Do diffusion models generalize?

Zahra Kadkhodaie
NYU

Eero Simoncelli
NYU & Flatiron Institute

Stéphane Mallat
Collège de France & Flatiron Institute

NYU Center for Data Science

FLATIRON INSTITUTE
Center for Computational Neuroscience

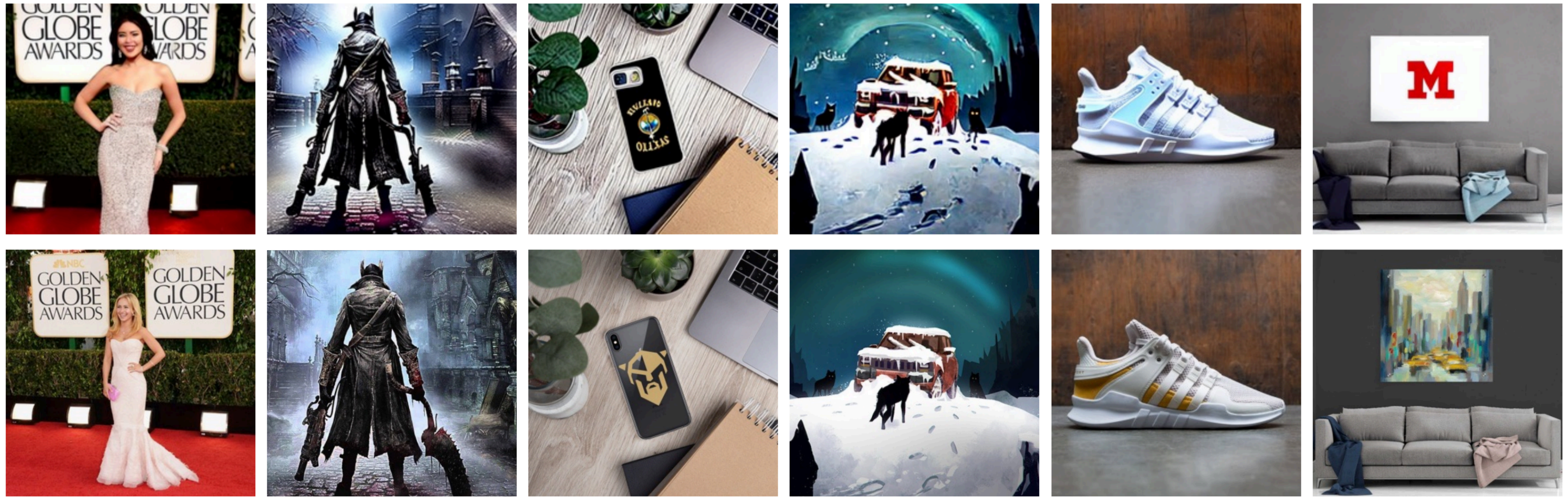
COLLÈGE DE FRANCE 1530

Generalization vs memorization

- Generative models can reproduce new images…
- …but also memorize their training set
- Does the learned model depend on the individual training samples?

[Ho et al, 2022]

[Carlini et al, 2023]

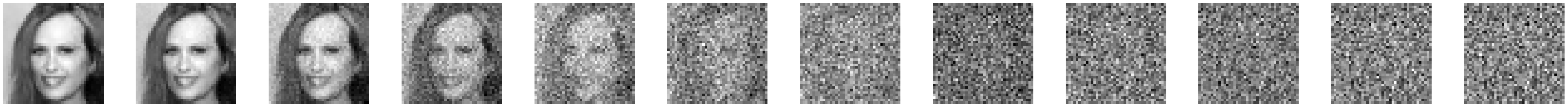


[Somepalli et al, 2023]

Generating images with the score

Forward process: diffuse images by adding noise

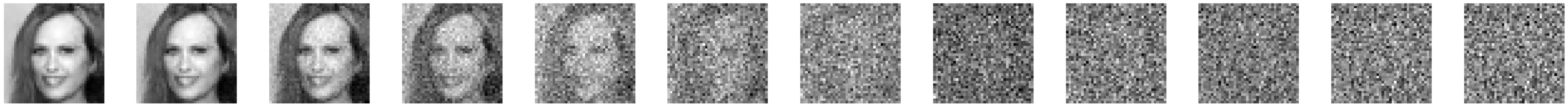
x_t



t

$x_0 \sim p_{\text{data}}$ $x_{t+\mathrm{d}t} \sim \mathcal{N}(x_t,\ \mathrm{d}t\,\mathrm{Id})$ $x_T \sim \mathcal{N}(0,\ T\,\mathrm{Id})$

