# Generalization in Diffusion Models

What we know and don't know about generalization

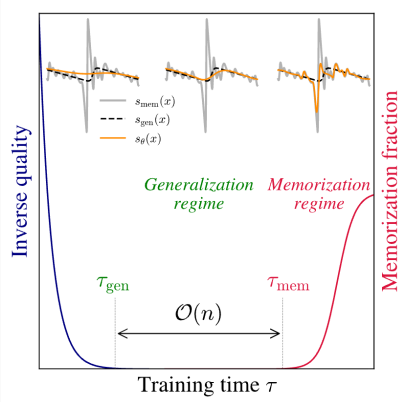
Yu-Han Wu
Supervised by: Gérard Biau, Claire Boyer, Pierre Marion, Quentin Berthet, Romuald Elie

04/06/2026

LPSM, Sorbonne University & Google DeepMind

# A bit of a background story ...

One of the best paper awards of last year's NeurIPS (Bonnaire et al. 2025)

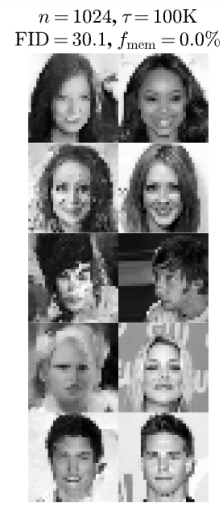



At a similar period of time, Favero et al. (2025) showed similar results.

# A closer look into the actual images

/!\ It is unclear if early stopping is actually "generalization"

Already, the images are $28 \times 28$ grey scale images.



# Understanding diffusion models requires rethinking (again) generalization

Pierre Marion[1*] & Yu-Han Wu[2*]

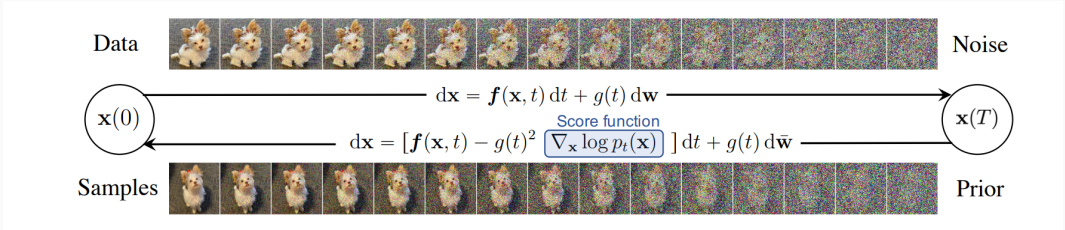
[1] Inria, École Normale Supérieure, PSL Research University

[2] LPSM, Sorbonne Université & Google DeepMind

*: Equal contribution

Preprint on arxiv:2605.06077

# Diffusion Models: The Big Picture




**Forward process** (data $\rightarrow$ noise):

$$dx = f(x,t)dt + g(t)dW$$

**Reverse process** (noise $\rightarrow$ data):

- $dx = [f(x,t) - g^2(t)\nabla_x \log p_t(x)]dt + g(t)d\bar{W}$

**Training:** Denoising score matching

$$\min_{\theta} \; \mathbb{E}_{x_0,\varepsilon,t}\left[\|s_\theta(x_t,t) - \varepsilon\|^2\right]$$

# The Memorization Problem in Diffusion Models

**Key observation:** The global minimum of the empirical score matching loss corresponds to the score of the **noised empirical measure**:

$$p_t^* = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(x_i, \sigma_t^2 I),$$

sampling from $p_t^*$ reproduce the training samples.

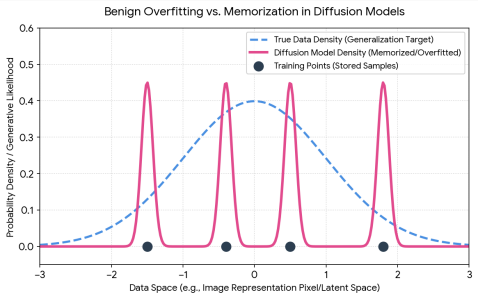
**Unlike supervised learning:**

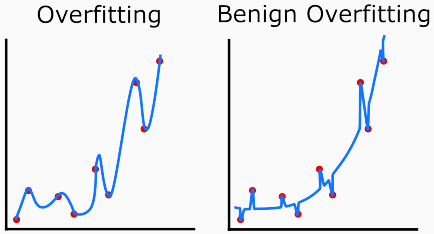
- In supervised: interpolation + generalization can coexist (*benign overfitting*)
- In diffusion: memorization $\neq$ generalization — they are **incompatible**

