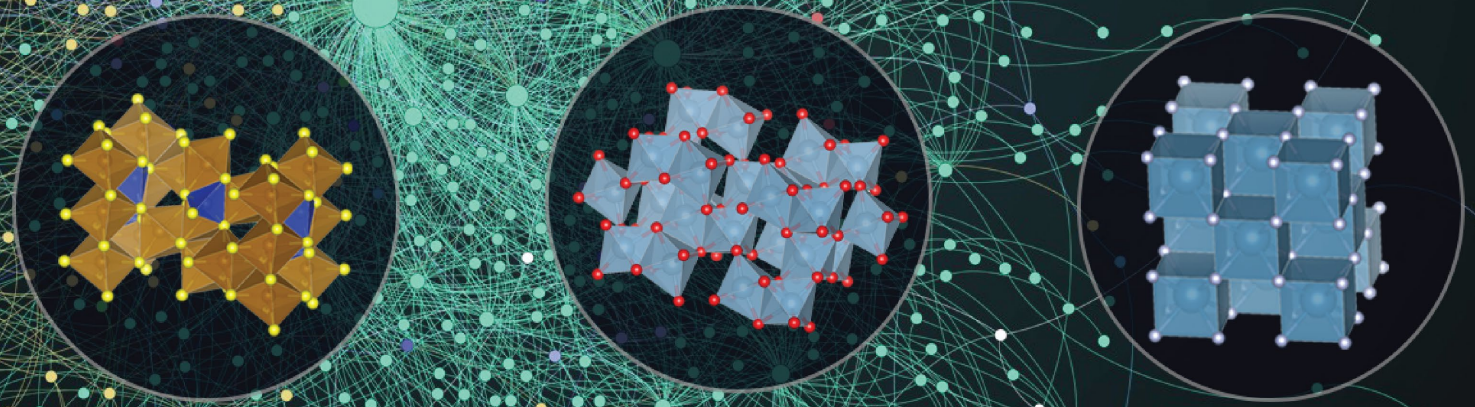
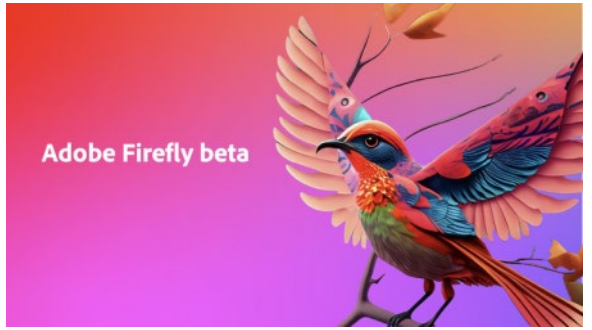
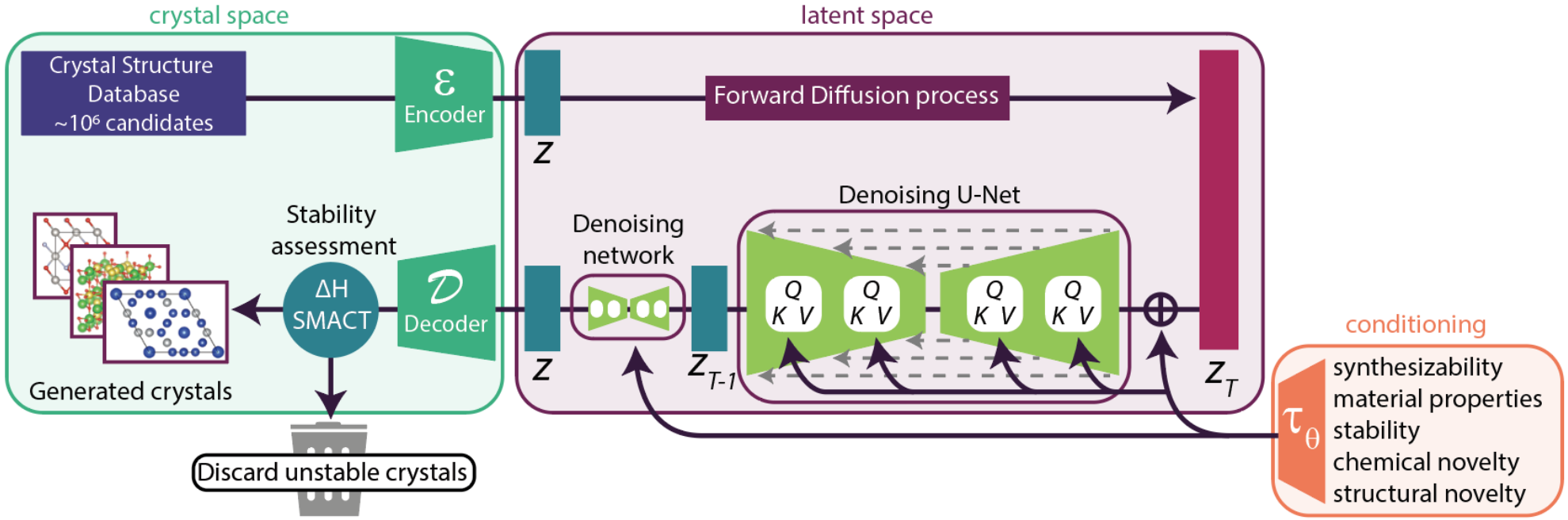
Diffusion models



Diffusion models are responsible for the most exciting AI-generated art





Diffusion models are recently being applied to materials research as well!

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2024, **3**, 62

Generative adversarial networks and diffusion models in material discovery

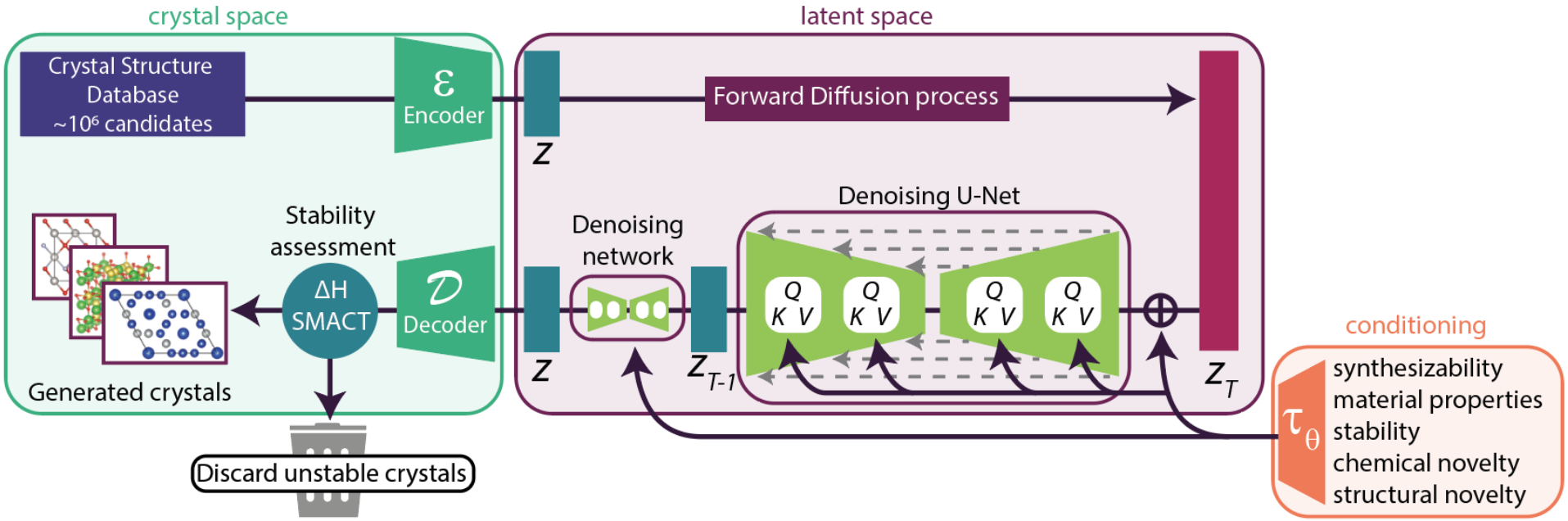
Michael Alverson, *[ac] Sterling G. Baird,[a] Ryan Murdock,[a] (Enoch) Sin-Hang Ho,[b] Jeremy Johnson [b] and Taylor D. Sparks [a]

