedia Tools and Applications 78 (2019): 20835-20854.

Hands Dataset



Afifi, Mahmoud. "11K Hands: Gender recognition and biometric identification using a large dataset of hand images." Multimedia Tools and Applications 78 (2019): 20835-20854.

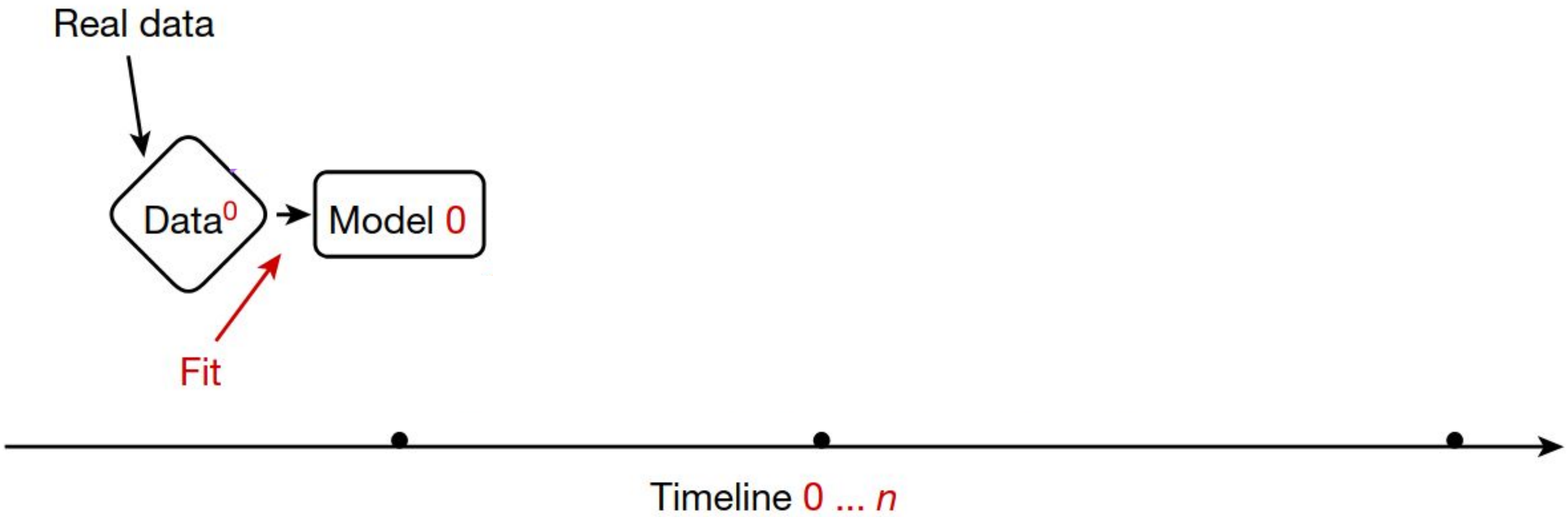
Hands Dataset



Recursive Model Training

Recursive Model Training

The internet is increasingly populated by more and more synthetic data.
Recursive training on synthetic data leads to mode collapse

