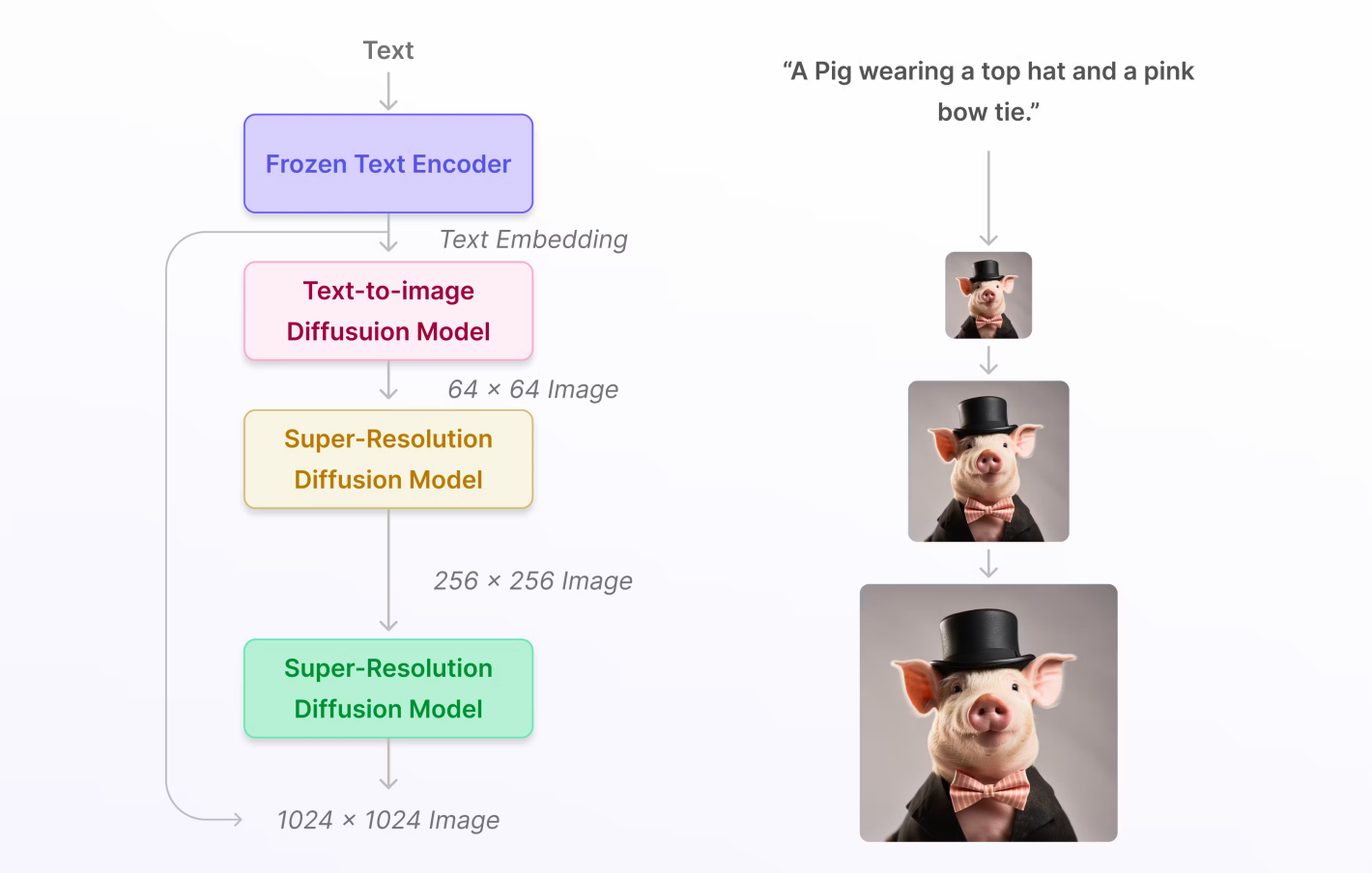
ENCORD

An Introduction to Diffusion Models for Machine Learning

Akruti Acharya • August 8, 2023 • 5 min read

< Back to blogs

Contents



Machine learning and artificial intelligence algorithms are constantly evolving to solve complex problems and enhance our understanding of data. One interesting group of models is diffusion models, which have gained significant attention for their ability to capture and simulate complex processes like data generation and image synthesis.

In this article, we will explore:

- What is diffusion?
- What are diffusion models?
- How do diffusion models work?
- Applications of diffusion models
- Popular diffusion models for image generation

What is Diffusion?

Diffusion is a fundamental natural phenomenon observed in various systems, including physics, chemistry, and biology.

In everyday life, this is readily noticeable. Consider the example of spraying perfume. Initially, the perfume molecules are densely concentrated near the point of spraying. As time passes, the molecules disperse.

Diffusion is the process of particles, information, or energy moving from an area of high concentration to an area of lower concentration. This happens because systems tend to reach equilibrium, where concentrations become uniform throughout the system.

In the context of machine learning and data generation, diffusion refers to a specific approach for generating data using a stochastic process similar to a **Markov chain**. In this context, diffusion models are used to create new data samples by starting with simple, easily generated data and then gradually transforming it into more complex and realistic data.

What are Diffusion Models in Machine Learning?

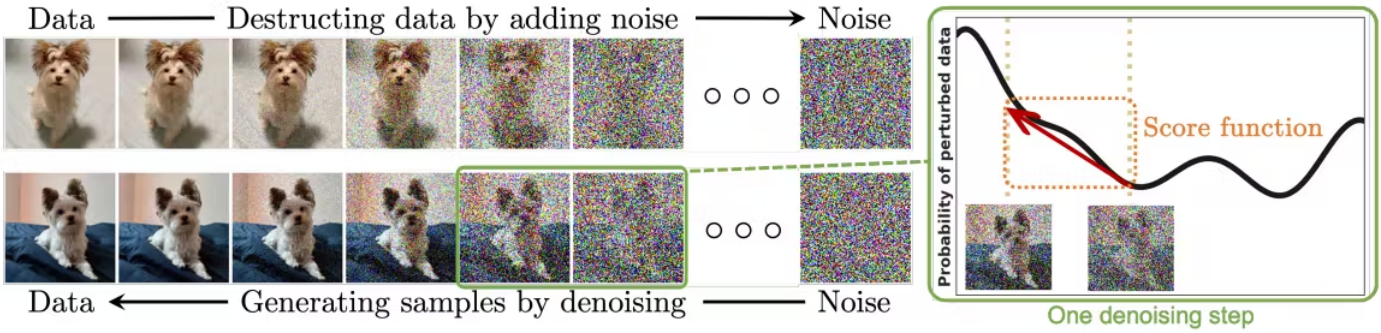
Diffusion models are generative models, which means that they generate new data based on the data they are trained on. For example, a diffusion model trained on a collection of human faces can generate new and realistic human faces with various features and expressions, even if those specific faces were not present in the original training dataset.

