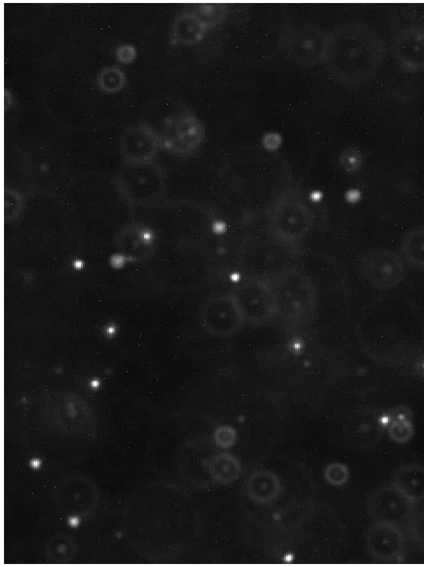
Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training

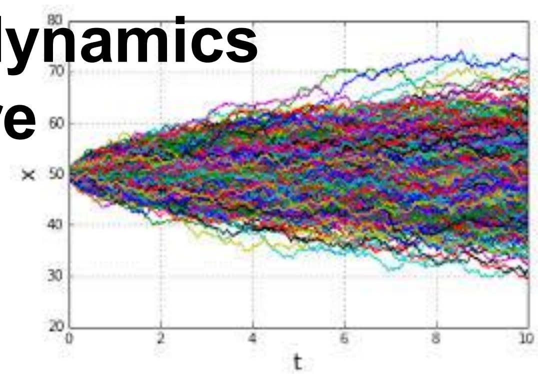
Bonnaire, Urfin, Biroli, Mézard

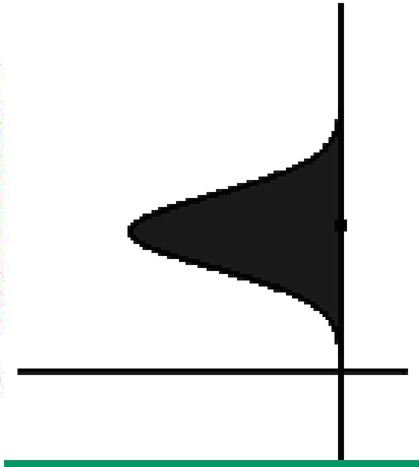
NeurIPS 2025

Diffusion

Brown: observes Brownian motion $\mathbf{B}(t)$
Laplace, Fourier: heat $H'(t) = \Delta H(t)$
Einstein: Avogadro number via BM
Fokker Planck equation, Langevin dynamics
Wiener: BM as a Fourier series
Polya: BM as a limit of Random Walk
Ornstein-Uhlenbeck $dx(t) = -x(t)dt + \mathbf{B}(t)$
Itô: Stochastic DEs $dt = (d\mathbf{B}(t))^2$
Feyman-Kac path integrals
Hairer: renormalixation for rough paths
Non-equilibrium thermodynamics
Diffusions are everywhere
BM/RW is universal
Trajectories vs flows
Fourier analysis
?? Rough paths ??







