Published in CodeX

Jair Ribeiro Follow

Aug 30 · 6 min read · ✨ · Listen

A quick look under the hood of Stable Diffusion Open Source architecture.

Analyzing the model that is powering a new wave of generation models.

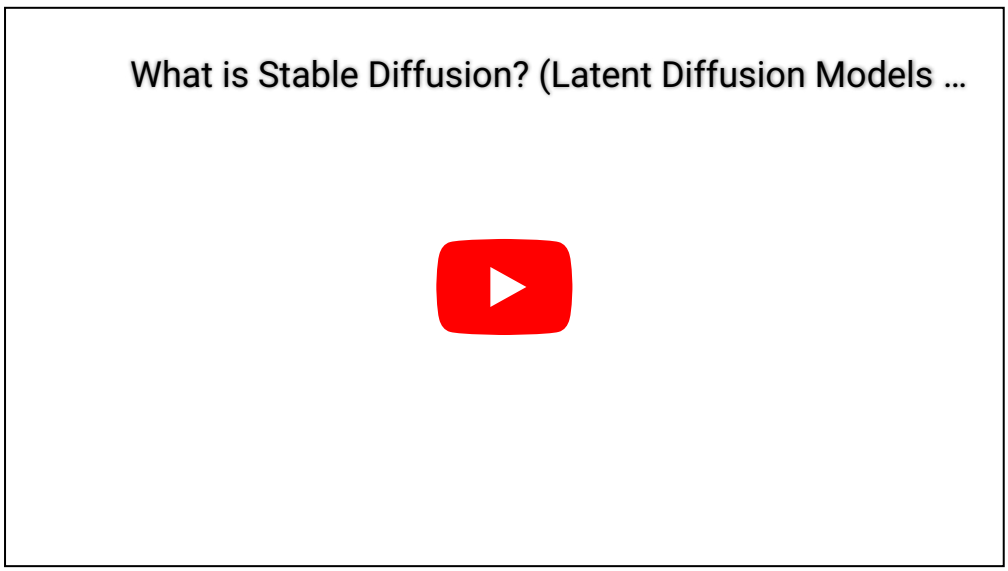
If you haven't lived under a rock for the past year, you've probably heard that the text-to-image generation space is undergoing a massive revolution. Last week, an AI startup called Stability AI unveiled its first version of its **Stable Diffusion** text-to-image synthesis model.

The good (I could say great) news is that they released it as an open-source model for free. That is significant!

And being an open-source model means that, if you have a sufficiently powerful graphics card, you can download and run the model on your computer (I did it and even already built a full creative project on it.. you can read about it **here**).

Because it is an open-source model, we can use it for non-commercial purposes and commercial under the terms of the license called **Creative ML OpenRAIL-M** — which is fine enough to impose some usage restrictions, such as not using it to break applicable laws, generating false information, discriminate against individuals, or provide medical advice.

I've seen an explosion of innovation in the last few days around what people can do with Stable Diffusion, which matches the quality of models like OpenAI's GLIDE and DALL-E 2, MidJourney of Google's Party, or Imagen by using an open source and hyperefficient architecture.