These models focus on modeling the step-by-step evolution of data distribution from a simple starting point to a more complex distribution. The underlying concept of diffusion models is to transform a simple and easily samplable distribution, typically a **Gaussian distribution**, into a more complex data distribution of interest. This transformation is achieved through a series of invertible operations. Once the model learns the transformation process, it can generate new samples by starting from a point in the simple distribution and gradually "diffusing" it to the desired complex data distribution.

Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs are a type of diffusion model used for probabilistic data generation. As mentioned earlier, diffusion models generate data by applying a sequence of transformations to random noise. DDPMs, in particular, operate by simulating a diffusion process that transforms noisy data into clean data samples.

To train DDPMs, the process entails acquiring knowledge of the diffusion process's parameters, effectively capturing the relationship between clean and noisy data during each transformation step. During inference (generation), DDPMs start with noisy data (e.g., noisy images) and iteratively apply the learned transformations in reverse to obtain denoised and realistic data samples.



Diffusion Models: A Comprehensive Survey of Methods and Applications

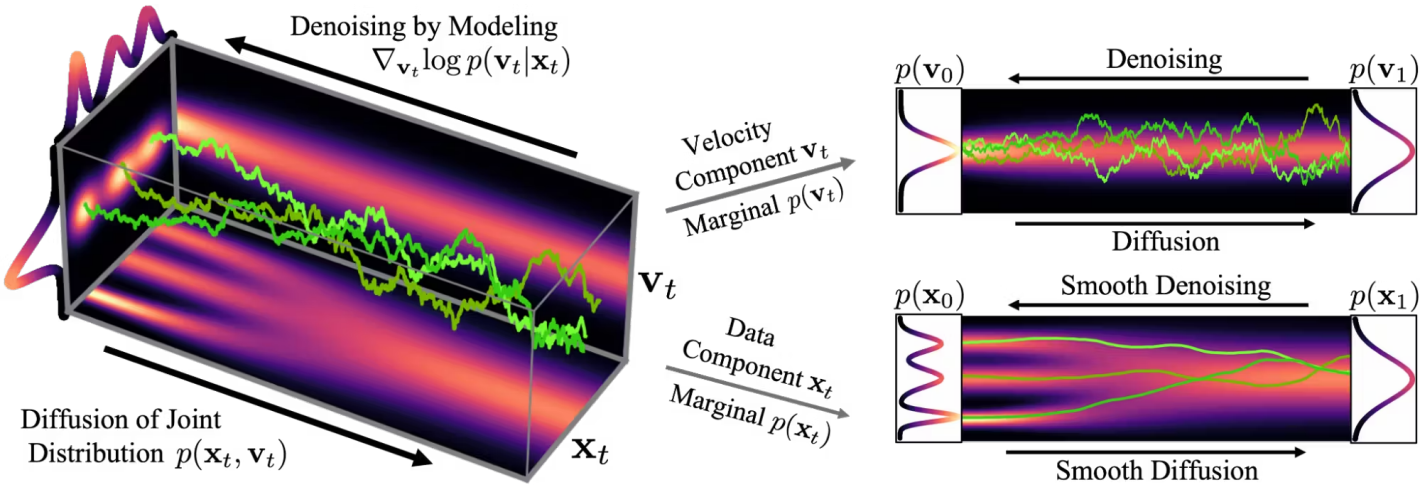
DDPMs are particularly effective for image-denoising tasks. They can effectively remove noise from corrupted images and produce visually appealing denoised versions. Moreover, DDPMs can also be used for image inpainting and super-resolution, among other applications.

Score Based Generative Models (SGMs)

Score Based Generative Models are a class of generative models that use the score function to estimate the likelihood of data samples. The score function, also known as the gradient of the log-likelihood with respect to the data, provides essential information about the local structure of the data distribution.

SGMs use the score function to estimate the probability density of the data at any given point. By doing so, they can effectively model complex and high-dimensional data distributions. Although the score function can be computed analytically for

some probability distributions, it is often estimated using techniques like automatic differentiation and neural networks.



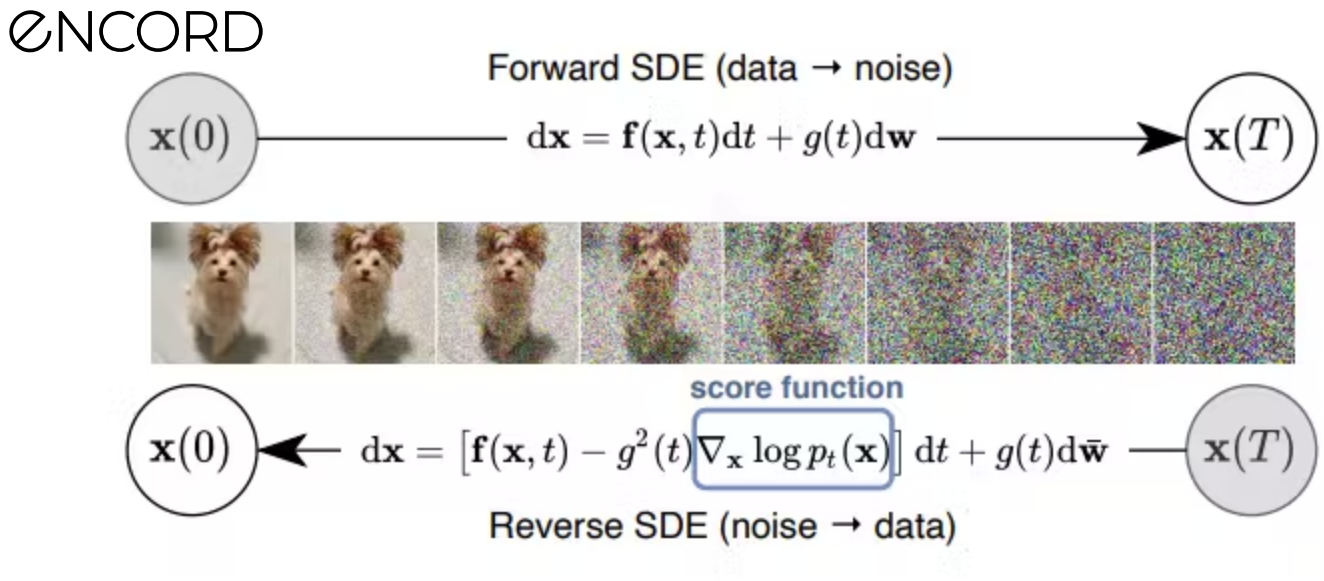
Score-Based Generative Modeling with Critically-Damped Langevin Diffusion

SGMs can generate data samples that resemble the training data distribution by utilizing the score function. This is achieved by iteratively updating them in the direction of the negative gradient of the log-likelihood.