Diffusion models

- **Sohl-Dickstein, Weiss, Maheswaranathan, Ganguli,(2015)**

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

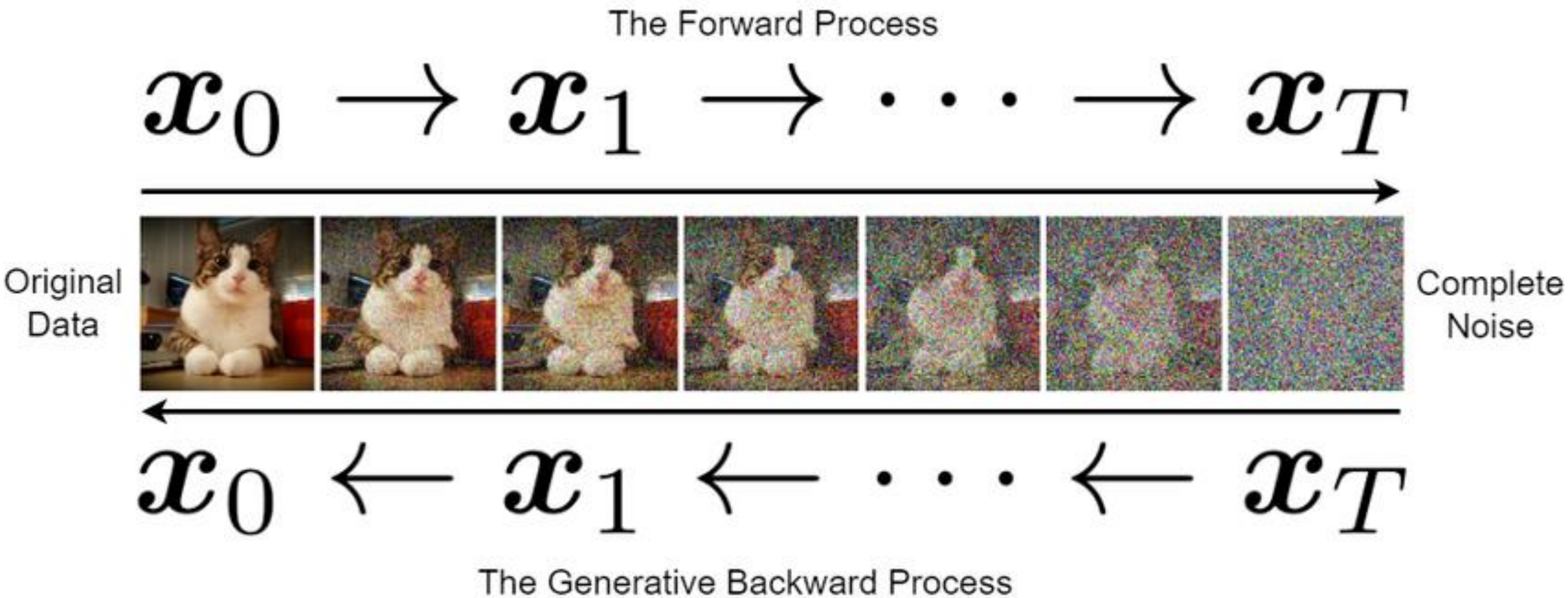
- **Ho, Jain, Abbeel,(2020).**

Denoising Diffusion Probabilistic Models

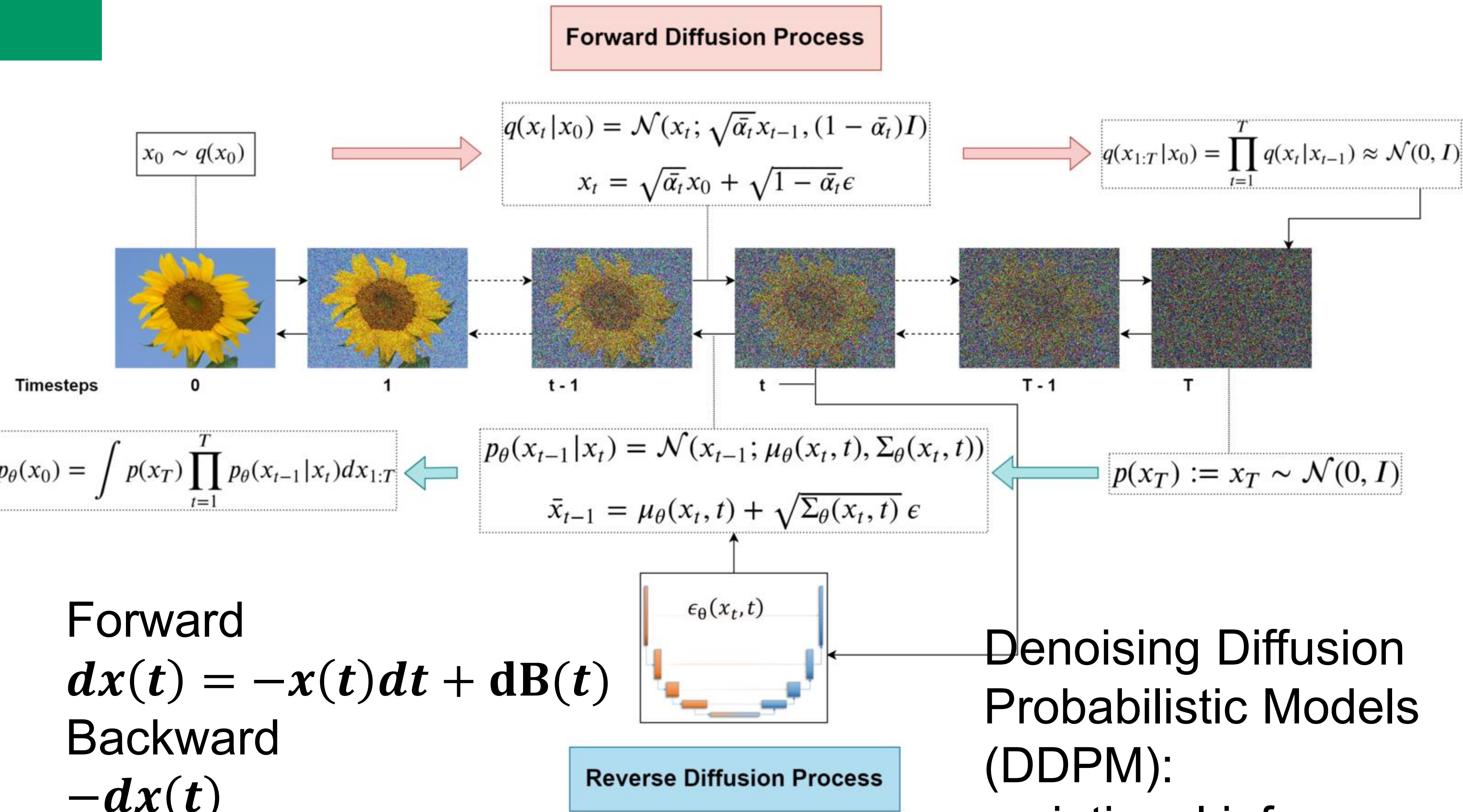
- **Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole (2021)**

Score-Based Generative Modeling through Stochastic Differential Equations

- Diffuse by Gaussian to get (almost) Boltzmann
- Integrate reverseSDE, try to mathc the score