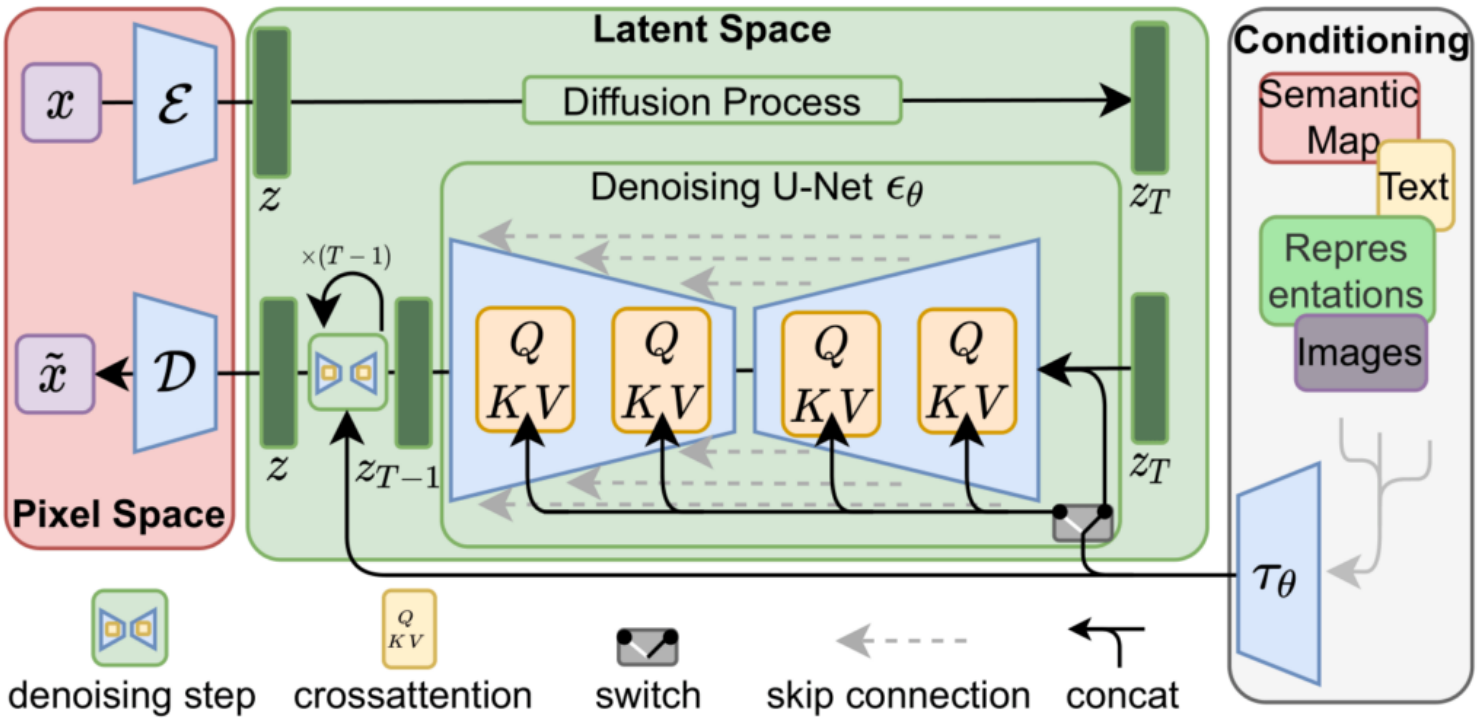
The Model Architecture

Stable Diffusion is powered by Latent Diffusion, a cutting-edge text-to-image synthesis technique. This method was described in a paper published by AI researchers at the Ludwig Maximilian University of Munich titled "**High-Resolution Image Synthesis with Latent Diffusion Models.**"

Latent diffusions are a simple way of saying that diffusion models (DMs) achieve state-of-the-art synthesis results on image data and more by breaking down the process of making an image into a series of applications of denoising autoencoders. Also, the way they are made

lets them be used immediately for image editing tasks like inpainting without having to be retrained.

But because these models usually work directly in pixel space, optimizing powerful DMs can take hundreds of GPU days, and inference is expensive because evaluations are done one at a time.