Stochastic Differential Equations (Score SDEs)

Stochastic Differential Equations (SDEs) are mathematical equations that describe how a system changes over time when subject to both deterministic and random forces. In generative modeling, Score SDEs are a way to parameterize the score-based models.

In Score SDEs, the score function is defined as a solution to a stochastic differential equation. By solving this differential equation, the model can learn a data-driven score function that adapts to the data distribution. In essence, the Score SDEs use stochastic processes to model the evolution of data samples and guide the generative process toward generating high-quality data samples.



Solving a reverse-time SDE yields a score-based generative model. *Score-Based Generative Modeling through Stochastic Differential Equations*

Score SDEs and score-based modeling can be combined to create powerful generative models that are capable of handling complex data distributions and generating diverse and realistic samples.

How do Diffusion Models Work?

Diffusion models are a class of generative models that operate based on the concept of "reverse diffusion" to simulate data generation.

Let's break down how diffusion models work step-by-step:

Data Preprocessing

The initial step involves preprocessing the data to ensure proper scaling and centering. Typically, standardization is applied to convert the data into a distribution with a mean of zero and a variance of one. This prepares the data for subsequent transformations during the diffusion process, enabling the diffusion models to effectively handle noisy images and generate high-quality samples.

Forward Diffusion

During forward diffusion, the model starts with a sample from a simple distribution, typically a Gaussian distribution, and applies a sequence of invertible transformations to "diffuse" the sample step-by-step until it reaches the desired complex data points distribution. Each diffusion step introduces more complexity to the data, capturing the intricate patterns and details of the original distribution. This process can be thought of as gradually adding Gaussian noise to the initial sample, resulting in the generation of diverse and realistic samples as the diffusion process unfolds.

Training the Model

Training a diffusion model involves learning the parameters of the invertible transformations and other components of the model. This process typically involves optimizing a loss function, which evaluates how effectively the model can transform samples from a simple distribution into ones that closely resemble the complex data distribution. Diffusion models are often called score-based models because the training process involves estimating the score function (gradient of the log-likelihood) of the data distribution with respect to the input data points.

The training process can be computationally intensive but advances in optimization algorithms and hardware acceleration have made it feasible to train diffusion models on various datasets.

Reverse Diffusion

Once the forward diffusion process generates a sample from the complex data distribution, the reverse diffusion process maps it back to the simple distribution through a sequence of inverse transformations.

Through this reverse diffusion process, diffusion models can generate new data samples by starting from a point in the simple distribution and diffusing it step-by-step to the desired complex data distribution. The generated samples exhibit a striking resemblance to the original data distribution, making diffusion models a powerful tool for tasks such as image synthesis, data completion, and denoising.

Benefits of Using Diffusion Models

Diffusion models offer several advantages over traditional generative models like **GANs** (Generative Adversarial Networks) and **VAEs** (Variational Autoencoders). These benefits stem from their unique approach to data generation and the use of reverse diffusion.

Image Quality and Coherence

Diffusion models are adept at generating high-quality images with fine details and realistic textures. By capturing the underlying complexity of the data distribution through reverse diffusion, diffusion models produce images with more coherent structures and fewer artifacts when compared to traditional generative models.

OpenAI's paper, **Diffusion Models Beat GANs on Image Synthesis** shows that diffusion models can achieve image sample quality superior to current state-of-the-art generative models.

Stable Training

Training diffusion models are generally more stable than training GANs, which are notoriously challenging to train. GANs require balancing the learning rates of the generator and discriminator networks, and mode collapse can occur when the generator fails to capture all aspects of the data distribution. In contrast, diffusion models use likelihood-based training, which tends to be more stable and avoids mode collapse.

Privacy-Preserving Data Generation

Diffusion models are suitable for applications in which data privacy is a concern. Since the model is based on invertible transformations, it is possible to generate synthetic data samples without exposing the underlying private information of the original data.

Handling Missing Data

Diffusion models can handle missing data during the generation process. Since reverse diffusion can work with incomplete data samples, the model can generate

coherent samples even when parts of the input data are missing.

ENCORD

Robustness to Overfitting

Traditional generative models like GANs can be prone to overfitting, where the model memorizes the training data and fails to generalize well to unseen data. Diffusion models have shown to be more robust to overfitting due to the use of likelihood-based training and the inherent properties of reverse diffusion, which encourages more coherent and diverse sample generation.

Interpretable Latent Space

Compared to GANs, diffusion models often have a more interpretable latent space. By introducing a latent variable into the reverse diffusion process, the model can capture additional variations and generate diverse samples. The reverse diffusion process maps the complex data distribution back to a simple distribution, allowing the latent space to represent meaningful features, patterns, and latent variables present in the data. This interpretability, coupled with the flexibility of the latent variable, can be valuable for understanding the learned representations, gaining insights into the data, and enabling fine-grained control over image generation.

Scalability to High-Dimensional Data

Diffusion models have demonstrated promising scalability to high-dimensional data, such as images with large resolutions. The step-by-step diffusion process allows the model to efficiently generate complex data distributions without getting overwhelmed by the high dimensionality of the data.

Applications of Diffusion Models

Diffusion models have shown promise in various applications across different domains due to their ability to model complex data distributions and generate high-quality samples. Let’s dive into some notable applications of diffusion models:

Text to Video

Make-A-Video: Text-to-Video Generation without Text-Video Data.

Diffusion models are a promising approach for text-to-video synthesis. The process involves first representing the textual descriptions and video data in a suitable format, such as word embeddings or transformer-based language models for text and video frames in a sequence format.

During the forward diffusion process, the model takes the encoded text representation and gradually generates video frames step-by-step, incorporating the semantics and dynamics of the text. Each diffusion step refines the rendered frames, transforming them from random noise into visually meaningful content that aligns with the given text. The reverse diffusion process then maps the generated video frames back to the simple distribution, completing the text-to-video synthesis.

This conditional generation enables diffusion models to create visually compelling videos based on textual prompts, with potential applications in video captioning,

storytelling, and creative content generation. However, challenges remain, including ensuring temporal coherence between frames, handling long-range dependencies in text, and improving scalability for complex video sequences.

Meta's **Make-A-Video** is a well-known example of leveraging diffusion models to develop machine learning models for text-to-video synthesis.

Image to Image

Diffusion models offer a powerful approach for image-to-image translation tasks, involving the transformation of images from one domain to another while preserving semantic information and visual coherence.

The process involves conditioning the diffusion model on a source image and using reverse diffusion to generate a corresponding target image that represents a transformed version of the source. To achieve this, both the source and target images are represented in a suitable format for the model, such as pixel values or embeddings.

During the forward diffusion process, the model gradually transforms the source image, capturing the desired changes or attributes specified by the target domain. This often involves upsampling the source image to match the resolution of the target domain and refining the generated image step-by-step to produce high-quality and coherent results.