Diffusion models



Forward

$$\boldsymbol{dx(t) = -x(t)dt + \mathrm{dB}(t)}$$

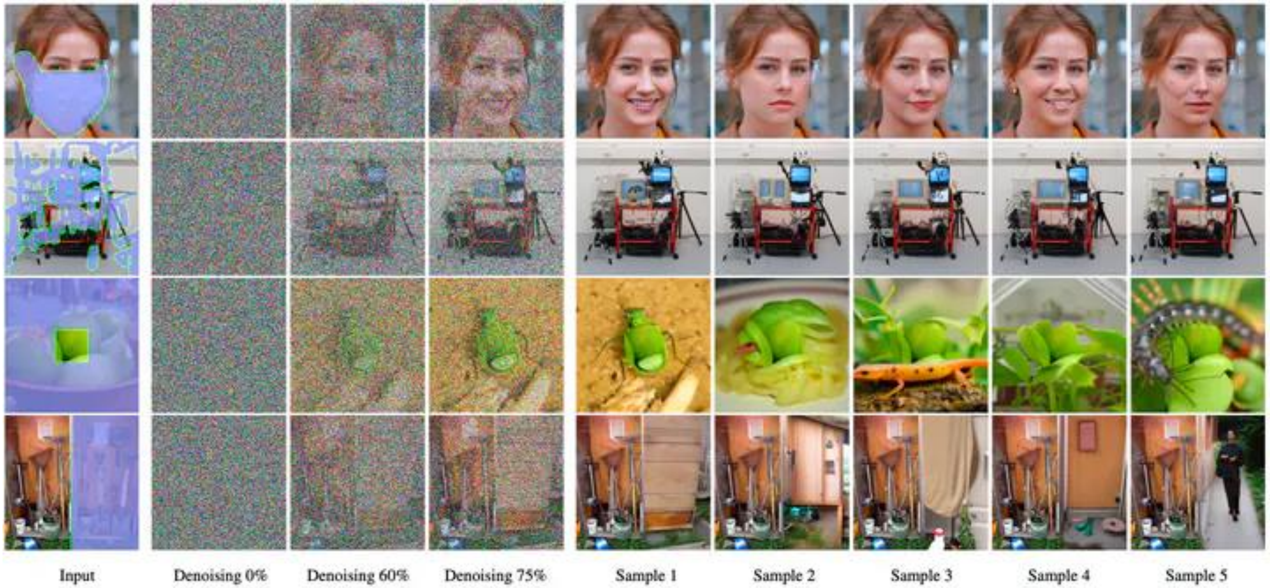
Backward

$$\boldsymbol{-dx(t) = (x(t) + 2\nabla_{\mathrm{x}}\log \mathrm{P_t}(\mathrm{x}))dt + \mathrm{B}(t)}$$

Denoising Diffusion Probabilistic Models (DDPM): variational inference

Pluses and minuses

- \+ **Training stability** No adversarial min–max optimization
- \+ **High sample quality** Excellent mode coverage Low artifacts
- \+ **Flexibility** Conditional generation Inpainting, super-resolution, editing
- \+ **Strong theoretical grounding** Explicit likelihood Well-defined SDE
- \- **Sampling cost** Requires tens to hundreds of denoising steps
- \- **Compute-intensive training** Large models, long training times
- \- **Less interpretable representations** Cf. VAEs / latent-variables



Question

Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training
Tony Bonnaire, Raphaël Urfin, Giulio Biroli, Marc Mezard
Implicit dynamical regularization during training gives diffusion models a generalization window that widens with the training set size, so stopping within this window prevents memorization.

Why don't diffusion models memorize training data?

In principle they could – overparametrization!

Yet, empirically, diffusion models generalize extremely well, generate novel samples, interpolate smoothly, **but** memorization only appears very late in training, if at all.

Main claim: that generalization in diffusion models is driven primarily by training dynamics, through a form of implicit dynamical regularization.

Empirical observations

$\mathbf{32 \times 32}$ portraits, $\boldsymbol{p = 4 \times 10^6}$ trainable parameters
FID = 2-Gaussian Earth Mover Distance to training set
$\boldsymbol{f_{mem}}$ = distance to the net of samples