The idea of materials discovery has excited and perplexed research scientists for centuries. Several different methods have been employed to find new types of materials, ranging from the arbitrary replacement of atoms in a crystal structure to advanced machine learning methods for predicting entirely new crystal structures. In this work, we pursue three primary objectives. (I) Introduce CrysTens, a crystal encoding that can be use
relative perform
innovative and e
(III) Show that th
symmetrical an
accomplish the
Pearson's Cryst

Received 24th July 2023
Accepted 30th November 2023

DOI: 10.1039/d3dd00137g

rsc.li/digitaldiscovery



Diffusion models were described by 4 key papers

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein JASCHA@STANFORD.EDU
Stanford University
Eric A. Weiss EAWEISS@BERKELEY.E
University of California, Berkeley
Niru Maheswaranathan NIRUM@STANFORD.E
Stanford University
Surya Ganguli SGANGULI@STANFORD.E
Stanford University

Abstract

A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps, as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference implementation of the algorithm.

1. Introduction

Historically, probabilistic models suffer from a tradeoff between two conflicting objectives: *tractability* and *flexibility*. Models that are *tractable* can be analytically evaluated and easily fit to data (e.g. a Gaussian or Laplace). However, these models are unable to aptly describe structure in datasets. On the other hand, models that are *flexible* can be molded to fit structure in arbitrary data. For example, we can define models in terms of any (non-negative) function $\phi(\mathbf{x})$ yielding the flexible distribution $p(\mathbf{x}) = \frac{\phi(\mathbf{x})}{Z}$. However, computing Z is a normalization constant. However, computing this normalization constant is generally intractable. Evaluating, training, or drawing samples from such flexible models typically requires a very expensive Monte Carlo process.

A variety of analytic approximations exist which can be fast, but do not remove, this tradeoff–for instance, mean field theory and its expansions (T, 1982; Tanaka, 1999), variational Bayes (Jordan et al., 1999), contrastive divergence (Welling & Hinton, 2002; Hinton, 2002), minimum probability flow (Sohl-Dickstein et al., 2011b;a), minimum KL contraction (Lyu, 2011), proper scoring rules (Gneiting & Raftery, 2007; Parry et al., 2012), score matching (Hyvärinen, 2005), pseudolikelihood (Besag, 1975), loopy belief propagation (Murphy et al., 1999), and many, many more. Non-parametric methods (Gershman & Blei, 2012) can also be very effective[1].

1.1. Diffusion probabilistic models

We present a novel way to define probabilistic models that allows:

1. extreme flexibility in model structure,
2. exact sampling,

[1]Non-parametric methods can be seen as transitioning smoothly between tractable and flexible models. For instance, a non-parametric Gaussian mixture model will represent a finite amount of data using a single Gaussian, but may represent infinite data as a mixture of an infinite number of Gaussians.

Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Denoising Diffusion Probabilistic Models

Jonathan Ho UC Berkeley jonathanho@berkeley.edu
Ajay Jain UC Berkeley ajayj@berkeley.edu
Pieter Abbeel UC Berkeley pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion.

1 Introduction

Deep generative models of all kinds have recently exhibited high quality samples in a wide variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [14, 27, 3, 58, 38, 25, 10, 32, 44, 57, 26, 33, 45], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 55].

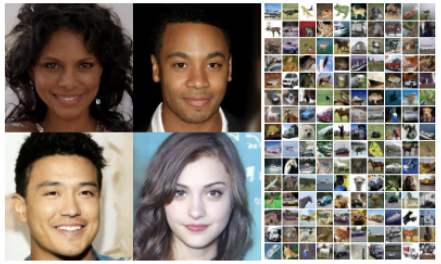


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Improved Denoising Diffusion Probabilistic Models

Alex Nichol [*1] **Prafulla Dhariwal** [*1]

Abstract

Denoising diffusion probabilistic models (DDPM) are a class of generative models which have recently been shown to produce excellent samples. We show that with a few simple modifications, DDPMs can also achieve competitive log-likelihoods while maintaining high sample quality. Additionally, we find that learning variances of the reverse diffusion process allows sampling with an order of magnitude fewer forward passes with a negligible difference in sample quality, which is important for the practical deployment of these models. We additionally use precision and recall to compare how well DDPMs and GANs cover the target distribution. Finally, we show that the sample quality and likelihood of these models scale smoothly with model capacity and training compute, making them easily scalable. We release our code at https://github.com/openai/improved-diffusion.

1. Introduction

Sohl-Dickstein et al. (2015) introduced diffusion probabilistic models, a class of generative models which match a data distribution by learning to reverse a gradual, multi-step noising process. More recently, Ho et al. (2020) showed an equivalence between denoising diffusion probabilistic models (DDPM) and score based generative models (Song & Ermon, 2019; 2020), which learn a gradient of the log-density of the data distribution using denoising score matching (Hyvärinen, 2005). It has recently been shown that this class of models can produce high-quality images (Ho et al., 2020; Song & Ermon, 2020; Jolicoeur-Martineau et al., 2020) and audio (Chen et al., 2020b; Kong et al., 2020), but it has yet to be shown that DDPMs can achieve log-likelihoods competitive with other likelihood-based models such as autoregressive models (van den Oord et al., 2016c) and VAEs (Kingma & Welling, 2013). This raises various questions, such as whether DDPMs are capable of capturing all the modes of a distribution. Furthermore, while Ho et al. (2020) showed extremely good results on the CIFAR-10 (Krizhevsky, 2009) and LSUN (Yu et al., 2015) datasets, it is unclear how well DDPMs scale to datasets with higher diversity such as ImageNet. Finally, while Chen et al. (2020b) found that DDPMs can efficiently generate audio using a small number of sampling steps, it has yet to be shown that the same is true for images.

In this paper, we show that DDPMs can achieve log-likelihoods competitive with other likelihood-based models, even on high-diversity datasets like ImageNet. To more tightly optimise the variational lower-bound (VLB), we learn the reverse process variances using a simple reparameterization and a hybrid learning objective that combines the VLB with the simplified objective from Ho et al. (2020).