## Fundamental Tension

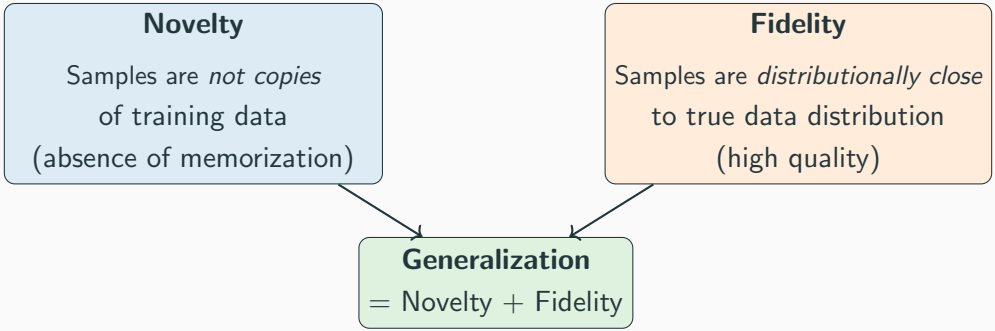
- Reaching global optimum $\Rightarrow$ memorization
- Yet practical models generate **novel, diverse** samples
- Something prevents full memorization
- **What?**

# Illustration



# Two Facets of "Generalization"




**Pure noise:** perfect novelty, zero fidelity. **Copy machine:** high fidelity, zero novelty.

# Family 1: Capacity vs. Dataset Size

**Idea:** Memorization occurs when the model has sufficient capacity relative to the dataset.

**Key results:**

- Yoon et al. (2023), Gu et al. (2023): small dataset $\Rightarrow$ memorization; large $\Rightarrow$ generalization
- Buchanan et al. (2025): memorization capacity for Gaussian mixtures
- Statistical consistency results (Oko et al., Azangulov et al., Tang et al., Lyu et al.)

**Connection:** echoes classical VC/Rademacher theory — generalization $\sim f(P/N)$

**Limitation**

These analyses typically assume **exact or near-exact minimization** of empirical risk.

In practice, optimization dynamics play a crucial role that capacity arguments alone cannot capture.

# Family 2: Implicit Regularization from Optimization

## Early Stopping

(Bonnaire et al., Favero et al.)

Memorization onset $\tau_{\text{mem}} \propto N$

For large datasets: memorization time exceeds any training budget

## Learning Rate

(Wu et al. 2025)

Large $\eta$ prevents convergence to sharp minima

*Minimum stability:* GD with $\eta$ can only reach minima with Hessian $\leq O(1/\eta)$

Connected to *edge of stability*

## Score Matching Bias

(Farghly et al., Li et al., Shen et al.)

Score matching itself favors manifold structure over individual points

Learning manifold support is easier than learning density on it

# Family 3: Inductive Bias of the Architecture

**Architecture shapes the function class:**

- Kadkhodaie & Simoncelli (2023): CNNs learn geometry-adaptive harmonic representations
- Kamb et al. (2024): convolutional weight-sharing and locality determine novelty capacity
- Cui et al. (2024): theoretical analysis of inductive biases in simple flow models

**Our Approach**

We **fix the architecture** (U-Net) to isolate the effects of:

- Dataset size
- Model size
- Batch size
- Learning rate

# Experimental Setup

**Data & Model:**

- CIFAR-10, first $N$ training samples
- U-Net architecture, rectified flow
- 250 sampling steps, no CFG
- 5,120 test samples for evaluation

**Metrics tracked:**

1. Denoising score matching loss
2. Memorization score
3. Sliced Wasserstein distance (`MIND`)
4. FID

**Hyperparameter sweeps:**

- $N \in \{\mathbf{2048}, 4096, 8192, 16384\}$
- $P \in \{3\text{M}, 8\text{M}, 26\text{M}, \mathbf{93M}, 201\text{M}\}$
- $B \in \{\mathbf{16}, 32, 64, 128\}$
- $\eta \in \{10^{-5}, 5{\cdot}10^{-5}, \mathbf{10^{-4}}, 4{\cdot}10^{-4}\}$

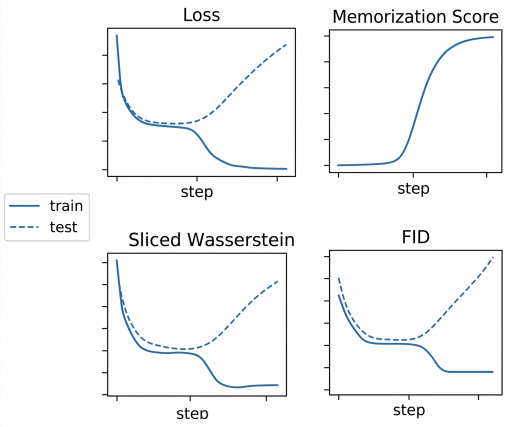
**Quantities of interest:**

- Memorization time $\tau_{\text{mem}}$
- Memorization rate (slope)
- Best fidelity (min `MIND` before $\tau_{\text{mem}}$)