By reversing time, we can generate new images if we know the score!



x_t

t

$x_0 \sim p_{\text{data}}$ $x_{t-\mathrm{d}t} \sim \mathcal{N}(x_t + \mathrm{d}t\,\nabla \log p(x_t),\ \mathrm{d}t\,\mathrm{Id})$ $x_T \sim \mathcal{N}(0,\ T\,\mathrm{Id})$

(Song & Ermon, 2019; Ho et al., 2020; Kadkhodaie & Simoncelli, 2020)

Learning the score by denoising

The score can be rewritten as a conditional expectation:

$$\nabla \log p(x_t) = \mathbb{E}[\,\nabla \log p(x_t \,|\, x_0) \,|\, x_t] = \frac{1}{t}(\mathbb{E}[x_0 \,|\, x_t] - x_t)$$

(marginalization) (Gaussianity)

We can learn it by least-squares regression (denoising)!

$$\min_f \; \mathbb{E}\left[\|x_0 - f(x_t)\|^2\right]$$

x_0 x_t $f(x_t)$

$$\nabla \log p(x_t) \approx \frac{1}{t}(f(x_t) - x_t)$$

(Miyasawa, 1961; Tweedie, via Robbins, 1956)

The dangers of memorization

- In practice, we approximate the 'true' p_{data} with an empirical distribution of training samples $\{x_1, \ldots, x_n\}$

- The optimal solution is then to learn a model of this empirical distribution: in other words, memorize the training set

$$f(y) = \sum_{i=1}^{n} w_i(y)\, x_i \qquad\qquad w_i(y) \propto e^{-\frac{\|y - x_i\|^2}{t}}$$

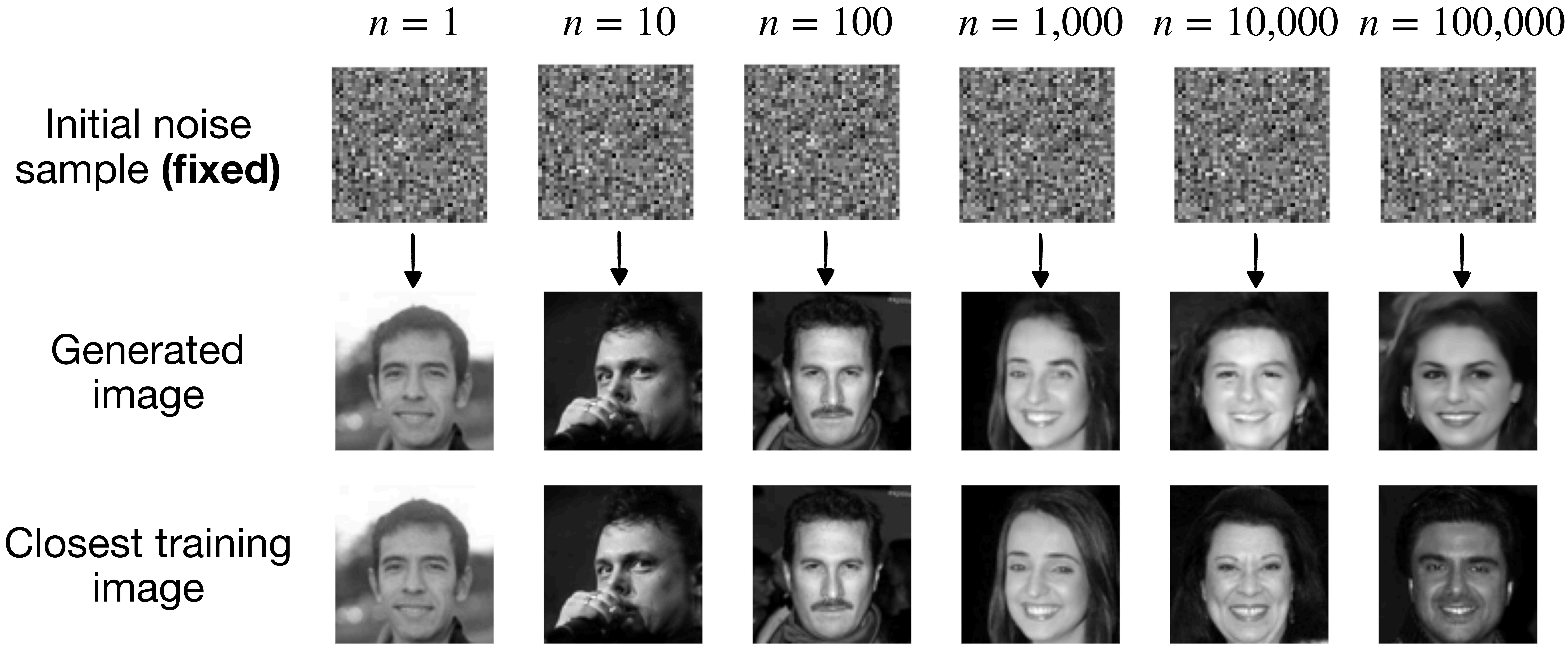
- The resulting network always generate one of the training images