We find surprisingly that, with our hybrid objective, our models obtain better log-likelihoods than those obtained by optimizing the log-likelihood directly, and discover that the latter objective has much more gradient noise during training. We show that a simple importance sampling technique reduces this noise and allows us to achieve better log-likelihoods than with the hybrid objective.

After incorporating learned variances into our model, we surprisingly discovered that we could sample in fewer steps from our models with very little change in sample quality. While DDPM (Ho et al., 2020) requires hundreds of forward passes to produce good samples, we can achieve good samples with as few as 50 forward passes, thus speeding up sampling for use in practical applications. In parallel to our work, Song et al. (2020a) develops a different approach to fast sampling, and we compare against their approach, DDIM, in our experiments.

While likelihood is a good metric to compare against other likelihood-based models, we also wanted to compare the distribution coverage of these models with GANs. We use the improved precision and recall metrics (Kynkäänniemi et al., 2019) and discover that diffusion models achieve much higher recall for similar FID, suggesting that they do indeed cover a much larger portion of the target distribution.

Finally, since we expect machine learning models to consume more computational resources in the future, we evaluate the performance of these models as we increase model size and training compute. Similar to (Henighan et al.,

[*]Equal contribution [1]OpenAI, San Francisco, USA. Correspondence to: <alex@openai.com>, <prafulla@openai.com>.

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal* OpenAI prafulla@openai.com
Alex Nichol* OpenAI alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

1 Introduction

Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

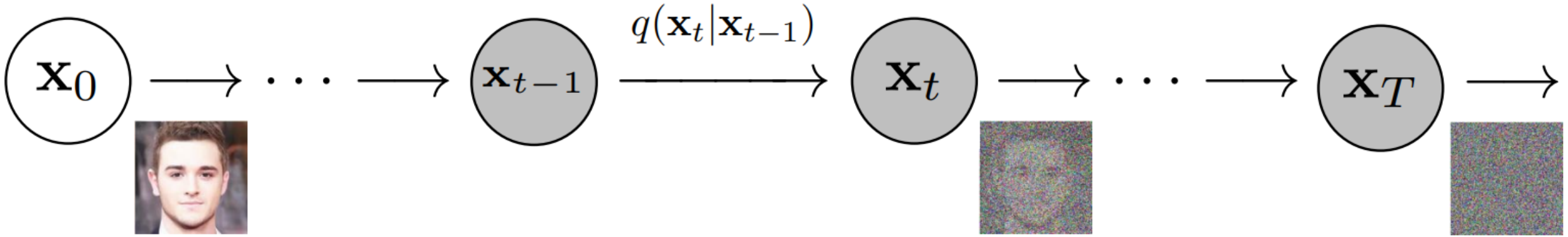
Over the past few years, generative models have gained the ability to generate human-like natural language [6], infinite high-quality synthetic images [5, 28, 51] and highly diverse human speech and music [64, 13]. These models can be used in a variety of ways, such as generating images from text prompts [72, 50] or learning useful feature representations [14, 7]. While these models are already

[*]Equal contribution

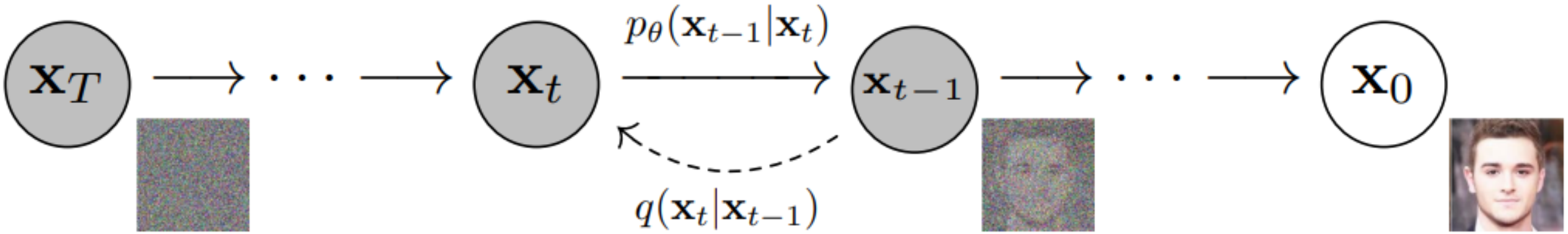
[#1 https://arxiv.org/pdf/1503.03585.pdf](https://arxiv.org/pdf/1503.03585.pdf)
[#2 https://arxiv.org/pdf/2006.11239.pdf](https://arxiv.org/pdf/2006.11239.pdf)
[#3 https://arxiv.org/pdf/2102.09672.pdf](https://arxiv.org/pdf/2102.09672.pdf)
[#4 https://arxiv.org/pdf/2105.05233.pdf](https://arxiv.org/pdf/2105.05233.pdf)

Diffusion models work in a different way than GANs, VAEs, and other generative models