# Prediction vs. Reality: Surprises!

**What we predicted:**

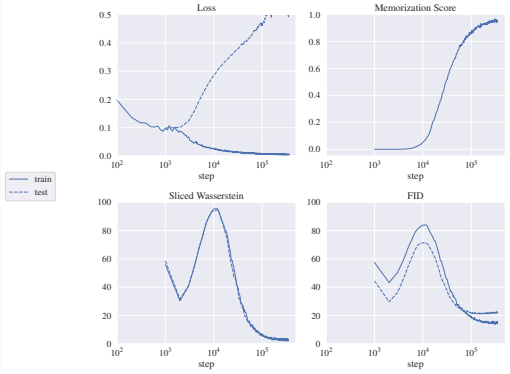
- Train/test loss decrease together until $\tau_{\text{mem}}$
- Then model overfits, memorization score rises
- Distributional distances (MIND, FID) track the loss



# Prediction vs. Reality: Surprises!

**What we observed:**

- ! Train–test gap in distributional distance is **uninformative**
- ! **Double descent** in MIND and FID
- ! Double descent for **both** train and test

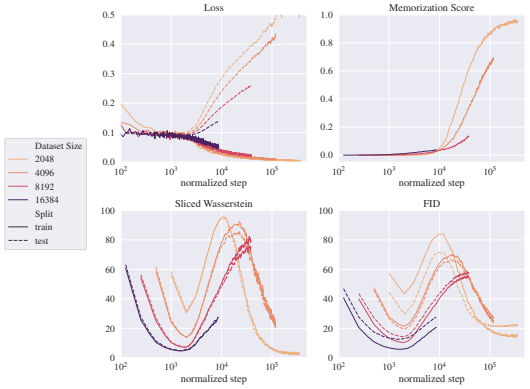


# Effect of Dataset Size

**Normalized step:** $T = \tau \cdot \frac{N_{\min}}{N} \cdot \frac{B}{B_{\min}} \cdot \frac{\eta}{\eta_{\min}}$

**When increasing $N$:**

1. **Curves collapse** before memorization
   $\Rightarrow$ confirms $\tau_{\text{mem}} \propto N$ (linear scaling)
2. **Memorization rate decreases**
   $\Rightarrow$ harder to memorize larger datasets
3. **Fidelity improves**
   $\Rightarrow$ lower minimum MIND with more data

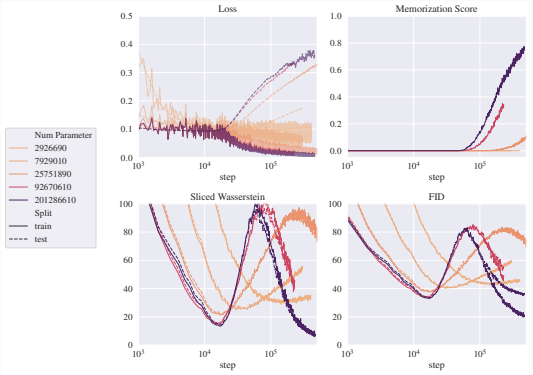


# Effect of Model Size

**When increasing model size $P$:**

1. **Memorization starts sooner**
   $\Rightarrow$ larger models memorize faster

2. **Peak fidelity improves**
   $\Rightarrow$ larger models achieve lower MIND before memorization




## Asymmetry with Dataset Size

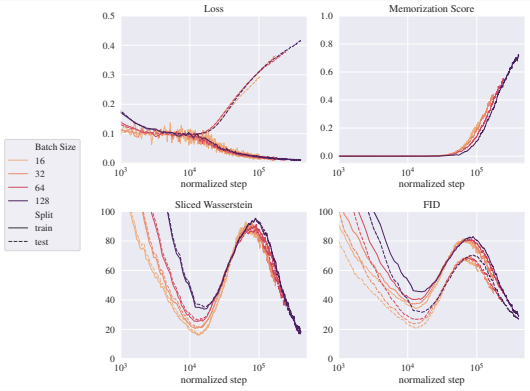
- More data: delays memorization, improves fidelity
- Bigger model: **accelerates** memorization, **also** improves fidelity

# Effect of Batch Size

**Normalized step:** $T = \tau \cdot \frac{N_{\min}}{N} \cdot \frac{B}{B_{\min}} \cdot \frac{\eta}{\eta_{\min}}$

**When increasing batch size $B$:**

1. **Normalized curves collapse**
   $\Rightarrow$ memorization onset $\propto 1/B$
2. **Memorization rate unchanged**
   $\Rightarrow$ *surprising*: more stochasticity doesn't slow memorization
3. **Smaller $B$ $\Rightarrow$ better fidelity**
   $\Rightarrow$ connects to minimum stability (flatter minima)




**But:** improved fidelity manifests *during training*, not at convergence — outside the scope of minimum stability theory.