- We rely on the network **not to perfectly minimize** the training loss!

From memorization to generalization

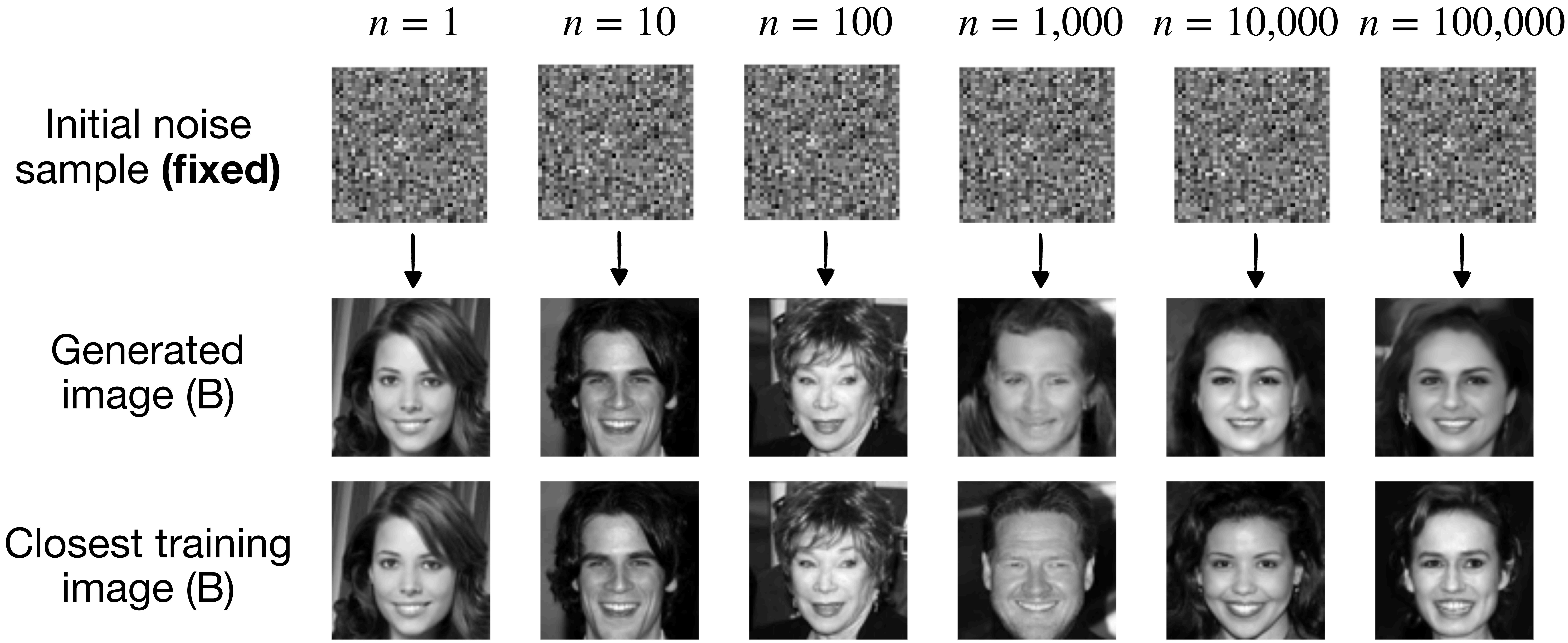
We train networks on n face images for increasing n, and compare the generated images with the training images.