First, we destroy data through successive addition of Gaussian noise…

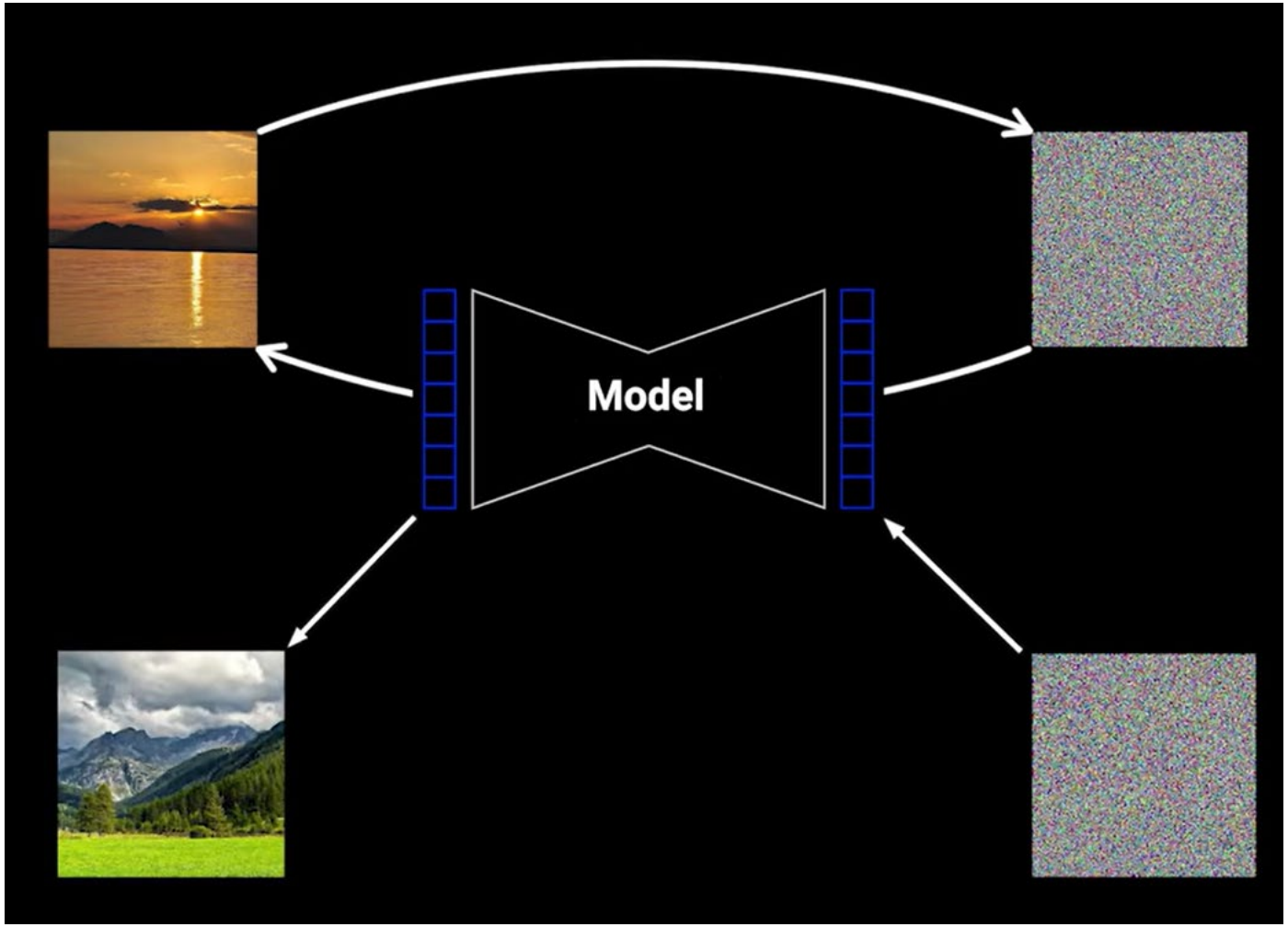


… and then we learn to recover the image by reversing the process

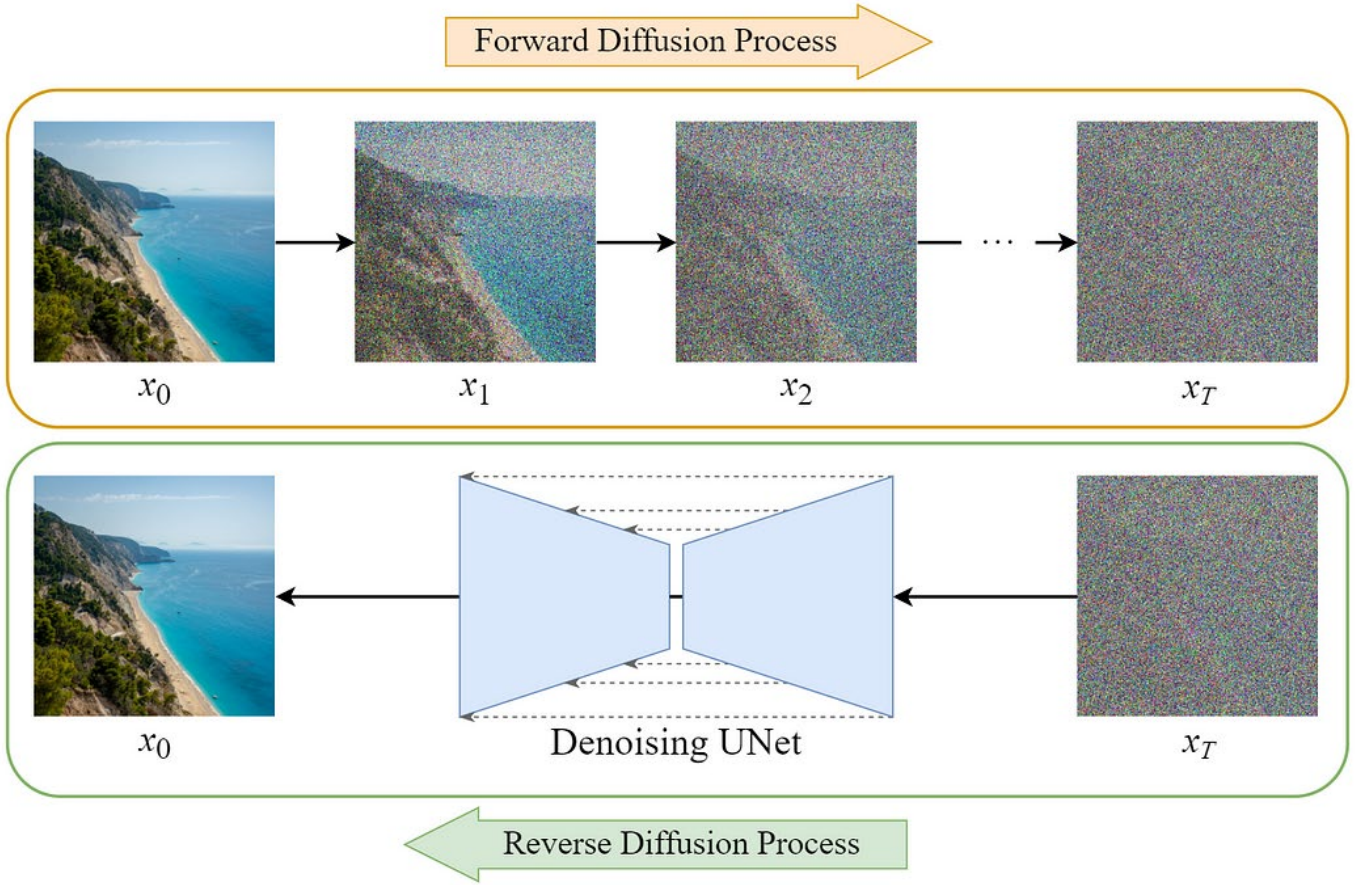


https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/

The Model must learn to go from noise to image, but not all at once



The input noise and output image are the same size, let's use a U-Net!