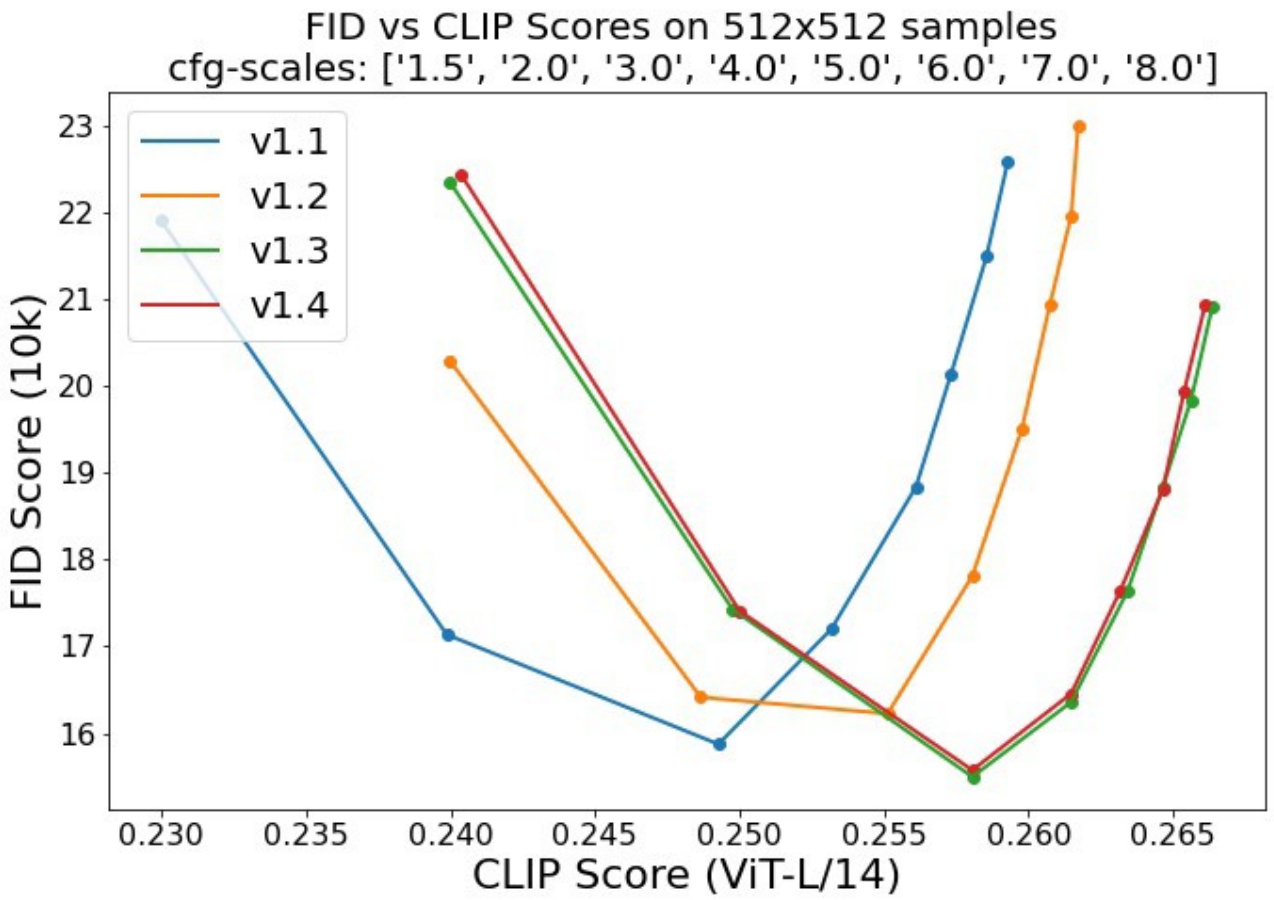


LDMs are conditioned either via concatenation or by a more general cross-attention mechanism. Source: High-Resolution Image Synthesis with Latent Diffusion Models

Stability AI put DMs in the latent space of powerful pre-trained autoencoders so that they could be trained with limited computing resources without losing their quality or flexibility.

Unlike previous work, training diffusion models on such a representation makes it possible for the first time to reach a near-optimal point between reducing complexity and downsampling space, which greatly improves visual fidelity.

By adding cross-attention layers to the model architecture, Stability AI turned diffusion models into powerful and flexible generators for general conditioning inputs like text or bounding boxes, and high-resolution synthesis became possible in a convolutional way.



Evaluations with different classifier-free guidance scales (1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0) and 50 PLMS sampling steps show the relative improvements of the checkpoint Source: High-Resolution Image Synthesis with Latent Diffusion Models

Stability AI latent diffusion models (LDMs) outperformed pixel-based DMs in several tasks, such as unconditional image generation, inpainting, and super-resolution while requiring much less computing power.

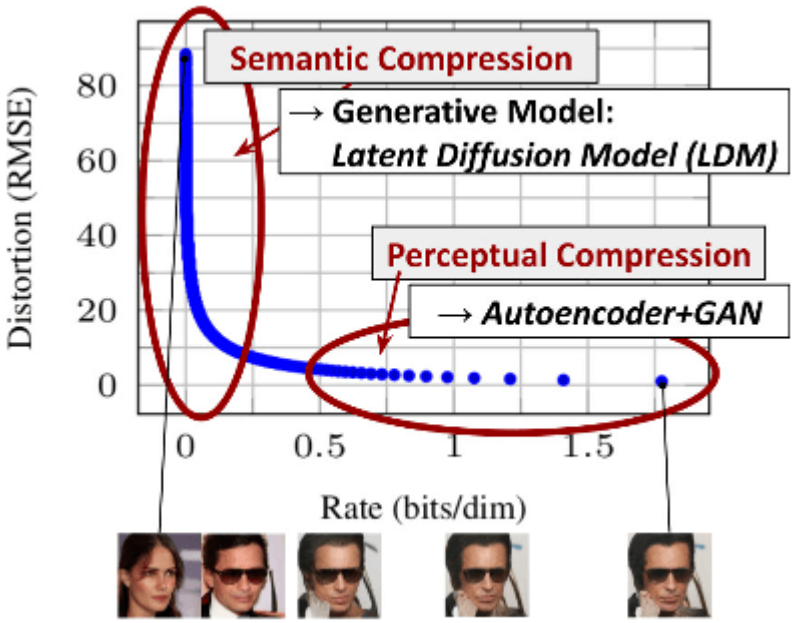
The model was trained on 4,000 A100 Ezra-1 AI ultraclusters for over one month, reaching 2.225,000 steps at resolution 512x512 on “laion-aesthetics v2 5+” and 10% dropping of the text-conditioning with improved classifier-free guidance sampling, and has over 1000 beta testers creating around 1.7 million images per day.

