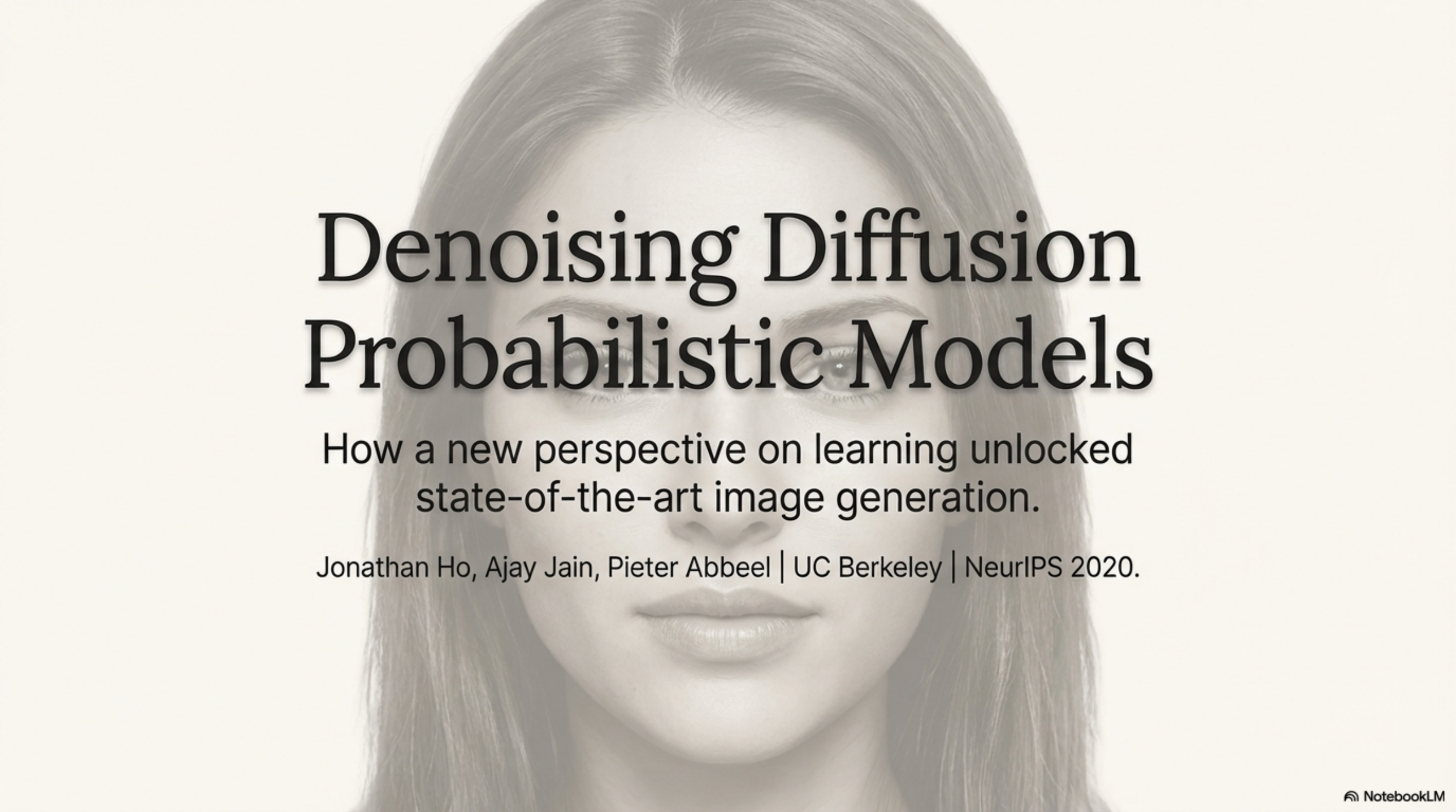


Denoising Diffusion Probabilistic Models

How a new perspective on learning unlocked state-of-the-art image generation.