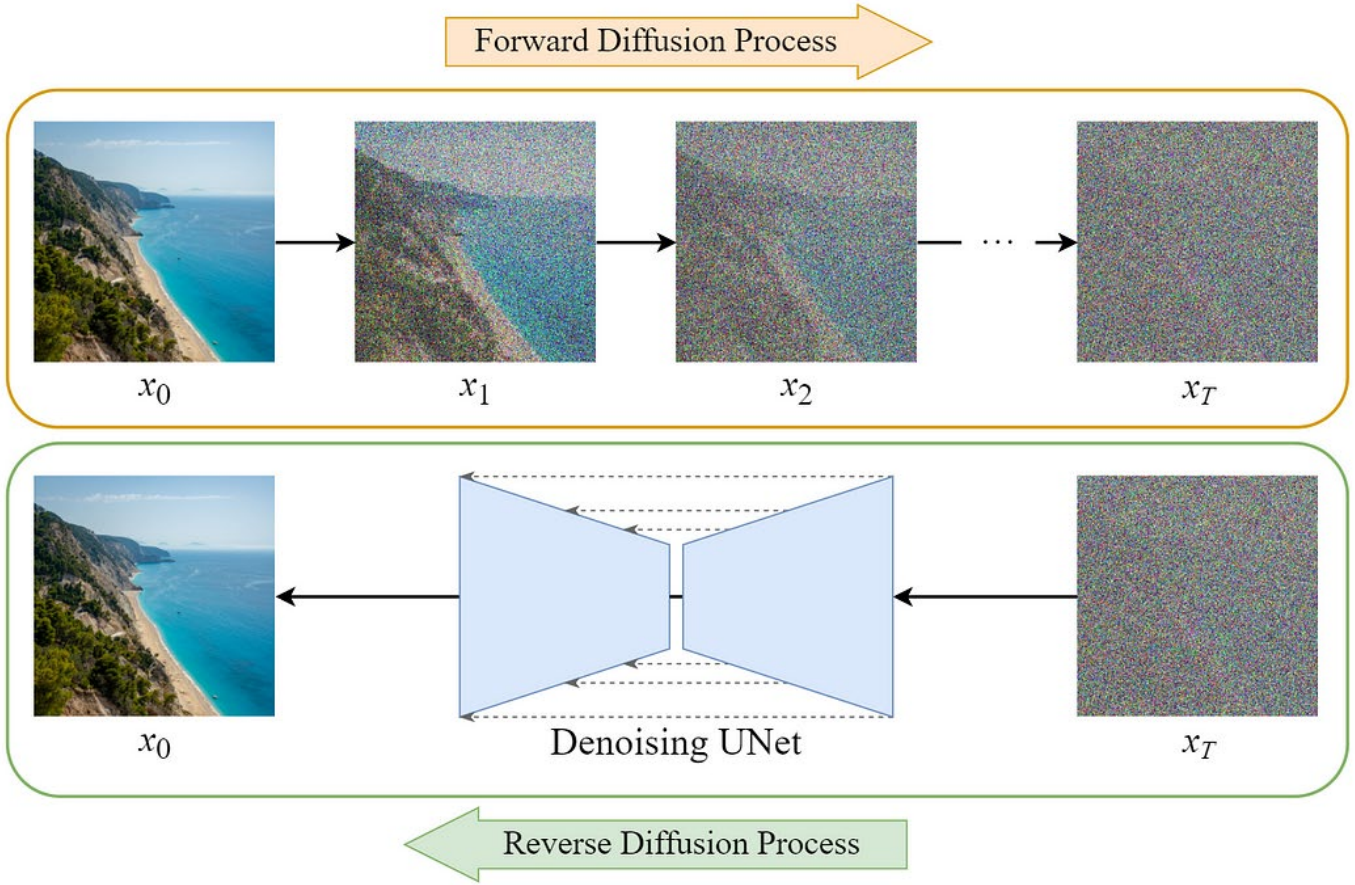


Current version of the image and the time are the inputs

The model learns the time step via sinusoidal embeddings

https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e

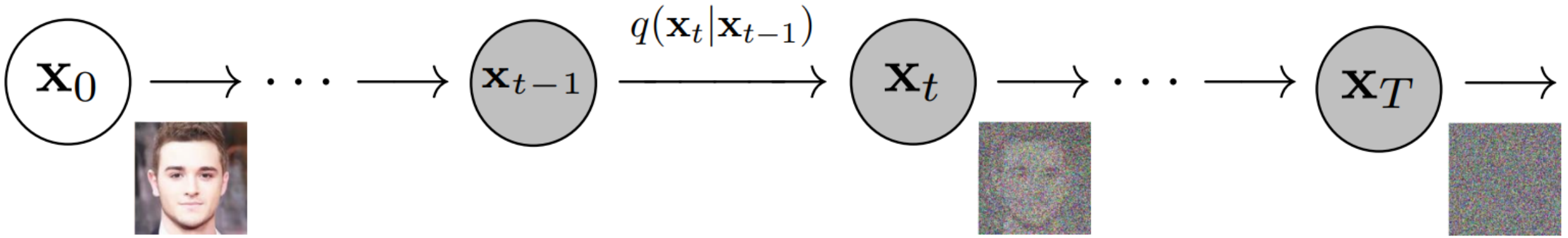
Mathematically, it's easier if we just predict the noise itself!



If this were just noise (not a beach+noise), we could learn the noise added during forward process

How AI Image Generators Work (Stable Diffusion / Dall-E), Computerphile, YouTube

The forward process is modeled as a Markov chain



We gradually add Gaussian noise over a fixed number of steps, T
Our initial data point, x_0, is transformed by this process $\{x_0, x_{t-1}, x_t, \dots, x_T\}$

If we look at a transformation we get

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t, \qquad \epsilon_t \sim N(0,1)$$

Here β_t are pre-defined variance schedules (small changes) that control how much noise gets added at each step, and ϵ_t is noise sampled from a Gaussian distribution

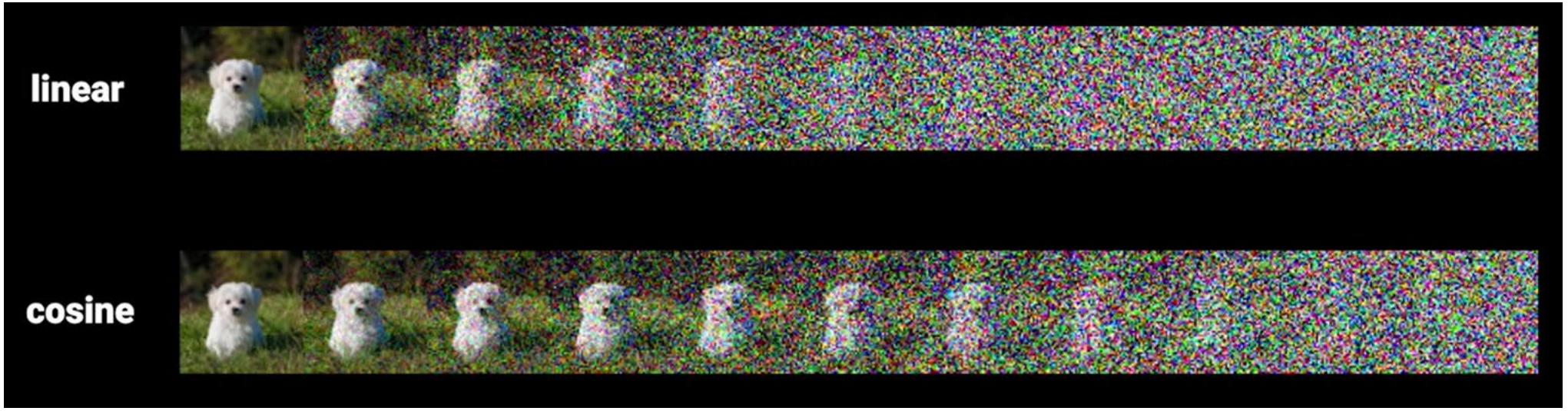
Variance schedules are very important in this process

β_t is not a single value, but a series of values for each step in the process $\{\beta_1, \ldots, \beta_T\}$

β_t values are typically small, but get larger over time and can be either fixed or learned

The cumulative effect of these t values determines the overall noise level at each step.