The reverse diffusion process then maps the generated target image back to the simple distribution, completing the image-to-image translation. This conditional generation allows diffusion models to excel in tasks like image colorization, style transfer, and image-to-sketch conversion.

The paper **Denoising Diffusion Probabilistic Models (DDPM)** which was initialized by Sohl-Dickstein et al and then proposed by Ho et al 2020 is an influential paper that showcases diffusion models as a potent neural network-based method for image generation tasks.

Image Search
ENCORD

Diffusion models can be applied to image search tasks as a powerful content-based image retrieval technique. To use diffusion models for image search, the first step is to encode the images in a latent space using the reverse diffusion process.

During reverse diffusion, the model maps each image to a point in the simple distribution. This latent representation retains the essential visual information of the image while discarding irrelevant noise and details, making it suitable for efficient and effective similarity searches. When a query image is given for image search, the model encodes the query image into the same latent space using the reverse diffusion process.

The similarity between the query image and the database images can be measured using standard distance metrics (e.g., Euclidean distance) in the latent space. Images with the most similar latent representations are retrieved, producing relevant and visually similar images to the query.

This application of diffusion models for image search enables accurate and fast content-based retrieval, useful in various domains such as image libraries, image databases, and reverse image search engines

Reverse Image Search

Diffusion models can be harnessed for reverse image search, also known as content-based image retrieval, to find the source or visually similar images based on a given query image.

To facilitate reverse image search with diffusion models, a database of images needs to be preprocessed by encoding each image into a latent space using the reverse diffusion process. This latent representation captures the essential visual characteristics of each image, allowing for efficient and accurate image retrieval.

When a query image is provided for reverse image search, the model encodes the query image into the same latent space using the reverse diffusion process. By measuring the similarity between the query image's latent representation and the database images' latent representations using distance metrics (e.g., Euclidean

distance), the model can identify and retrieve the most visually similar images from the database.

This application of diffusion models for reverse image search facilitates fast and reliable content-based retrieval, making it valuable for various applications, including image recognition, plagiarism detection, and multimedia databases.

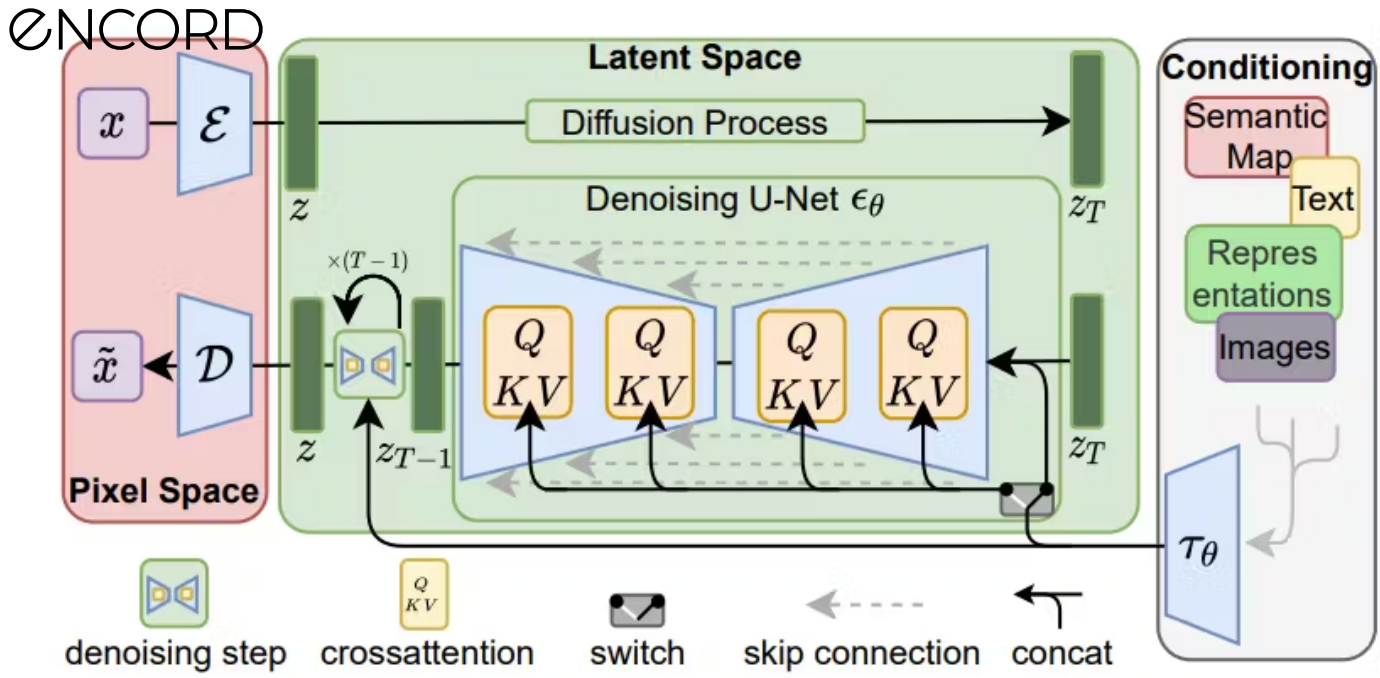
Well-known Diffusion Models for Image Generation

Stable Diffusion

High-Resolution Image Synthesis with Latent Diffusion Models

Stable diffusion is a popular approach for image generation that uses diffusion models (DMs) and the efficiency of latent space representation. The method introduces a two-stage training process to enable high-quality image synthesis while overcoming the computational challenges associated with directly operating in pixel space.

In the first stage, an autoencoder is trained to compress the image data into a lower-dimensional latent space that maintains perceptual equivalence with the original data. This learned latent space serves as an efficient and scalable alternative to the pixel space, providing better scaling properties with respect to spatial dimensionality. By training diffusion models in this latent space, known as Latent Diffusion Models (LDMs), Stable Diffusion achieves a near-optimal balance between complexity reduction and detail preservation, leading to a significant boost in visual fidelity.



High-Resolution Image Synthesis with Latent Diffusion Models

Stable diffusion introduces cross-attention layers into the model architecture, enabling the diffusion models to become robust and flexible generators for various types of conditioning inputs, such as text or bounding boxes. This architectural enhancement opens up new possibilities for image synthesis and allows for high-resolution generation in a convolutional manner.

The approach of stable diffusion has demonstrated remarkable success in image inpainting, class-conditional image synthesis, text-to-image synthesis, unconditional image generation, and super-resolution tasks. Moreover, it achieves state-of-the-art results while considerably reducing the computational requirements compared to traditional pixel-based diffusion models.

The code for stable diffusion has been made publicly available on **GitHub**.