(Yoon et al, 2023)



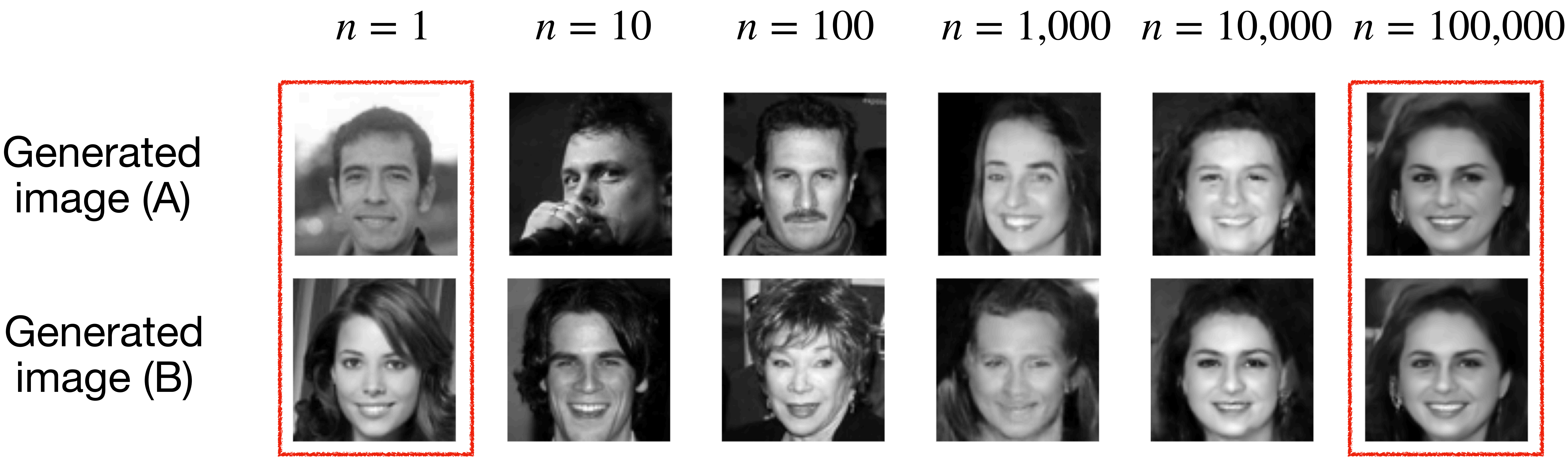
From memorization to generalization (bis)

We repeat the analysis with networks trained on another, **non-overlapping** set of face images.



From memorization to generalization (ter)

Let us compare the mages generated by the two networks **from the same noise sample.**



Memorized images from respective training sets

Identical generated image from neither of the training sets

Strong evidence of generalization.

Which inductive biases allow the networks to beat the curse of dimensionality?

Inductive biases: teaser

- Direct link between generalization and optimality of denoising

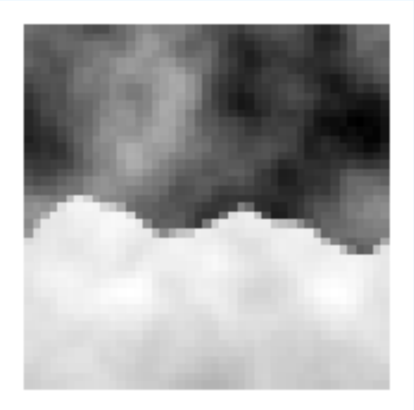
$$D_{\mathrm{KL}}(p(x)\,\|\,p_\theta(x)) \leq \int_0^\infty \Big(\mathrm{MSE}(f_\theta,\sigma^2) - \mathrm{MSE}(f^\star,\sigma^2)\Big)\,\sigma^{-3}\,\mathrm{d}\sigma,$$

- Focus on synthetic datasets where we know (approximately) the optimal denoiser
- Deviations from optimality tell us about the inductive biases of the network!