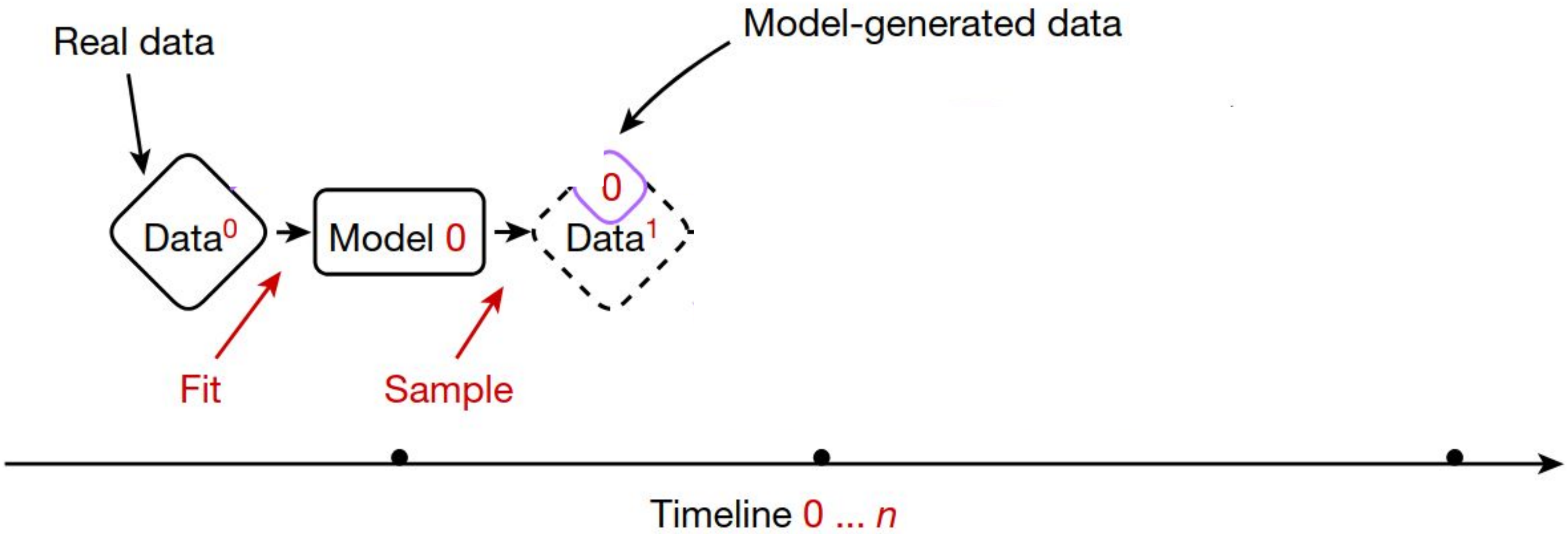


Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Recursive Model Training

The internet is increasingly populated by more and more synthetic data.

Recursive training on synthetic data leads to mode collapse

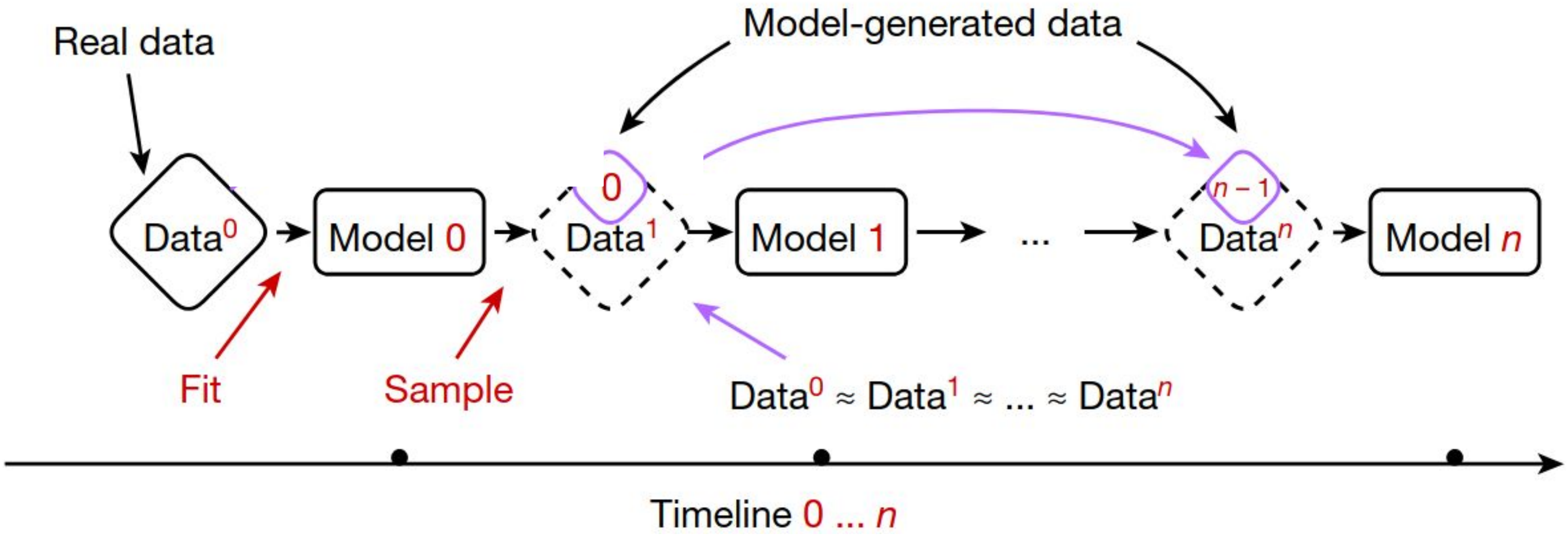


Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Recursive Model Training

The internet is increasingly populated by more and more synthetic data.