If chosen correctly, the variance schedule ensures that x_T is isotropic Gaussian for a sufficiently large value of T



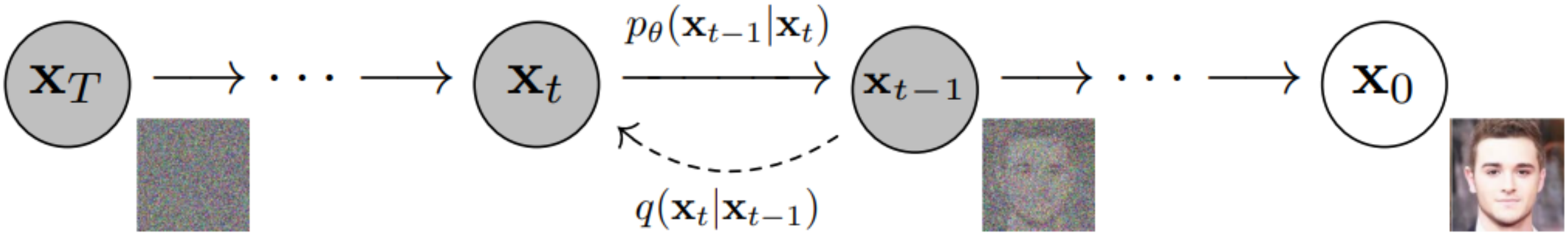
The reverse process tries to recover the original data

The steps for de-noising are as follows

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

Now $\epsilon_t(x_t, t)$ is the noise predicted by the model (parameterized by θ)

$$\alpha_t = 1 - \beta_t$$



We train the diffusion model based on the noise

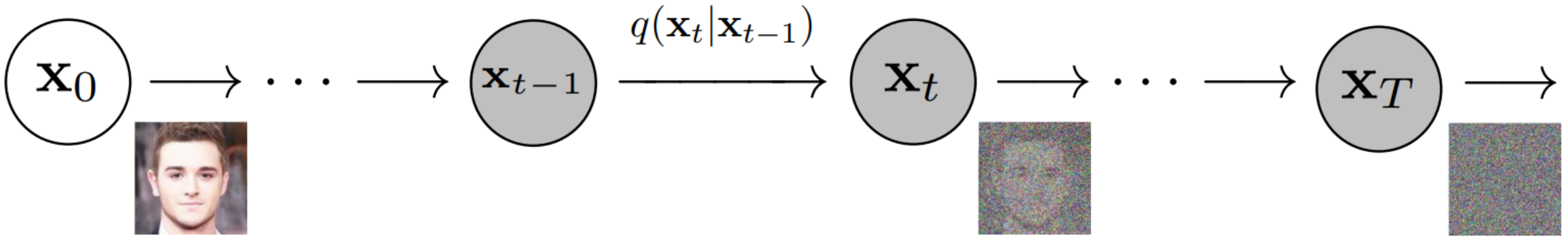
The training objective is to minimize the difference between the true noise ϵ_t and the noise predicted by the model $\epsilon_\theta(x_t, t)$.

This can be done using a loss function like the mean squared error (MSE):

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,\epsilon_t}\left[\left\|\epsilon_t - \epsilon_\theta\left(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t, t\right)\right\|^2\right]$$

Where $\alpha_t = \prod_{s=1}^{T}(1-\beta_s)$

Diffusion models are often described in a probabilistic framework



$$\{x_0, x_{t-1}, x_t, \dots, x_T\}$$

Formally, the distribution of x_t given x_{t-1} is a Gaussian

$$x_t|x_{t-1} \sim N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$$

This expresses x_t as a conditional Gaussian with mean $\sqrt{1-\beta_t}x_{t-1}$ and variance $\beta_t I$ where I is the identity matrix.

Entire sequence becomes a joint distribution

$$p(x_0, x_1, \dots, x_T) = p(x_0)\prod_{t=1}^{T} p(x_t|x_{t-1})$$

Rather than doing this step by step for many steps, we can do a math trick in one step!

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} a_s$$

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

$$= \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon$$

$$= \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \epsilon$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\, \epsilon$$

$$= \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}\, x_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\, \epsilon$$

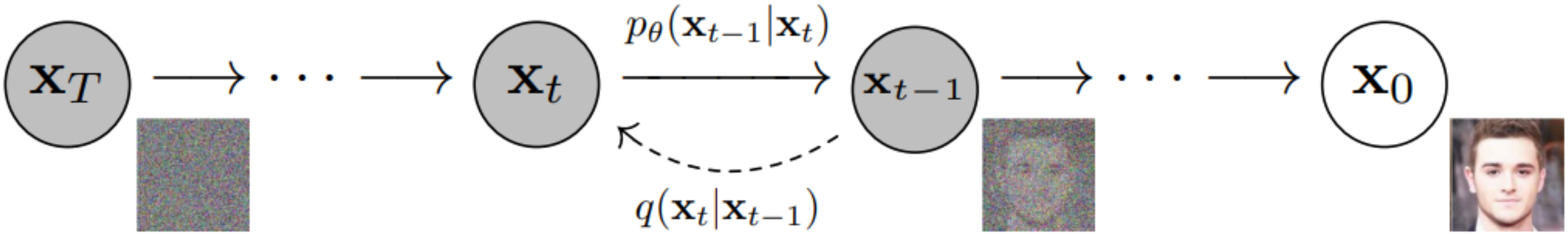
$$= \sqrt{\alpha_t\alpha_{t-1}\ldots\alpha_1\alpha_0}\, x_0 + \sqrt{1-\alpha_t\alpha_{t-1}\ldots\alpha_1\alpha_0}\, \epsilon$$

$$\boxed{= \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon}$$

Denoising in probabilistic framework becomes

$$p(x_{t-1}|x_t) \sim N\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta (x_t, t)\right)$$

Now $\mu_\theta(x_t, t)$ and $\sum_\theta(x_t, t)$ are the mean and the covariance parameterized by the model



Training in terms of distributions becomes variational lower bound

Maximizing the evidence lower bound (ELBO) of the log likelihood of the data, $\log(p(x_0))$.

$$\mathcal{L}(\theta) = \mathbb{E}\left[-\log p_\theta(x_0|x_1) - \sum_{t=2}^{T} \log p_\theta(x_{t-1}|x_t) + \log p(x_T)\right]$$

As mentioned previously, it is possible[1] to rewrite L_{vlb} almost completely in terms of KL divergences:

$$L_{vlb} = L_0 + L_1 + ... + L_{T-1} + L_T$$

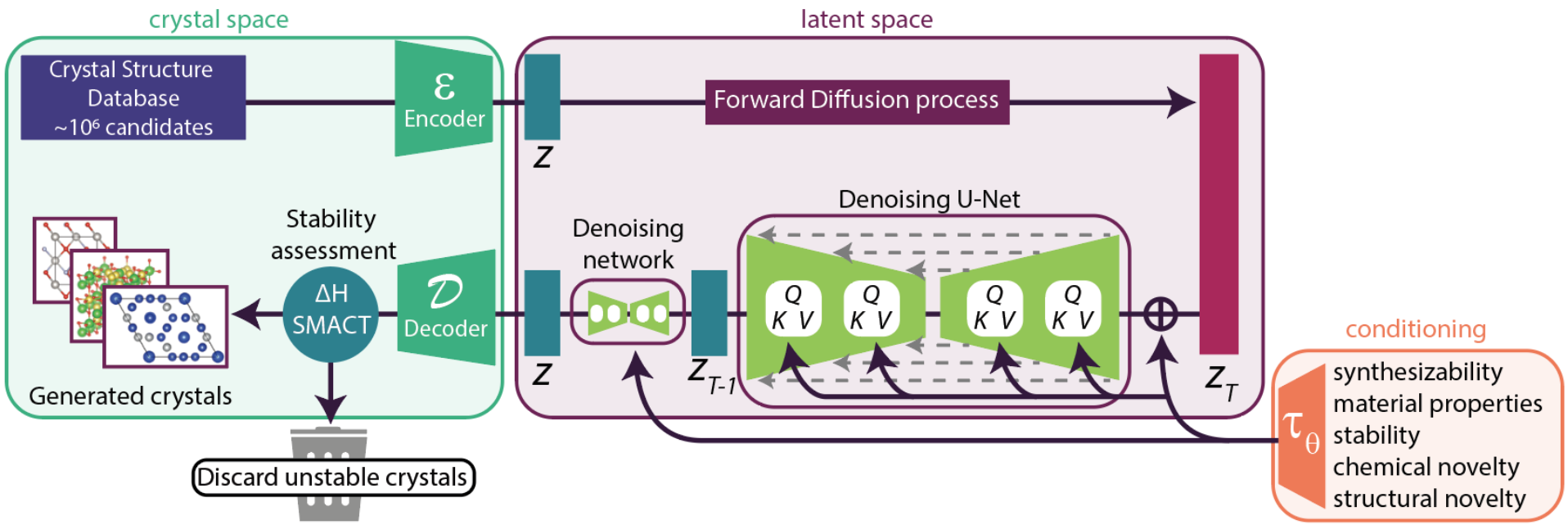
where

$$L_0 = -\log p_\theta(x_0|x_1)$$

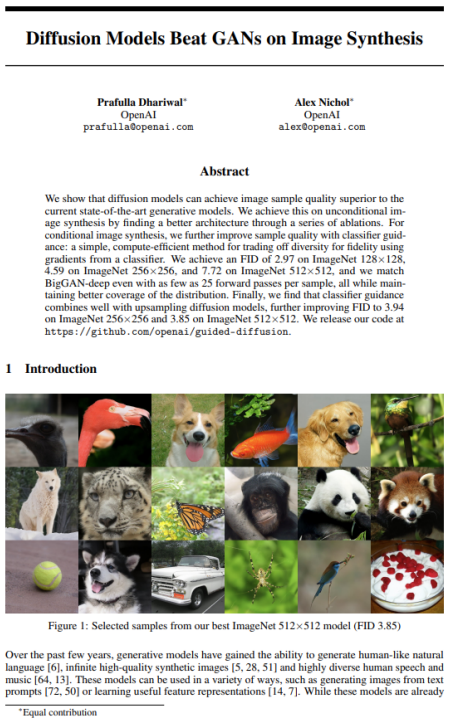
$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0) \,||\, p_\theta(x_{t-1}|x_t))$$

$$L_T = D_{KL}(q(x_T|x_0) \,||\, p(x_T))$$

We can condition our diffusion models to achieve “Guided Diffusion”



We can condition our diffusion models to achieve "Guided Diffusion"



Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal* OpenAI prafulla@openai.com

Alex Nichol* OpenAI alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

1 Introduction

Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

Over the past few years, generative models have gained the ability to generate human-like natural language [6], infinite high-quality synthetic images [5, 28, 51] and highly diverse human speech and music [64, 13]. These models can be used in a variety of ways, such as generating images from text prompts [72, 50] or learning useful feature representations [14, 7]. While these models are already

*Equal contribution

First proposed in "Diffusion Models Beat GANs on Image Synthesis"
Conditioner can be…. Anything! (text, image, class label etc)

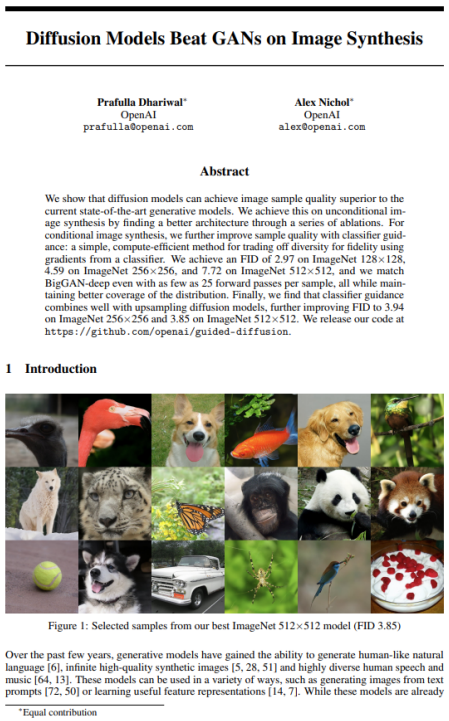
Train a classifier to predict class y given (x,t)
Classifier trains on noisy images

Noise prediction becomes

$$\hat{\epsilon} = \epsilon_\theta - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t}\log f_\phi(y|x_t)$$

"classifier guidance"

A separate classifier model isn't actually necessary though



arXiv:2105.05233v4 [cs.LG] 1 Jun 2021

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal* OpenAI prafulla@openai.com

Alex Nichol* OpenAI alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

1 Introduction

Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

Over the past few years, generative models have gained the ability to generate human-like natural language [6], infinite high-quality synthetic images [5, 28, 51] and highly diverse human speech and music [64, 13]. These models can be used in a variety of ways, such as generating images from text prompts [72, 50] or learning useful feature representations [14, 7]. While these models are already

*Equal contribution

Also proposed in "Diffusion Models Beat GANs on Image Synthesis"

Avoid a classifier and instead take a class label y during training, but during training, replace label y with a null label.

During sampling

$$\hat{\epsilon} = \epsilon_\theta(x_t|y) + s\big(\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)\big)$$

"classifier-free guidance"

Text can also be used for guided diffusion

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

Alex Nichol * Prafulla Dhariwal * Aditya Ramesh * Pranav Shyam Pamela Mishkin Bob McGrew
Ilya Sutskever Mark Chen

Abstract

Diffusion models have recently been shown to generate high-quality synthetic images, especially when paired with a guidance technique to trade off diversity for fidelity. We explore diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance. We find that the latter is preferred by human evaluators for both photorealism and caption similarity, and often produces photorealistic samples. Samples from a 3.5 billion parameter text-conditional diffusion model using classifier-free guidance are favored by human evaluators to those from DALL-E, even when the latter uses expensive CLIP reranking. Additionally, we find that our models can be fine-tuned to perform image inpainting, enabling powerful text-driven image editing. We train a smaller model on a filtered dataset and release the code and weights at https://github.com/openai/glide-text2im.

1. Introduction

Images, such as illustrations, paintings, and photographs, can often be easily described using text, but can require specialized skills and hours of labor to create. Therefore, a tool capable of generating realistic images from natural language can empower humans to create rich and diverse visual content with unprecedented ease. The ability to edit images using natural language further allows for iterative refinement and fine-grained control, both of which are critical for real world applications.

Recent text-conditional image models are capable of synthesizing images from free-form text prompts, and can compose unrelated objects in semantically plausible ways (Xu et al., 2017; Zhu et al., 2019; Tao et al., 2020; Ramesh et al., 2021; Zhang et al., 2021). However, they are not yet able to generate photorealistic images that capture all aspects of their corresponding text prompts.

On the other hand, unconditional image models can synthesize photorealistic images (Brock et al., 2018; Karras et al., 2019a;b; Razavi et al., 2019), sometimes with enough fidelity that humans can't distinguish them from real images (Zhou et al., 2019). Within this line of research, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020b) have emerged as a promising family of generative models, achieving state-of-the-art sample quality on a number of image generation benchmarks (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2021).