DALL-E 2

a dolphin in an astronaut suit on saturn, artstation

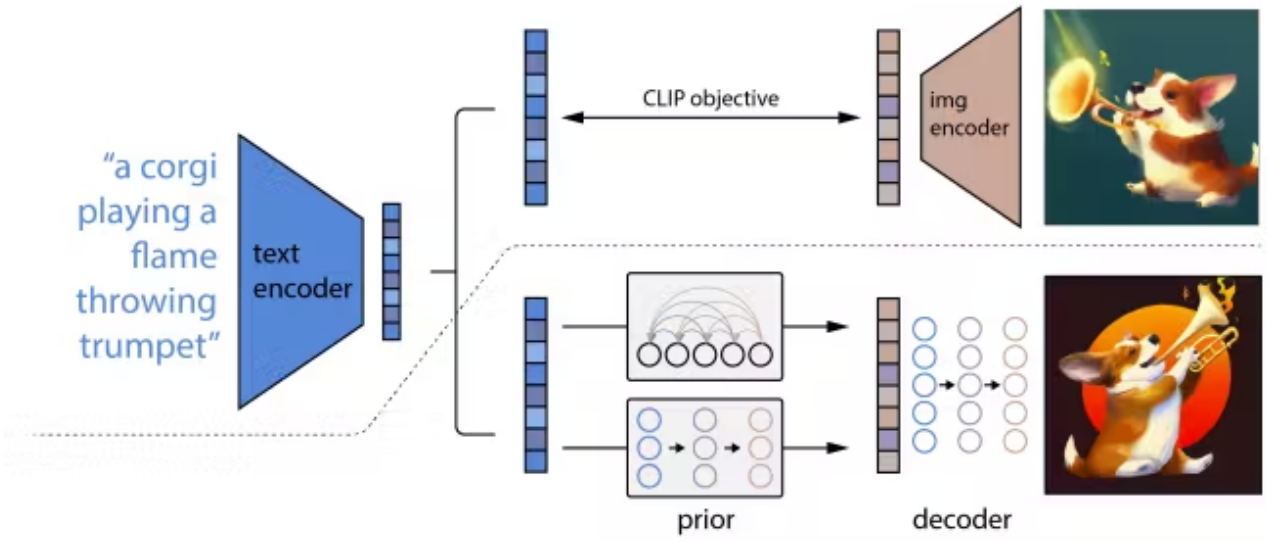
a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Hierarchical Text-Conditional Image Generation with CLIP Latents

DALL-E 2 utilizes contrastive models like CLIP to learn robust image representations which capture the semantics and style. It has a 2-staged model which consists of a prior stage that generates a CLIP image embedding based on a given text caption and a decoder stage.

The decoders in the model use diffusion models. These models are conditioned on image representations and produce variations of an image that preserve both its semantics and style while altering non-essential details.

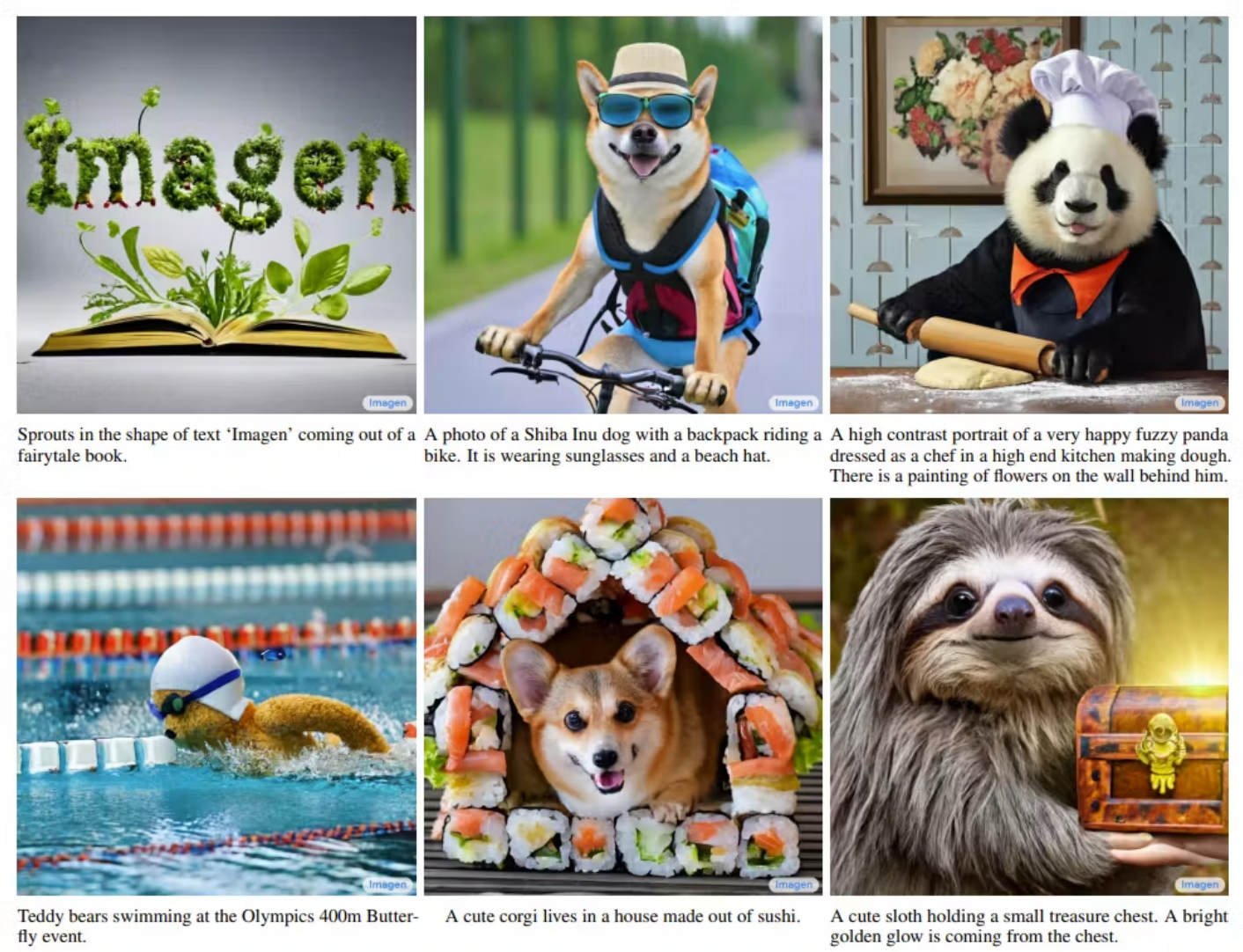


Hierarchical Text-Conditional Image Generation with CLIP Latents

The joint embedding space of CLIP enables language-guided image manipulations in a zero-shot manner, allowing the diffusion model to generate images based on textual description without explicit supervision.

Read the paper **Hierarchical Text-Conditional Image Generation with CLIP Latents** here.