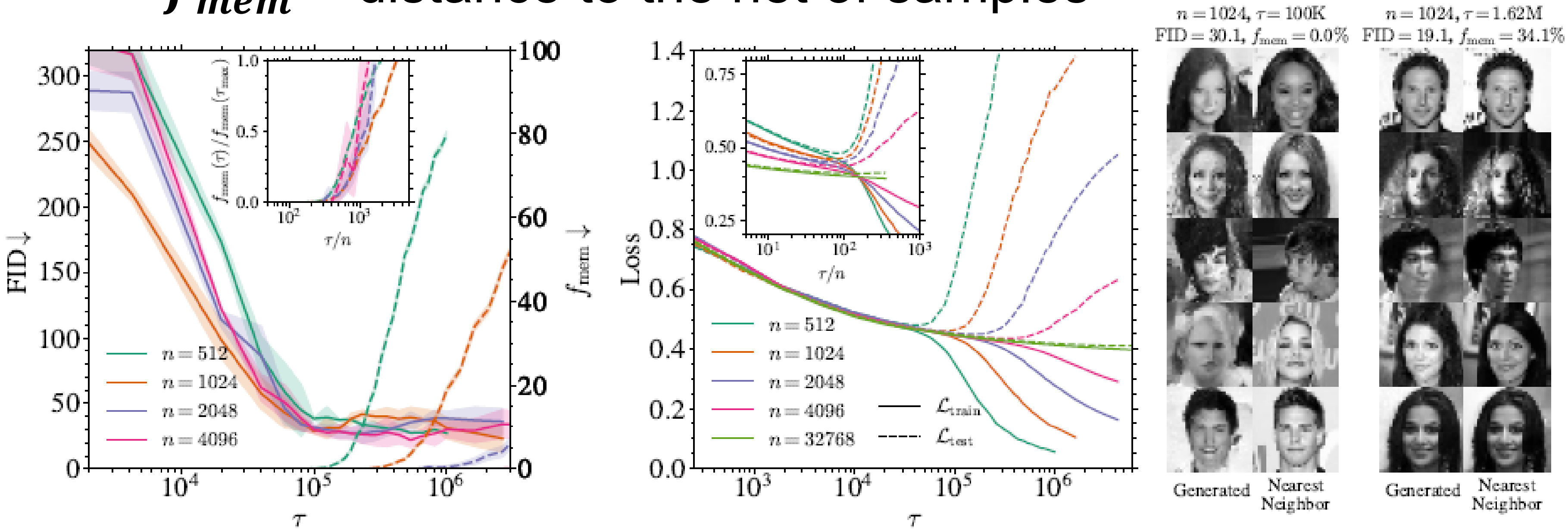
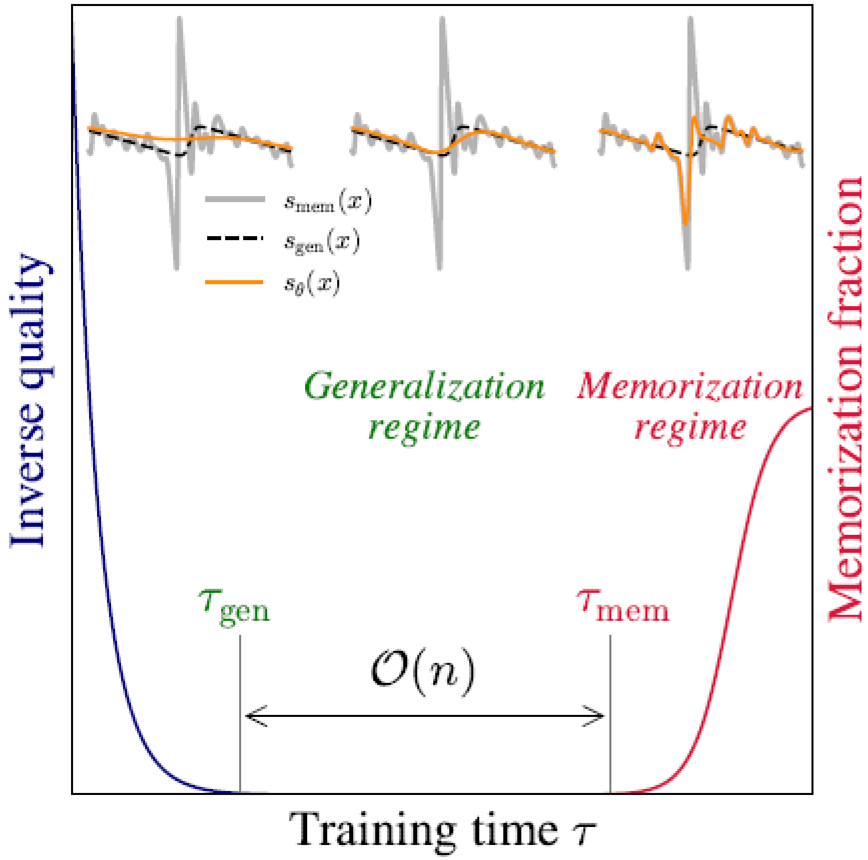