The perceptual compression phase uses autoencoders to remove high-frequency details, while the semantic compression phase derives knowledge from the data’s composition.

The core architecture of LDMs revolves around separating the compressive and generative learning phases. LDMs rely on an autoencoder to learn a lower-dimension representation of the pixel space to accomplish this. Then, this latent representation is passed through the diffusion process, which adds noise at each step.

That phase’s output is fed into a denoising network based on the U-Net architecture with cross-attention layers. This denoising network employs some additional inputs, such as

semantic maps or additional image or text representations, in addition to the latent representation.



Illustrating perceptual and semantic compression: Most bits of a digital image correspond to imperceptible details. While DMs allow to suppress this semantically meaningless information by minimizing the responsible loss term, gradients (during training) and the neural network backbone (training and inference) still need to be evaluated on all pixels, leading to superfluous computations and unnecessarily expensive optimization and inference. We propose latent diffusion models (LDMs) as an effective generative model and a separate mild compression stage that only eliminates imperceptible details. Source: High-Resolution Image Synthesis with Latent Diffusion Models

LDMs are one of the most significant innovations in the text-to-image synthesis space. The release of Stable Diffusion may help advance research and development in this critical area of deep learning. Perhaps this release will compel OpenAI, Google, and Meta to speed up the open-source release of models such as DALL-E 2 or Imagen.

Because one of the main advantages of Stable Diffusion is its relatively lightweight architecture, you can run the model on your hardware or Google Colab, for example, if you have enough computing power.

The current version can run 10 GB of VRAM on consumer GPUs and generate images with 512x512 pixels in a matter of seconds. Of course, if you have less computing power, you will have to wait a little longer, but trust me... it is well worth the wait, considering the result.

How about possible biases and misuses?

Even though it's impressive to be able to turn text into an image, and I'm very excited about this AI revolution, I want everyone to know that this model may produce content that reinforces or exaggerates societal biases.

The model was trained on an unfiltered version of the LAION-400M dataset, which scraped non-curated image-text-pairs from the internet (of course, the researchers did some work to remove illegal content from the dataset) and is meant to be used for research.

Stable Diffusion is a very new area from an ethical point of view. Other AI systems that make art, like OpenAI's DALL-E 2, have strict filters for pornographic content.

As I said before, the open-source Stable Diffusion license forbids some uses, like exploiting minors, but the model itself isn't limited on a technical level.

Also, unlike Stable Diffusion, many people cannot make art of famous people. When these two things are combined, it could be dangerous because bad actors could make pornographic "deepfakes" that could, in the worst case, keep the abuse going or make someone look like they did something they didn't.

Women are often the most likely to be hurt by this. For example, a study done in 2019 found that about 90% of the non-consensual deepfakes, which make up 90% to 95% of all deepfakes, are by women.

Let's keep this in mind when working with these new powerful tools.

Other Articles you may want to read.

- These 9 Research Papers are changing how I see Artificial Intelligence this year.
- Are We Witnessing the Next Evolution of Artificial Intelligence?
- DALL-E 2: When AI transforms words into images
- The most impressive Youtube Channels for you to Learn A.I., Machine Learning, and Data Science.
- An Overview of Pathways Autoregressive Text-to-Image Model
- These are some of the best Youtube channels where you can learn PowerBI and Data Analytics for free.
- These 10 Algorithms Can Change Your Life — If You Work With Data
- About Dante, Michelangelo, and Stable Diffusion- Reimagining the Divine Comedy with AI
- 5 amazing books about A.I. that you must read.

- The Best MIT Online Resources for You to Learn A.I. and Machine Learning for Free.

Links, Resources, and References

- FROM RAIL TO OPEN RAIL: TOPOLOGIES OF RAIL LICENSES — https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses
- Stable Diffusion Model Card — https://huggingface.co/CompVis/stable-diffusion
- High-Resolution Image Synthesis with Latent Diffusion Models (A.K.A. LDM & Stable Diffusion) — https://ommer-lab.com/research/latent-diffusion-models/