Recursive training on synthetic data leads to mode collapse



Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

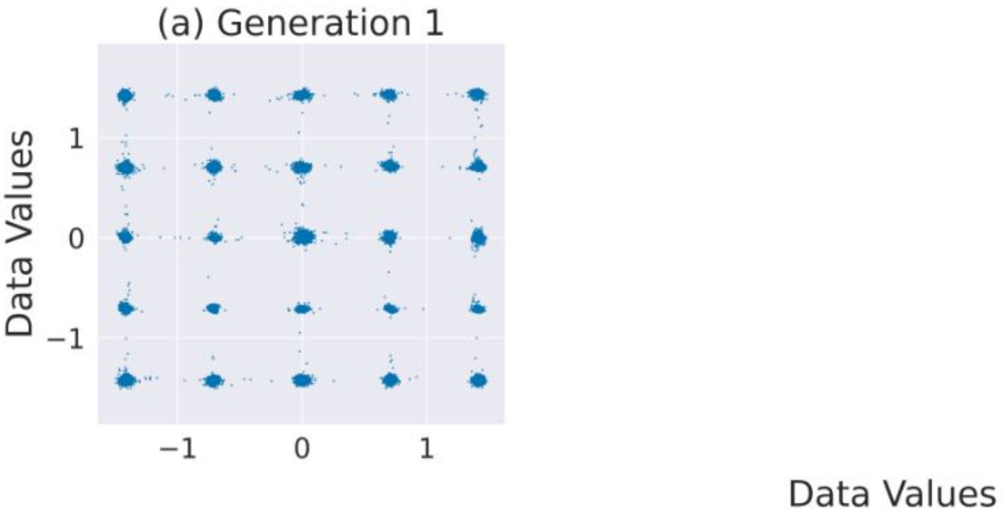
Recursive Model Training: Model Collapse

Past work has focused on model collapse without considering the interaction between the modes

Alemohammad, Sina, et al. "Self-consuming generative models go mad." arXiv preprint arXiv:2307.01850 (2023).

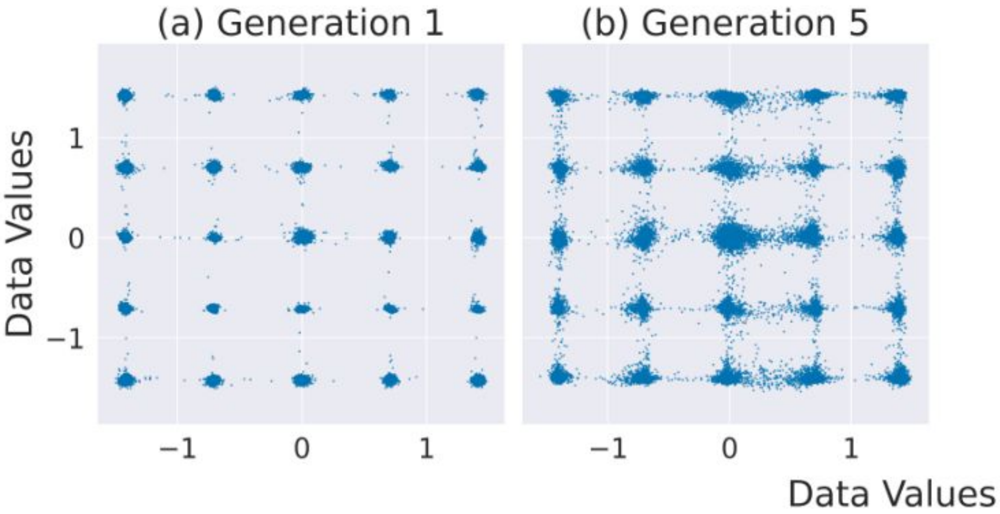
Recursive Model Training

Recursively training a DDPM on its own generated data using a square grid of 2D Gaussians