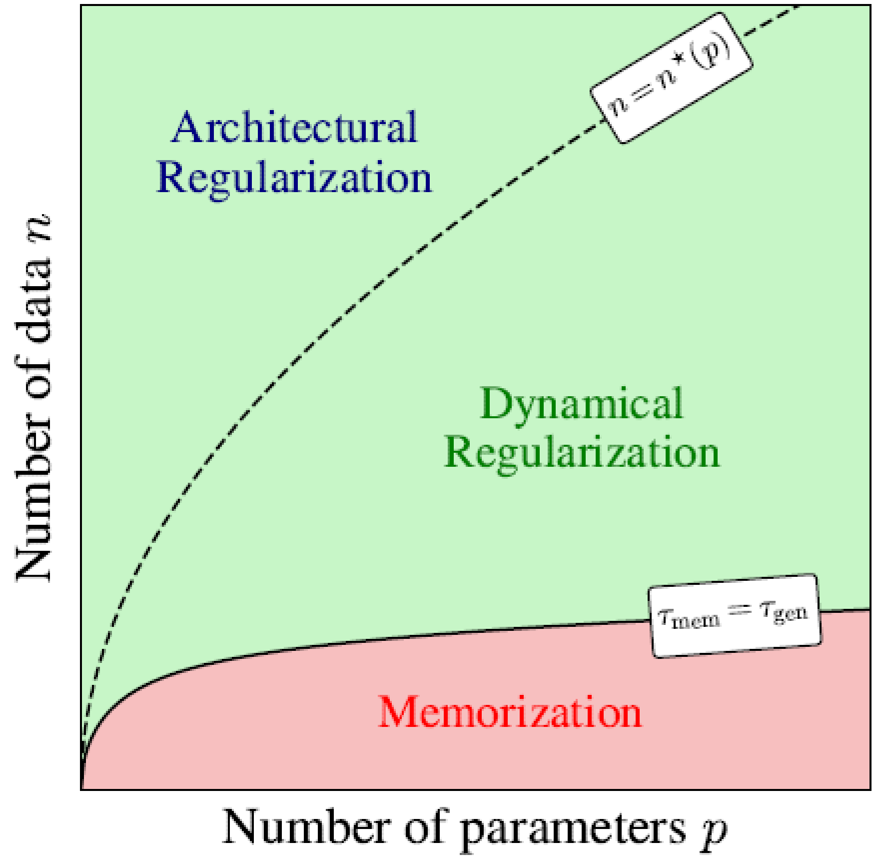