Imagen

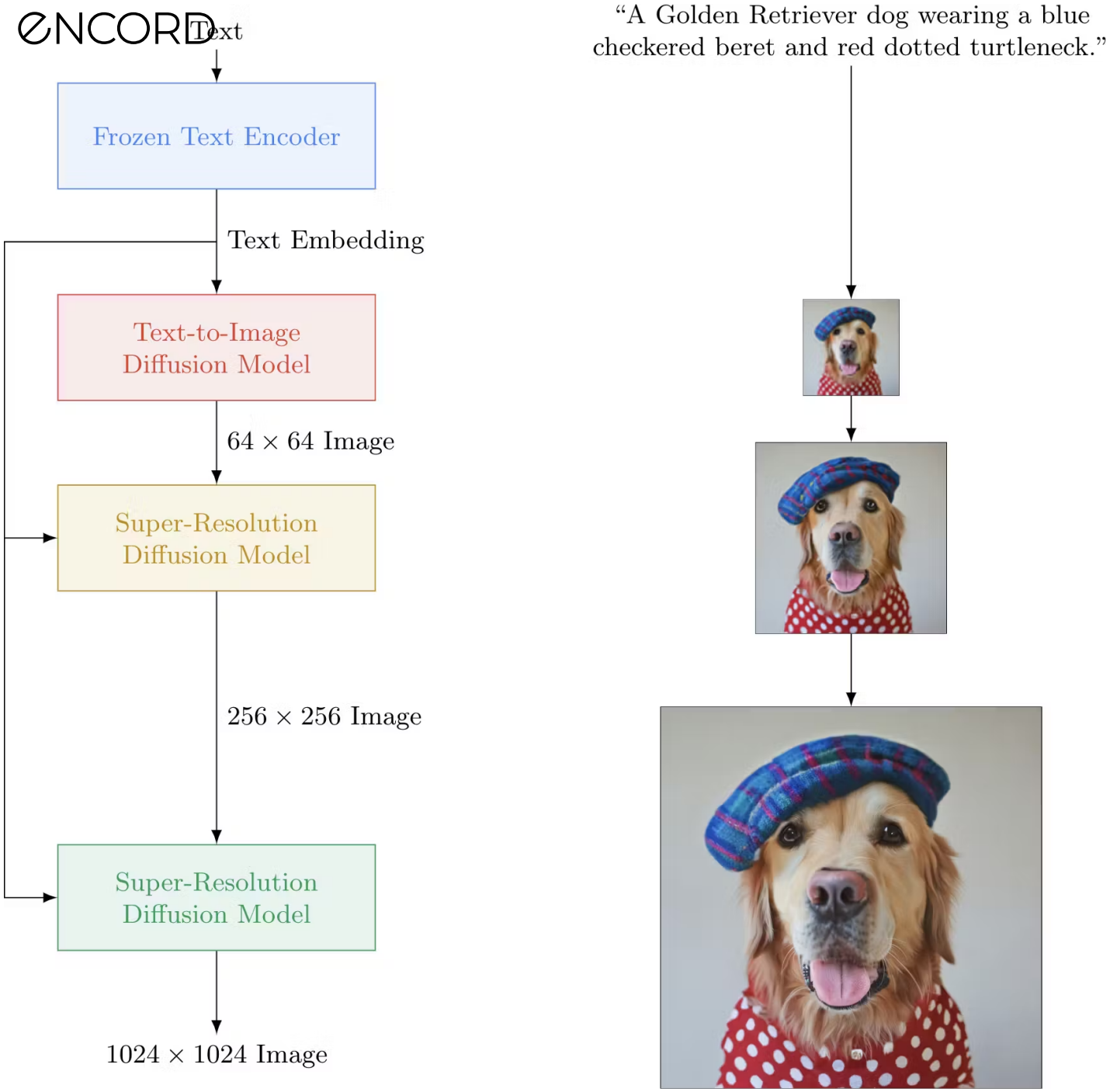


Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Imagen is a text-to-image diffusion model that stands out for its exceptional image generation capabilities. The model is built upon two key components - large pretrained frozen text encoders and diffusion models. Leveraging the strength of

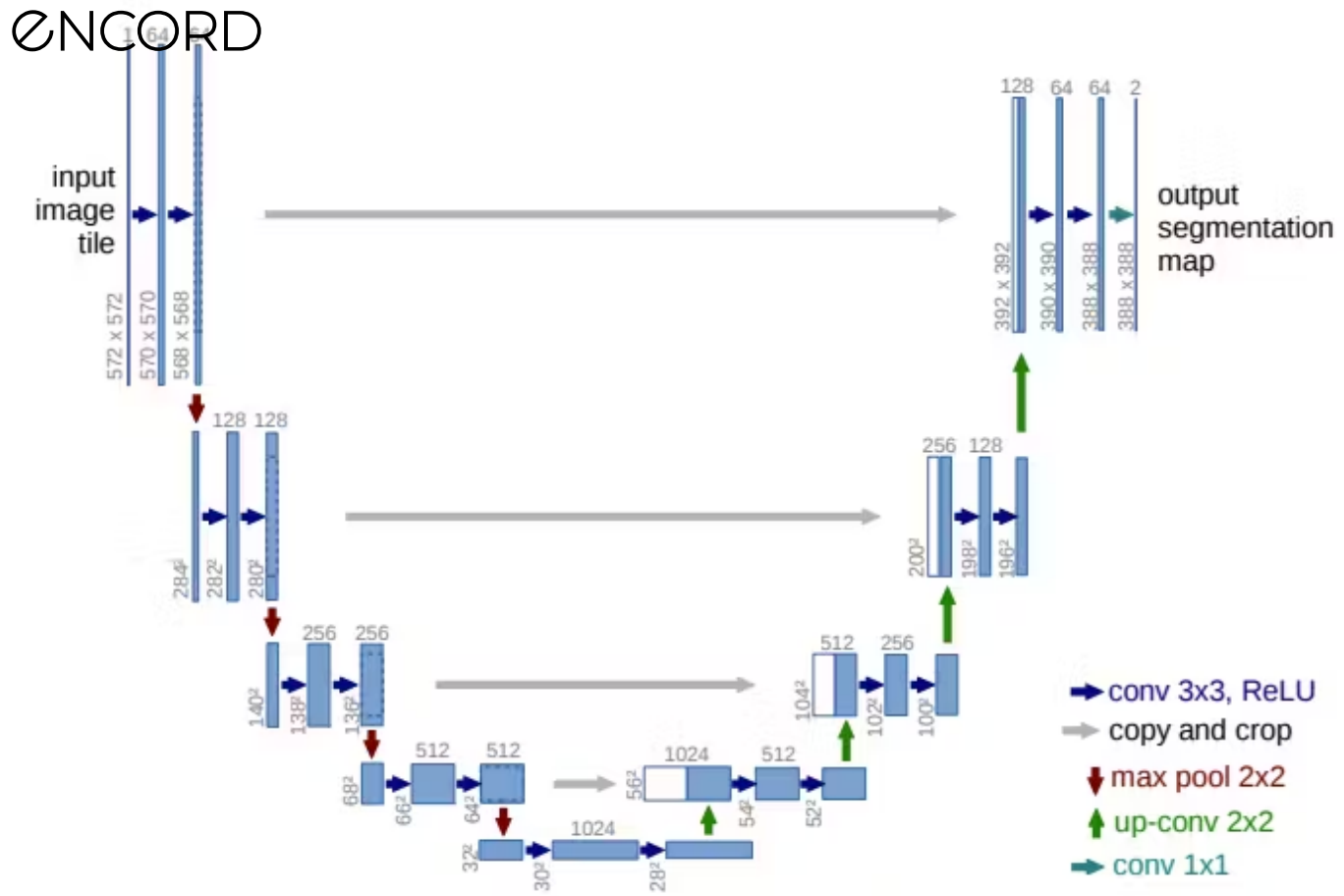
transformer-based language models, such as T5, Imagen showcases remarkable proficiency in understanding textual descriptions and encoding them effectively for image synthesis.