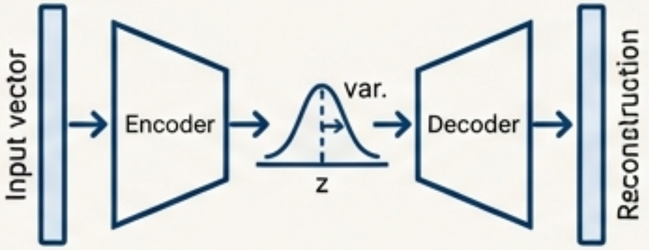
Jonathan Ho, Ajay Jain, Pieter Abbeel | UC Berkeley | NeurIPS 2020.

The Generative Landscape Circa 2020

Generative Adversarial Networks (GANs)

Unmatched sample quality but known for training instability.

Variational Autoencoders (VAEs)



Stable training and **principled likelihood-based approach**, but often producing blurrier samples.

Autoregressive Models & Flows

Excellent log-likelihoods but computationally intensive sampling.

An Overlooked Contender

“Diffusion probabilistic models, first introduced in 2015, were a known class of latent variable models. However, to the best of our knowledge, there has been no demonstration that they are capable of generating high quality samples.”

A Breakthrough in Image Quality from an Unlikely Source

Prior Perception

A theoretical curiosity, not competitive on high-fidelity image synthesis.

The 2020 Result

State-of-the-art FID Score: **3.17** on unconditional CIFAR10

"Sample quality sometimes better than published results on other types of generative models."

Inception Score: **9.46**

"Evoiiriting in most eamerent forencen, and dealmarter broach."

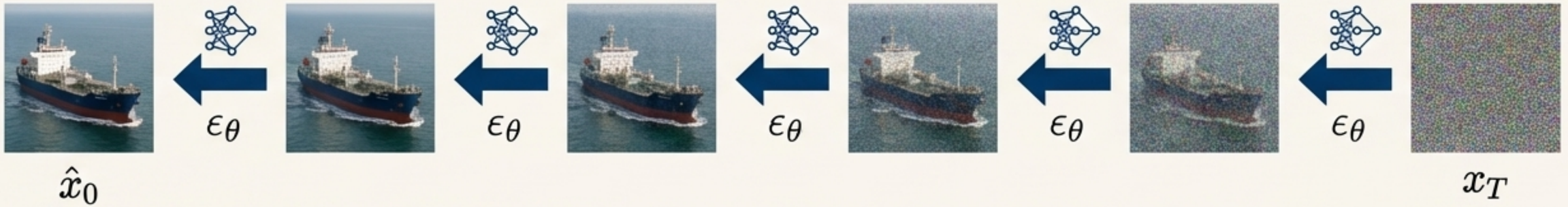
How did this paper transform diffusion models into a **state-of-the-art method?**

The Core Idea: Systematically Corrupt, then Learn to Reverse

Part 1: The Forward Process (Fixed)

A fixed process that gradually adds Gaussian noise, destroying the data.

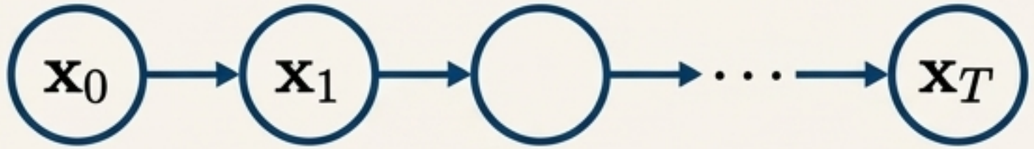
Part 2: The Reverse Process (Learned)



A learned process that reverses the corruption to generate data.

The Forward Process: A Fixed Markov Chain of Noise Addition

The forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ gradually adds Gaussian noise to an image $\mathbf{x}_0$ over T timesteps according to a fixed variance schedule β_t.



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right),$$

At each step t, we scale the previous image and add a small amount of Gaussian noise.

A remarkable property is that we can sample $\mathbf{x}_t$ at any timestep t directly from $\mathbf{x}_0$ in a single step.

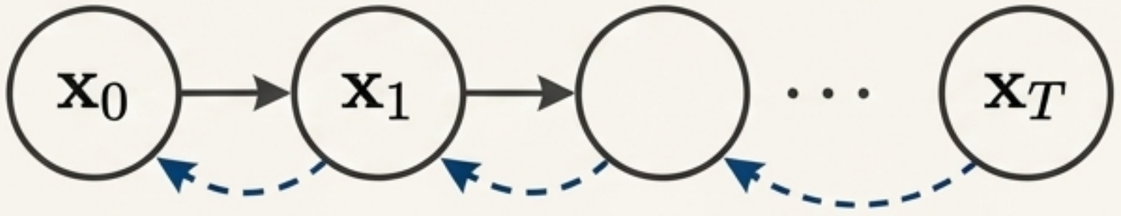
$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

where $\bar{\alpha}_t = \prod_{s=1..t}^{t}(1-\beta_s)$.