Figure 2: **Memorization transition as a function of the training set size n for U-Net score models on CelebA.** *(Left)* FID (solid lines, left axis) and memorization fraction f_{mem} (dashed lines, right axis) against training time τ for various n. Inset: normalized memorization fraction $f_{\mathrm{mem}}(\tau)/f_{\mathrm{mem}}(\tau_{\max})$ with the rescaled time τ/n. *(Middle)* Training (solid lines) and test (dashed lines) loss with τ for several n at fixed $t = 0.01$. Inset: both losses plotted against τ/n. Error bars on the losses are imperceptible. *(Right)* Generated samples from the model trained with $n = 1024$ for $\tau = 100\mathrm{K}$ or $\tau = 1.62\mathrm{M}$ steps, along with their nearest neighbors in the training set.

Empirical observations

$\tau_{gen} \approx$ const: onset of good sample quality
$\tau_{mem} \asymp$ dataset size: onset of memorization
Hence growing window without memorization





Theoretic rationale

Memorization is ultimately driven by the overfitting of the empirical score.
Initially L_{train} and L_{test} are indistinguishable, but beyond a critical time, L_{train} continues to decrease while L_{test} increases, with generalization loss depending on n.
Memorization is not due to data repetition –even if at fixed t all models have processed each sample equally often, larger n postpone memorization.

Instead, we see **implicit dynamical regularization**: regularization arises indirectly from the **optimization dynamics themselves**, via **spectral bias**.

Smooth, low-frequency components are learned quickly. Highly oscillatory, high-frequency components are learned slowly.

Toy model for analysis

Score: **linear random-features model** $s(x) = \sum_k w_k \phi_k(x)$

- Train **full-batch gradient descent** on the denoising
- Features ϕ_k are fixed; only weights w_k are trained
- Data drawn i.i.d. from a population distribution p_0

Training dynamics is exactly solvable

- Gradient flow reduces to **linear regression**
- Diagonalize dynamics in the eigenbasis of the feature covariance $C = \mathbf{E}[\phi(x)\phi(x)^T]$, with eigenvalues λ_k

Closed-form solution

Modes k evolve independently: $w_{k(t)} = w_k\left(1 - e^{-\lambda_k t}\right)$

Conlsuion

Large λ_k (smooth) modes learned fast → generalization
Small λ_k (fine-scale) modes learned slow → delayed memorization

Takeaways

- Explains robustness of diffusion models
- Early stopping is theoretically justified
- Generalization driven by dynamics

Do we want to study DM at LM?

\+ A lot of SDE has not been used yet

\- Need big compute?

Some things to do

- Modify diffusion component
- Control high frequencies with Fourier
- Try to project on training set directly