Imagen uses a new thresholding sampler, which enables the use of very large classifier-free guidance weights. This enhancement further enhances the guidance and control over the image generation process, resulting in improved photorealism and image-text alignment.



Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

To address computational efficiency, the researchers introduce a novel Efficient **U-Net** architecture, which is optimized for better computing and memory efficiency, leading to faster convergence during training.



U-Net: Convolutional Networks for Biomedical Image Segmentation

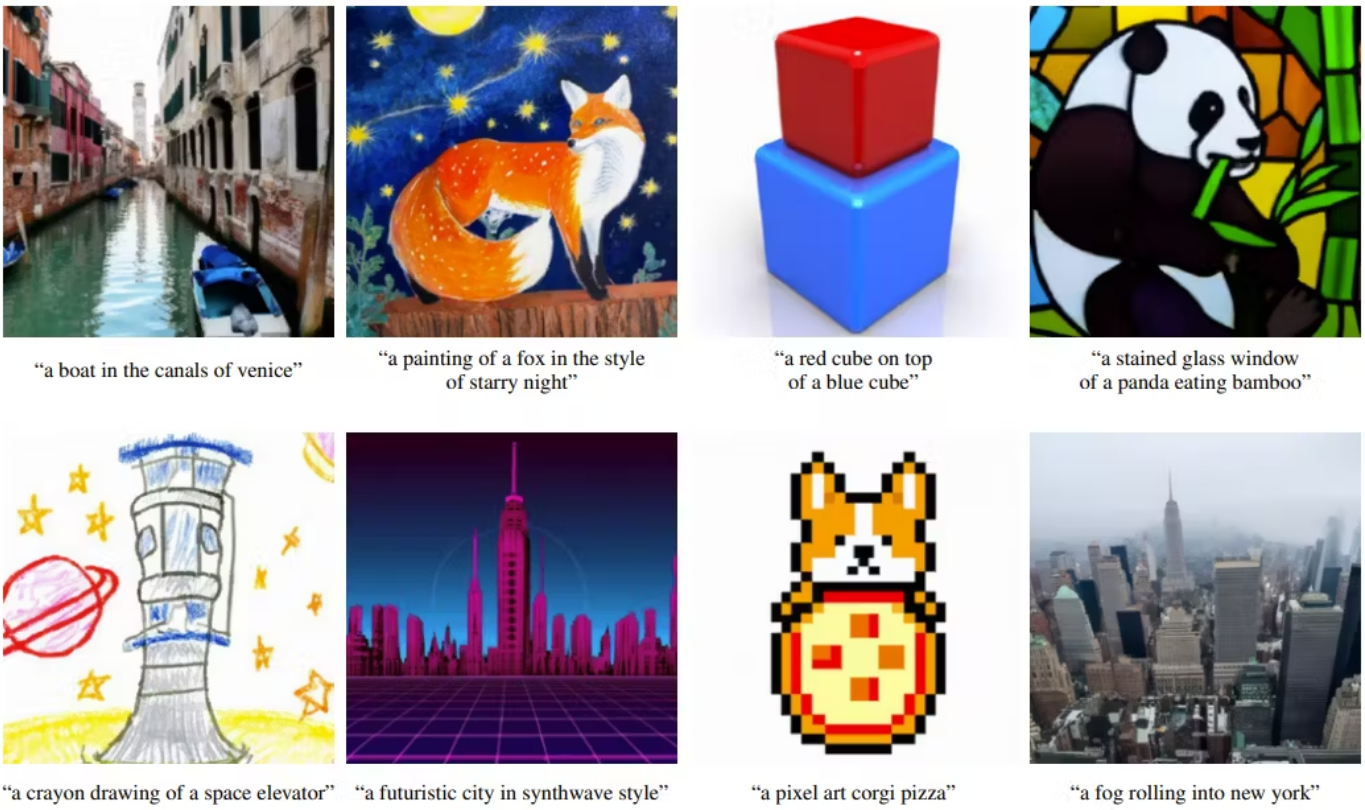
A significant finding of the research is the importance of scaling the pretrained text encoder size for the image generation task. Increasing the size of the language model in Imagen has a substantial positive impact on both the fidelity of generated samples and the alignment between images and corresponding text descriptions. This highlights the effectiveness of large language models (LLMs) in encoding meaningful representations of text, which significantly influences the quality of the generated images.

The PyTorch implementation of Imagen can be found on **GitHub**.

GLIDE

Guided Language to Image Diffusion for Generation and Editing (GLIDE) is another powerful text-conditional image synthesis model by OpenAI. It is a computer vision model based on diffusion models. It leverages a 3.5 billion parameter diffusion model

that utilizes a text encoder to condition natural language descriptions. The primary goal of GLIDE is to generate high-quality images based on textual prompts while also offering editing capabilities to improve model samples for complex prompts.



