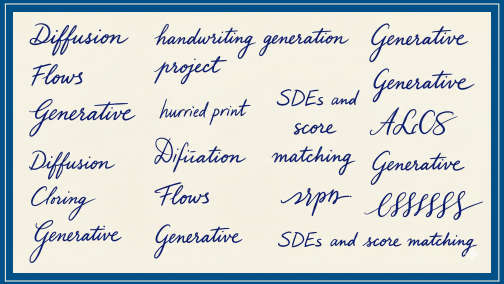
# Generative Handwriting via Conditional Diffusion & Flows

A Directed Reading Program Project
on SDEs, Score Matching, and Generative AI

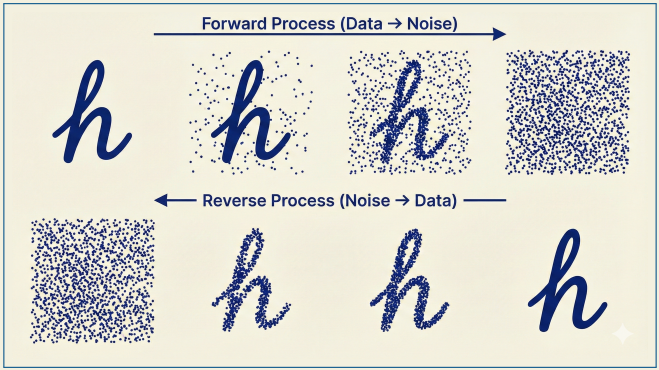
**Aditya Dutta**

# The Challenge




- **Goal:** Generate realistic, variable-length handwriting from text input.
- **Constraint:** Must mimic specific writing styles (User ID conditioning).
- **The Math Problem:** Modeling a high-dimensional distribution $p_{\text{data}}(x)$ given only samples.

# The Diffusion Process




**Forward:** Systematically destroy data with Gaussian noise.
**Reverse:** Learn to denoise step-by-step to recover data.

# Forward Process (Diffusion)

We define a fixed Markov chain that adds Gaussian noise to data $x_0$ using a schedule $\beta_t$.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I})$$

Data $\rightarrow$ Noise

**Key Property:** We can jump to any timestep $t$ in closed form:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

# Reverse Process (Generation)

Noise $\rightarrow$ Data

We approximate the intractable posterior $q(x_{t-1}|x_t)$ with a model $p_\theta$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

In practice, we train a U-Net to predict the **noise** $\epsilon$ that was added, effectively learning to point "backwards" towards the data manifold.

# Training Objective: The ELBO

Maximizing the Evidence Lower Bound (ELBO) simplifies to a weighted MSE between actual noise $\epsilon$ and predicted noise $\epsilon_\theta$.

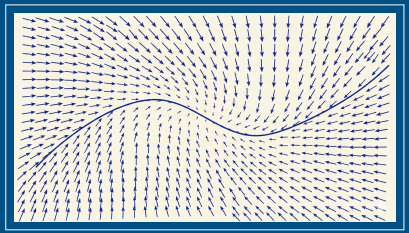
$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]$$

*"Given a noisy image, guess the noise that corrupted it."*

# Continuous Time: SDEs & Flows

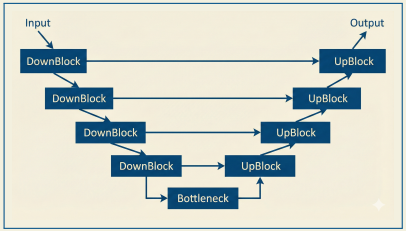
As $\Delta t \to 0$, the process becomes an SDE. Associated with it is a deterministic **Probability Flow ODE**:

$$dx = \left[ f(x,t) - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x) \right] dt$$



- $\nabla_x \log p_t(x)$ is the **Score Function**.
- Model learns score: $\epsilon_\theta \approx -\sigma \nabla \log p$.
- This allows for deterministic sampling (Flows).

# Architecture: The U-Net




We use a 2D U-Net conditioned on time and style.

- **Downsampling:** Compresses image to capture context.
- **Bottleneck:** Deepest understanding of content.
- **Upsampling:** Reconstructs fine ink details.
- **Skip Connections:** Preserves spatial structure.

# Conditioning the Flow

**Text Conditioning** Frozen

BERT encoder.
Injected via **Cross-Attention**.

*(Controls WHAT to write)*

**Style Conditioning** CNN Style

Encoder.
Added to **Time Embeddings**.

*(Controls HOW to write)*

# Implementation Details

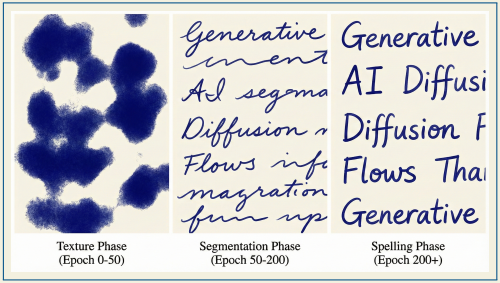
- **Dataset:** IAM Handwriting Database.
- **Tools:** PyTorch, Diffusers, Accelerate.
- **Optimization:** Mixed Precision (FP16) & Metal (MPS).

### The VRAM Crash

**Problem:** Attention is $O(N^2)$. High-res input caused VRAM explosion.

**Solution:** Swapped initial layers to 'DownBlock2D' (Convolutions only).

# Training Dynamics




1. **Texture Phase (Epoch 0-50):**
   Learns ink density and stroke width.
2. **Segmentation Phase (Epoch 50-200):**
   Learns to break waves into words.
3. **Spelling Phase (Epoch 200+):**
   Cross-attention aligns text to strokes.

# Conclusion & Future Work

**Summary** Successfully implemented a conditional diffusion model approximating $p_{\text{data}}(x)$, bridging SDE theory with practical engineering.

**Future Directions** Flow Matching (regress vector field directly) and Consistency Models (1-step generation).

# Sources & References


- Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*.
- Song, J., & Ermon, S. (2019). *Generative Modeling by Estimating Gradients of the Data Distribution*.
- Song, J., Sohl-Dickstein, J., et al. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*.
- Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.
- IAM Handwriting Database – University of Bern.
- Various Diffusion & Flow references from Lilian Weng's "Understanding Diffusion Models".