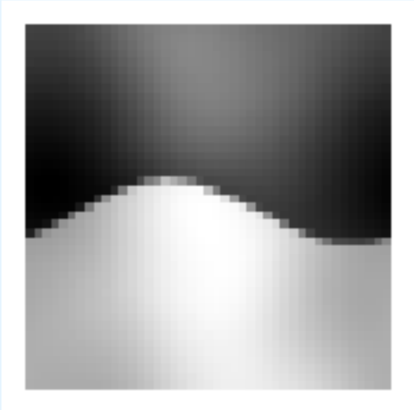
Optimality (aligned inductive biases)

Geometric C^α images

$\alpha = 2$



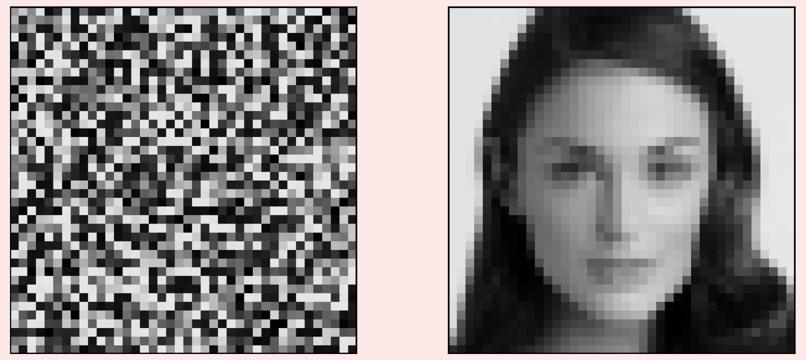
$\alpha = 4$



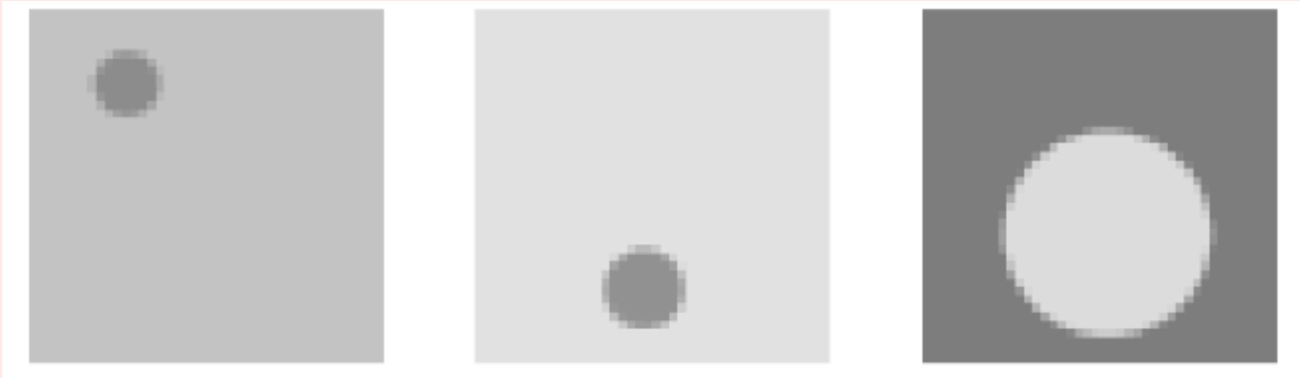
(Korostelev & Tsybakov, 1993; Donoho, 1999; Peyré & Mallat, 2008)

Suboptimality (misaligned inductive biases)

Shuffled faces



Low-dimensional manifolds



More details: arXiv:2310.02557

Summary

- Diffusion models transition from memorization to generalization when the training set size increases
  - Note: the critical training size depends on the network architecture, image resolution, etc…
- Strong generalization: we learn the same probability model independently of the training samples!
  - The networks learn the same underlying function
- This generalization relies on inductive biases towards high-dimensional geometric structures (see paper for more details)

Kadkhodaie, FG, Simoncelli, and Mallat. *Generalization in diffusion models arises from geometry-adaptive harmonic representations*. arXiv:2310.02557, *ICLR*, 2024.