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

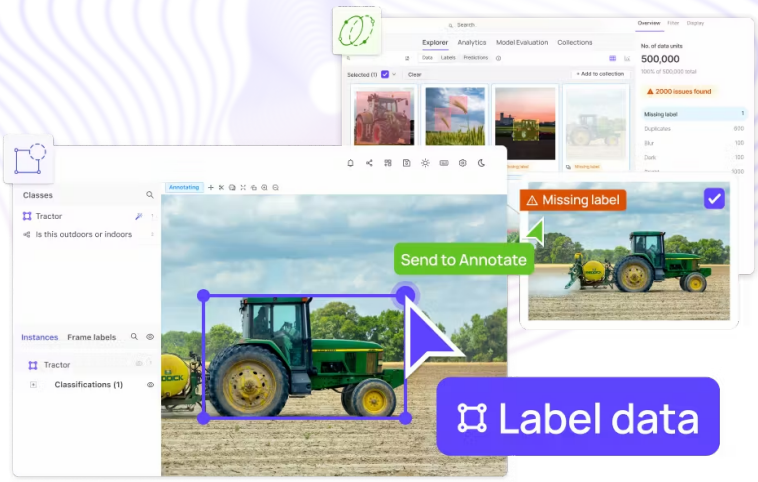
In the context of text-to-image synthesis, GLIDE explores two different guidance strategies: CLIP guidance and classifier-free guidance. Through human and automated evaluations, the researchers discover that classifier-free guidance yields higher-quality images compared to the alternative. This guidance mechanism allows GLIDE to generate photorealistic samples that closely align with the given text descriptions.

One notable application of GLIDE in computer vision is its potential to significantly reduce the effort required to create disinformation or Deepfakes. However, to address ethical concerns and safeguard against potential misuse, the researchers have released a smaller diffusion model and a noised CLIP model trained on filtered datasets.

OpenAI has made the codebase for the small, filtered data GLIDE model publicly available on **GitHub**.

Diffusion Models: Key Takeaways

- Diffusion models are a class of generative models that simulate the data generation process by transforming a simple starting distribution into the desired complex data distribution through a sequence of invertible operations.
- Compared to traditional generative models, diffusion models have better image quality, interpretable latent space, and robustness to overfitting.
- Diffusion models have diverse applications across several domains, such as text-to-video synthesis, image-to-image translation, image search, and reverse image search.
- Diffusion models excel at generating realistic and coherent content based on textual prompts and efficiently handling image transformations and retrievals. Popular models include Stable Diffusion, DALL-E 2, and Imagen.



Written by Akruti Acharya

Akruti is a data scientist and technical content writer with a M.Sc. in Machine Learning & Artificial Intelligence from the University of Birmingham. She enjoys exploring new things and applying her technical and analytical skills to solve challenging problems and sharing her knowledge and... see more

View more posts →

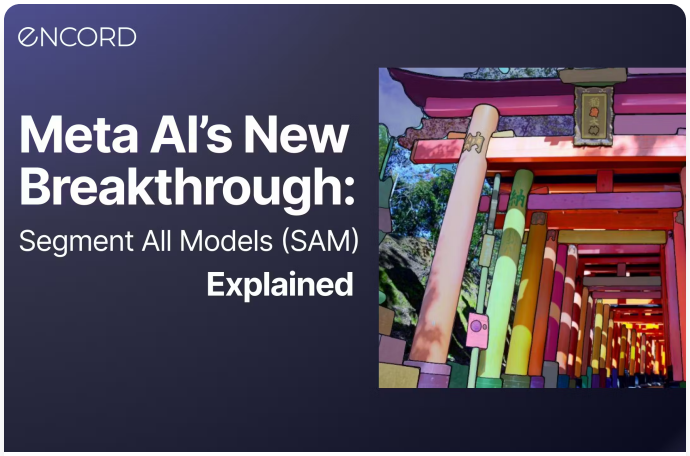
ENCORD

Discuss this blog on Slack

Join the Encord Developers community to discuss the latest in computer vision, machine learning, and data-centric AI

Join the community

Related Blogs

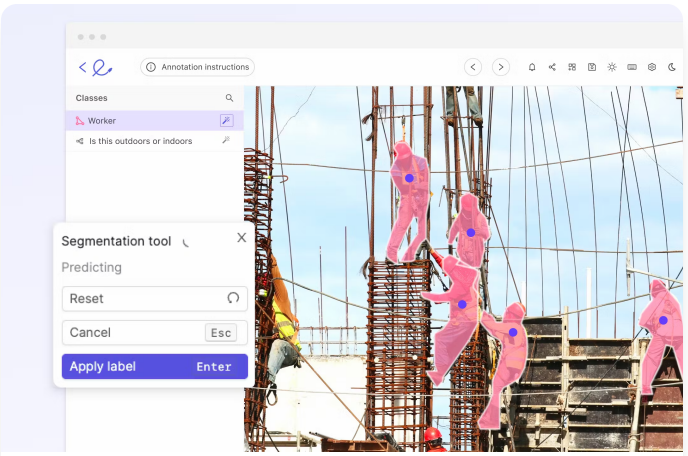


MACHINE LEARNING

Meta AI's New Breakthrough: Segment Anything Model...

If you thought the AI space was already moving fast with ChatGPT, GPT4, and Stable Diffusion, then strap in and get ready for the...

April 6 6 min

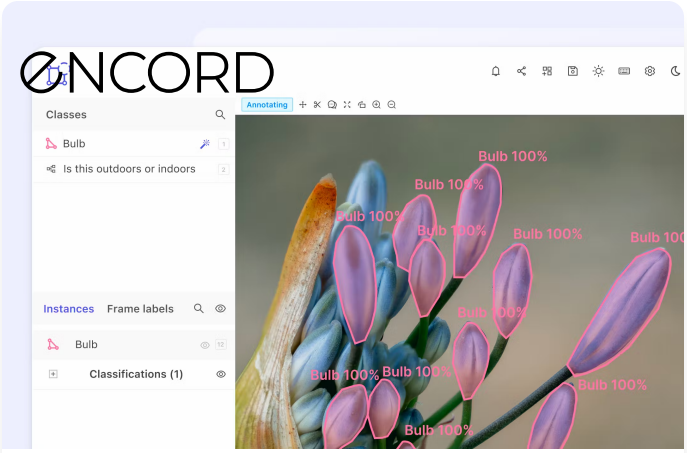


TUTORIALS

How To Fine-Tune Segment Anything

Computer vision is having its ChatGPT moment with the release of the Segment Anything Model (SAM) by Meta last week....

April 13 10 min



DATA OPERATIONS

Best Image Annotation Tools for Computer Vision [Updated...

{{gray_callout_start}} Guide to the most popular image annotation tools that you need to know about in 2024. Compare the feature...

February 8 10 min

Software To Help You Turn Your Data Into AI

Forget fragmented workflows, annotation tools, and Notebooks for building AI applications. Encord Data Engine accelerates every step of taking your model into production.

ENCORD

Your work email

Book a demo

Terms · Privacy Policy

Product

Image

Video

DICOM

SAR

Automation

API & Python SDK

Quality Assessment

Encord Active

Industries

Aerospace & Defense

Agriculture

Computer Vision

Energy

Healthcare & Medical

Insurance

Life Sciences & Biotech

Logistics

Manufacturing

Media, Gaming & Entertainment

Retail & E-commerce

Sports

Technology & Software

Company

About

Careers

Customers

Contact Us

Documentation

Glossary

Blog

Press

Pricing

Security

Subscribe