To achieve photorealism in the class-conditional setting, Dhariwal & Nichol (2021) augmented diffusion models with *classifier guidance*, a technique which allows diffusion models to condition on a classifier's labels. The classifier is first trained on noised images, and during the diffusion sampling process, gradients from the classifier are used to guide the sample towards the label. Ho & Salimans (2021) achieved similar results without a separately trained classifier through the use of *classifier-free guidance*, a form of guidance that interpolates between predictions from a diffusion model with and without labels.

Motivated by the ability of guided diffusion models to generate photorealistic samples and the ability of text-to-image models to handle free-form prompts, we apply guided diffusion to the problem of text-conditional image synthesis. First, we train a 3.5 billion parameter diffusion model that uses a text encoder to condition on natural language descriptions. Next, we compare two techniques for guiding diffusion models towards text prompts: CLIP guidance and classifier-free guidance. Using human and automated evaluations, we find that classifier-free guidance yields higher-quality images.

We find that samples from our model generated with classifier-free guidance are both photorealistic and reflect a wide breadth of world knowledge. When evaluated by human judges, our samples are preferred to those from DALL-E (Ramesh et al., 2021) 87% of the time when evaluated for photorealism, and 69% of the time when evaluated for caption similarity.

*Equal contribution. Correspondence to alex@openai.com, prafulla@openai.com, aramesh@openai.com

Shown in GLIDE paper

CLIP was used to measure the similarity between a text input and an image input

The dot product between the image and the text input is used as the score for the gradient

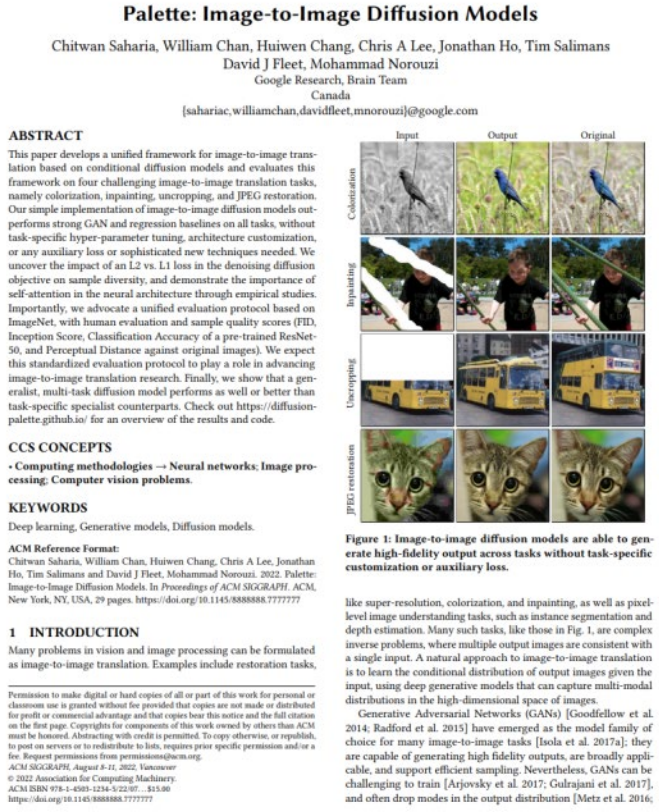
In GLIDE, they retrain CLIP on noisy images

Text can also be used for guided diffusion

Figure 10. Comparison of GLIDE to two CLIP guidance strategies applied to pre-trained ImageNet diffusion models. On the left, we use a vanilla CLIP model to guide the 256×256 diffusion model from Dhariwal & Nichol (2021), using a combination of engineered perceptual losses and data augmentations (Crowson, 2021a). In the middle, we use our noised ViT-B CLIP model to guide the ImageNet 64×64 diffusion model from Dhariwal & Nichol (2021), then apply a diffusion upsampler. On the right, we show random samples from GLIDE with classifier-free guidance scale 3.0.

https://arxiv.org/pdf/2112.10741.pdf

We can also do image-to-image diffusion



Palette: Image-to-Image Diffusion Models

Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans
David J Fleet, Mohammad Norouzi
Google Research, Brain Team
Canada
{sahariac,williamchan,davidfleet,mnorouzi}@google.com

ABSTRACT

This paper develops a unified framework for image-to-image translation based on conditional diffusion models and evaluates this framework on four challenging image-to-image translation tasks, namely colorization, inpainting, uncropping, and JPEG restoration. Our simple implementation of image-to-image diffusion models outperforms strong GAN and regression baselines on all tasks, without task-specific hyper-parameter tuning, architecture customization, or any auxiliary loss or sophisticated new techniques needed. We uncover the impact of an L2 vs. L1 loss in the denoising diffusion objective on sample diversity, and demonstrate the importance of self-attention in the neural architecture through empirical studies. Importantly, we advocate a unified evaluation protocol based on ImageNet, with human evaluation and sample quality scores (FID, Inception Score, Classification Accuracy of a pre-trained ResNet-50, and Perceptual Distance against original images). We expect this standardized evaluation protocol to play a role in advancing image-to-image translation research. Finally, we show that a generalist, multi-task diffusion model performs as well or better than task-specific specialist counterparts. Check out https://diffusion-palette.github.io/ for an overview of the results and code.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Image processing**; **Computer vision problems**.

KEYWORDS

Deep learning, Generative models, Diffusion models.

ACM Reference Format:
Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans and David J Fleet, Mohammad Norouzi. 2022. Palette: Image-to-Image Diffusion Models. In *Proceedings of ACM SIGGRAPH*. ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/8888888.7777777

Figure 1: Image-to-image diffusion models are able to generate high-fidelity output across tasks without task-specific customization or auxiliary loss.

1 INTRODUCTION

Many problems in vision and image processing can be formulated as image-to-image translation. Examples include restoration tasks, like super-resolution, colorization, and inpainting, as well as pixel-level image understanding tasks, such as instance segmentation and depth estimation. Many such tasks, like those in Fig. 1, are complex inverse problems, where multiple output images are consistent with a single input. A natural approach to image-to-image translation is to learn the conditional distribution of output images given the input, using deep generative models that can capture multi-modal distributions in the high-dimensional space of images.

Generative Adversarial Networks (GANs) [Goodfellow et al. 2014; Radford et al. 2015] have emerged as the model family of choice for many image-to-image tasks [Isola et al. 2017a]; they are capable of generating high fidelity outputs, are broadly applicable, and support efficient sampling. Nevertheless, GANs can be challenging to train [Arjovsky et al. 2017; Gulrajani et al. 2017], and often drop modes in the output distribution [Metz et al. 2016;

ACM SIGGRAPH, August 8–11, 2022, Vancouver

ACM ISBN 978-1-4503-1234-5/22/07…$15.00
https://doi.org/10.1145/8888888.7777777

An image now becomes our condition

The condition image gets concatenated

The dot product between the image and the text input is used as the score for the gradient

$$\hat{\epsilon} = \epsilon_\theta(x_t|y) + s\big(\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)\big)$$

In GLIDE, they retrain CLIP on noisy images

Bayesian inference