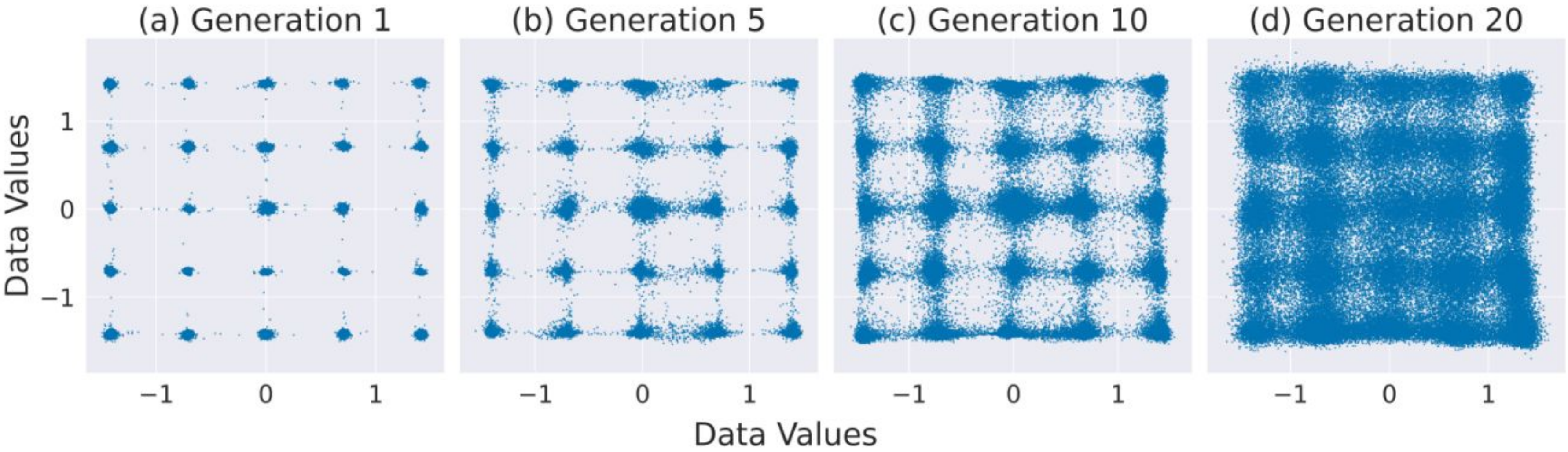
Recursive Model Training

Recursively training a DDPM on its own generated data using a square grid of 2D Gaussians



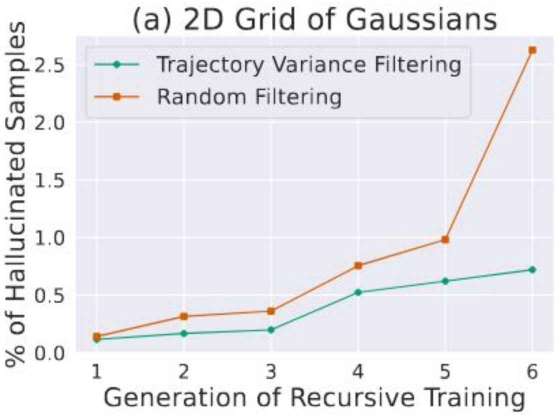
Recursive Model Training

Recursively training a DDPM on its own generated data using a square grid of 2D Gaussians



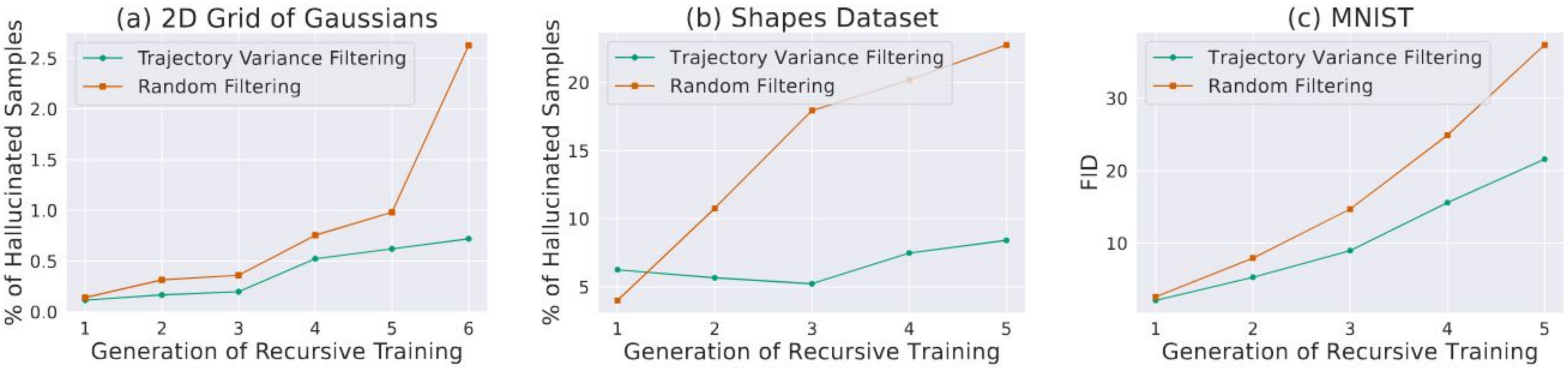
Mitigating Hallucinations with Pre-emptive Detection

Filter out hallucinated samples using the metric before training on samples from the previous generation of the diffusion model



Mitigating Hallucinations with Pre-emptive Detection

Filter out hallucinated samples using the metric before training on samples from the previous generation of the diffusion model



Summary

- Introduce a failure-mode of diffusion models: mode interpolation
- Explanation of why mode interpolation occurs
- Metric to detect hallucinations in diffusion models
- Potential hypothesis for inaccurate modeling of hands/limbs in modern text-to-image generative models.
- Novel Perspective on the Recursive Training of Generative Models