# Effect of Learning Rate

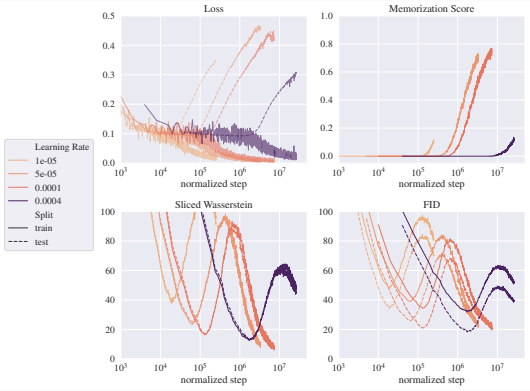
**Normalized step:** $T = \tau \cdot \frac{N_{\min}}{N} \cdot \frac{B}{B_{\min}} \cdot \frac{\eta}{\eta_{\min}}$

**Two effects when increasing $\eta$:**

1. **Normalized training time increases**
   $\Rightarrow$ consistent with edge-of-stability: beyond a threshold, larger $\eta$ doesn't accelerate training

2. **Larger $\eta \Rightarrow$ better fidelity**
   $\Rightarrow$ consistent with Wu et al. (2025)

Improvement manifests **along the trajectory**, not at convergence.
$\eta = 4 \times 10^{-4}$ is close to divergence.



# Summary of Findings: What We Know

1. **The memorization transition is systematic.**
   Observable with careful calibration of dataset size. Gradual, with well-defined $\tau_{\text{mem}}$.
2. **Memorization onset scales linearly with $N$.**
   Explains why practical models don't memorize. This question is largely resolved.
3. **Model size and dataset size play asymmetric roles.**
   Both improve fidelity, but dataset size *delays* while model size *accelerates* memorization. Classical $P/N$ ratios are insufficient.
4. **Optimization hyperparameters improve fidelity during training.**
   Smaller $B$, larger $\eta \Rightarrow$ better peak fidelity. Effects manifest along the trajectory, not at convergence.

# Open Questions

1. **What metrics capture novelty, fidelity, and their interaction?**
   - FID, MIND fail to distinguish copy-vs-generalize
   - Need metrics sensitive to this distinction, robust to dimensionality
   - What does "generalization" formally mean for generative models?

2. **How does implicit regularization operate along the trajectory?**
   - Minimum stability: characterizes convergence, not the path
   - Need theory for *intermediate iterates* of SGD on score matching
   - Promising direction: central flow theory (Cohen et al.)

3. **How does data geometry shape generalization?**
   - How do manifold structure & regularity interact with optimization?
   - Even linear models show nontrivial interactions
   - Exciting direction: text-conditioned models (unseen prompts)

# The Bigger Picture: Rethinking Generalization

**Supervised Learning:**

- Training loss $\approx$ prediction accuracy
- Generalization gap is the natural object
- Benign overfitting explains interpolation + generalization

**Generative Modeling:**

- Training loss (score matching) $\neq$ sample quality
- Train-test gap is **uninformative**
- Memorization and generalization are **incompatible**
- Need **new conceptual frameworks**

# Our Position

**Position Statement**

Understanding generalization in diffusion models requires going **beyond classical statistical learning theory** and the **benign overfitting** paradigm.

The question of *why* diffusion models do not memorize is **largely resolved**: early stopping + linear scaling of $\tau_{\text{mem}}$ with $N$.

The deeper question: **What is actually learned during the pre-memorization phase?** How do novelty and fidelity jointly emerge from optimization, architecture, and data geometry?

# Conclusion

**Generalization in Diffusion Models**

- Memorization $\neq$ generalization in generative models (unlike supervised)
- "Why don't models memorize?" is largely answered: $\tau_{\text{mem}} \propto N$
- Model size & dataset size play **asymmetric** roles
- Optimization shapes fidelity **during training**, not at convergence
- **Open:** proper metrics, trajectory-level theory, data geometry interactions