This closed-form expression allows us to randomly sample any noisy version of an image during training without iterating, making the process highly efficient.

The Reverse Process: Learning to Denoise

The reverse process $p_\theta(\mathbf{x}_{0:T})$ is a learned Markov chain that starts with pure noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it to produce a sample $\mathbf{x}_0$.



Learned Reverse Process $p_\theta(\mathbf{x}_{0:T})$

The Goal: We need to learn the transition probabilities $p_\theta(x_{t-1}|x_t)$. The paper models these as Gaussians:

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

The Learning Problem: The core task is to train a neural network to predict the mean $\boldsymbol{\mu}_\theta(x_t, t)$ of the denoised image at step t-1, given the noisy image x_t.

The Variational Bound: Training optimizes the variational bound L on the negative log likelihood. The key term for training the reverse process at each step t is a KL divergence:

$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t))$$

This term compares our learned reverse step to the true posterior of the forward process, which is known and tractable.

The Key Insight: Stop Predicting the Image, Start Predicting the Noise

The Conventional Approach

The most direct way to parameterize the mean $\mu_\theta(x_t, t)$ is to train a network to predict the true posterior mean $\tilde{\mu}_t$.

The loss term looks like this: $L \approx ||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2$. This task is complex.

↓ The Breakthrough Reparameterization

The Breakthrough Reparameterization

The authors re-parameterize the mean μ_θ in terms of a function ϵ_θ that predicts the noise ϵ added at that step.

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t) \right)$$

The Simplified Objective

This reparameterization dramatically simplifies the loss term into a simple mean-squared error between the true and predicted Gaussian noise.

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

Why it Works: This simplified objective re-weights the loss to focus the network on more difficult denoising tasks at larger t... leading to better sample quality.

The Power of Predicting Noise: An Ablation Study

Objective Type	Training Objective	FID Score (Lower is better)	Inception Score
$\tilde{\mu}$ **prediction (Baseline)**	Variational Bound `$\mathcal{L}$`	13.22	8.06
	$\|\tilde{\mu} - \tilde{\mu}_\theta\|^2$	Unstable / Poor Samples	-
ε **prediction (Ours)**	Variational Bound `$\mathcal{L}$`	13.51	7.67
	$\|\varepsilon - \varepsilon_\theta\|^2$ ($\mathcal{L}_{\text{simple}}$)	**3.17**	**9.46**

Predicting `$\varepsilon$` with the simplified objective ($\mathcal{L}_{\text{simple}}$) is not just an alternative—it's the key that unlocked state-of-the-art performance, dramatically outperforming the more intuitive baseline of predicting the denoised image mean.

The Result: A New State of the Art in Image Generation

Key Metrics

Unconditional CIFAR10

FID: **3.17** (better than most class-conditional models at the time)

IS: **9.46**

Visual Gallery

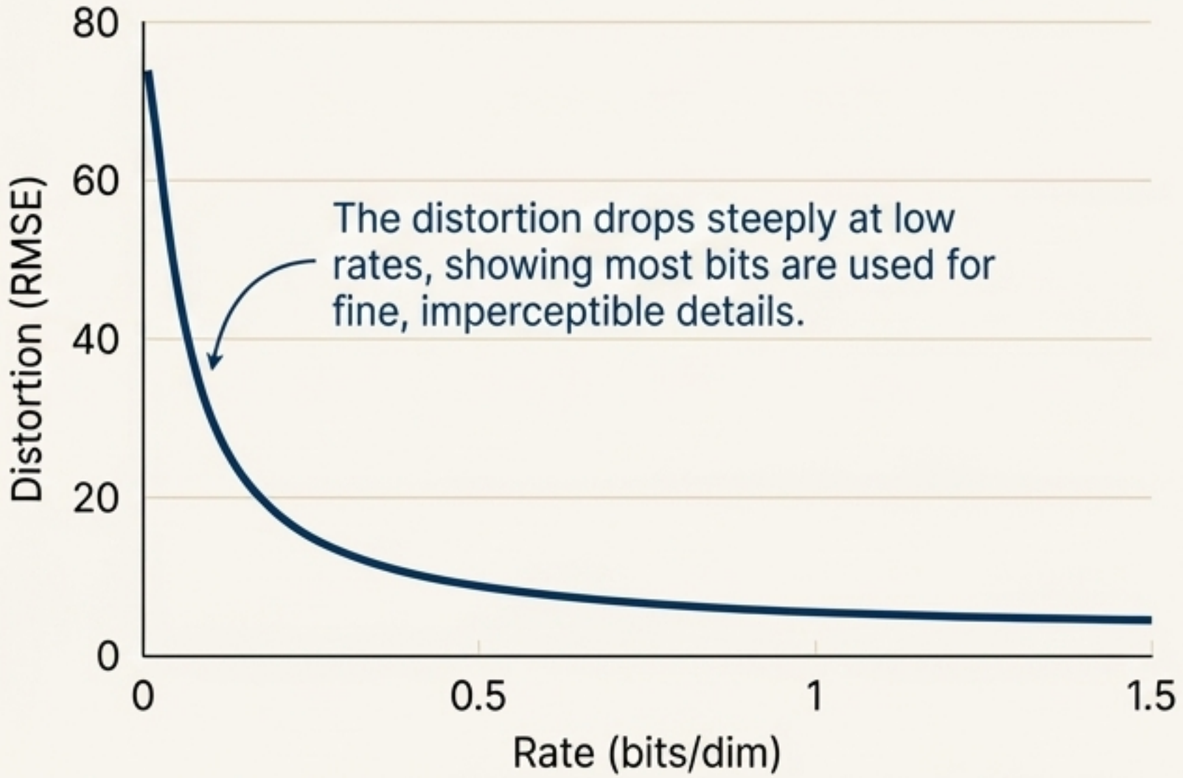
LSUN Church (FID 7.89)

LSUN Bedroom (FID 4.90)

More Than a Generator: An Excellent Lossy Compressor

The model's high sample quality, despite non-competitive log likelihoods, suggests a strong inductive bias for lossy compression.

For the highest quality CIFAR10 model, the rate is **1.78 bits/dim** and distortion is **1.97 bits/dim**. More than half of the lossless codelength describes imperceptible distortions.

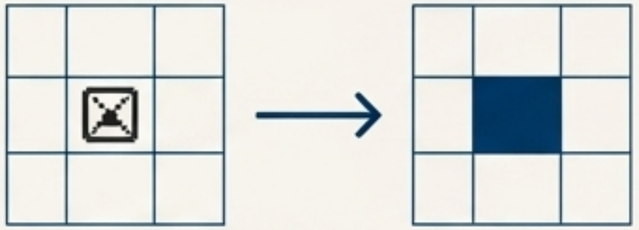


Progressive generation reveals the model's structure: large-scale features appear first, with fine details emerging last.

A Connection to Autoregressive Decoding

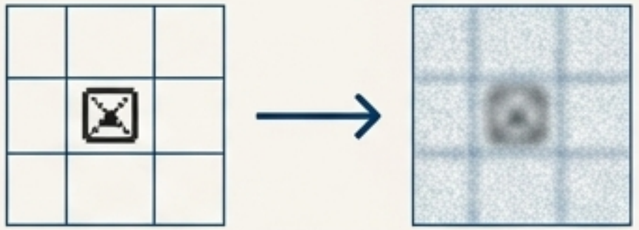
The diffusion process can be interpreted as a type of autoregressive model with a generalized bit ordering.

Standard Autoregressive Model



Predicts masked pixels in a fixed order.

Gaussian Diffusion Model



Denoises all pixels simultaneously—a generalized ordering.

Consider a diffusion process that masks one pixel at a time. Training the reverse process to predict the masked pixel is equivalent to training a standard autoregressive model.

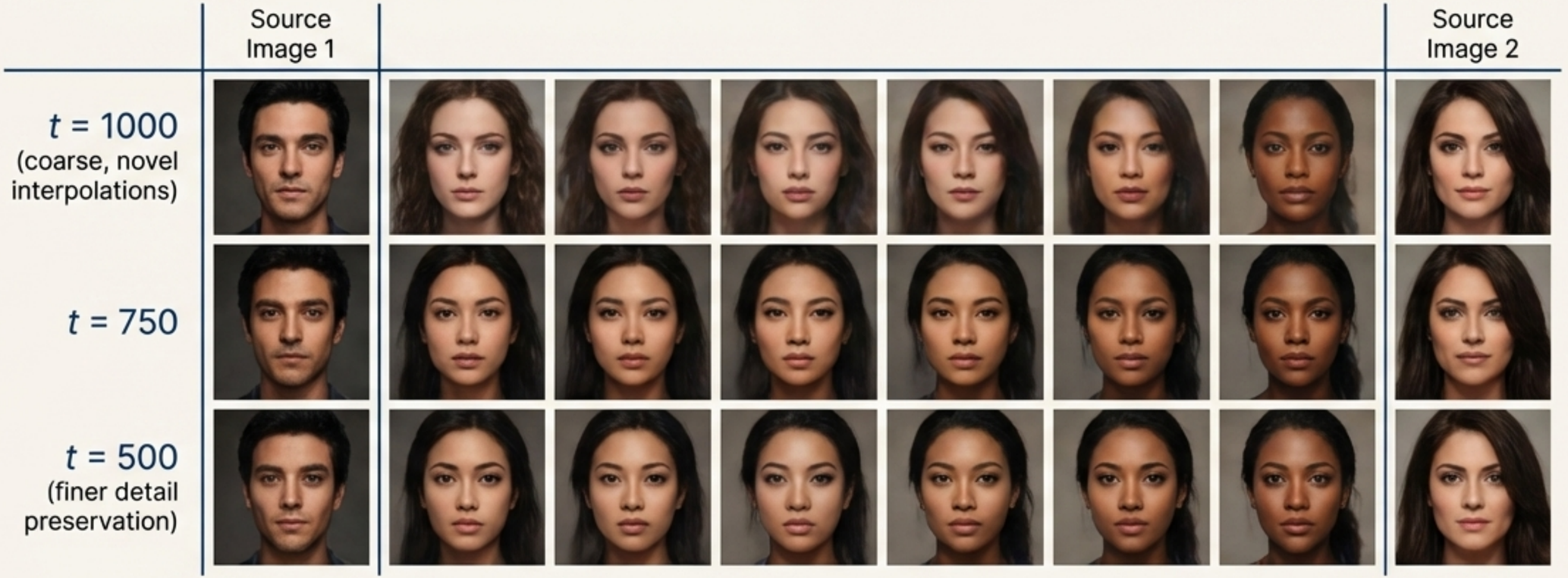
The Gaussian diffusion model we use is analogous, but with a generalized genealized “ordering” that isn’t tied to pixel coordinates. Gaussian noise is a more natural corruption for images than masking.

Benefit

This generalized ordering provides a powerful inductive bias for images. The diffusion length T is also a flexible hyperparameter, not fixed to the data dimension, allowing a trade-off between sampling speed and model expressiveness.

Exploring the Latent Space: Smooth and Meaningful Interpolations

We can encode two source images (x_0, x'_0) into noisy latents (x_t, x'_t), linearly interpolate between them, and then decode back to image space using the reverse process.



The reverse process produces high-quality reconstructions and plausible interpolations that smoothly vary attributes like pose, skin tone, and hairstyle. Interpolating at larger *t* results in coarser, more varied interpolations, creating novel samples at *t*=1000.

The Legacy: A Paradigm Shift for Generative Models

1. **Unlocked SOTA Quality**: Demonstrated for the first time that diffusion models could achieve state-of-the-art sample quality.

2. **A New Learning Objective**: Showed that parameterizing the model to predict noise (ε`) with a simplified objective was the key to this success.

3. **Revealed Deeper Connections**: Established an explicit link between diffusion models, denoising score matching, and annealed Langevin dynamics.

The Impact

This work transformed diffusion models from a theoretical curiosity into a dominant, practical, and highly effective tool for generative modeling, setting the stage for the next wave of high-fidelity image synthesis.