Image Generative Model

Stable Diffusion

Credit to TA.Karin, TA.Nat

What are Diffusion Models?

- It is generative deep learning model using noise reduction method
- It reverses the process of adding noise to an image
- Usually use for text-to-image, but also img-to-img and inpainting
- More stable than GAN (no mode collapse)

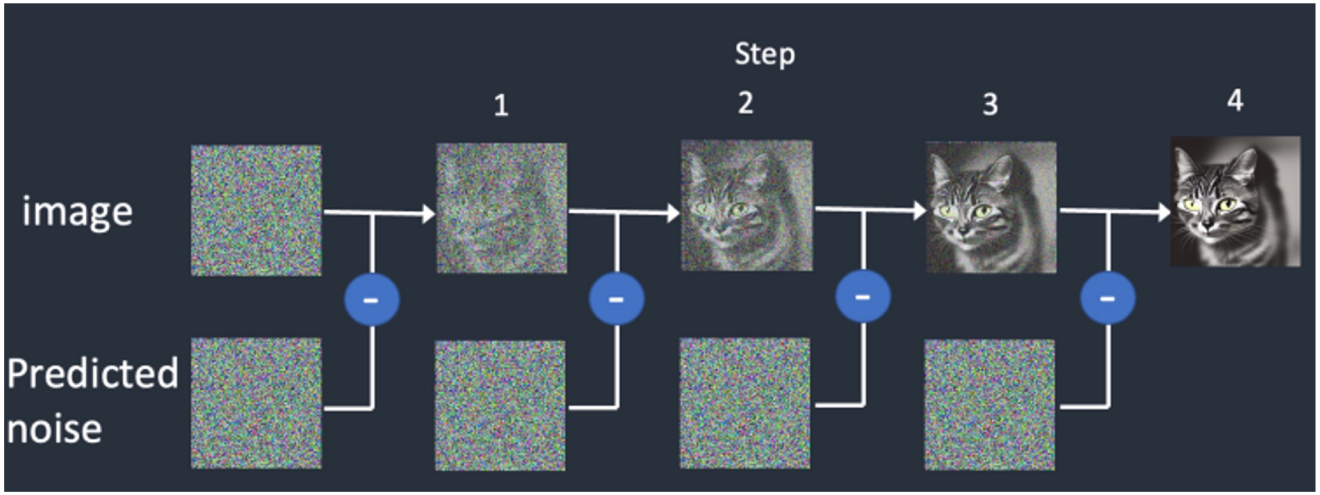


Figure 1 The Process of Noise Reduction

Stable diffusion

- Text-to-Image
- Latent diffusion model
- Open access
- Trained on 512 × 512 images from a subset of the LAION-5B database
- Uses a frozen CLIP ViT-L/14 text encoder
- Popular model → **stable-diffusion-v1-5**
- there is also **stable-diffusion-v2-1** (less popular)

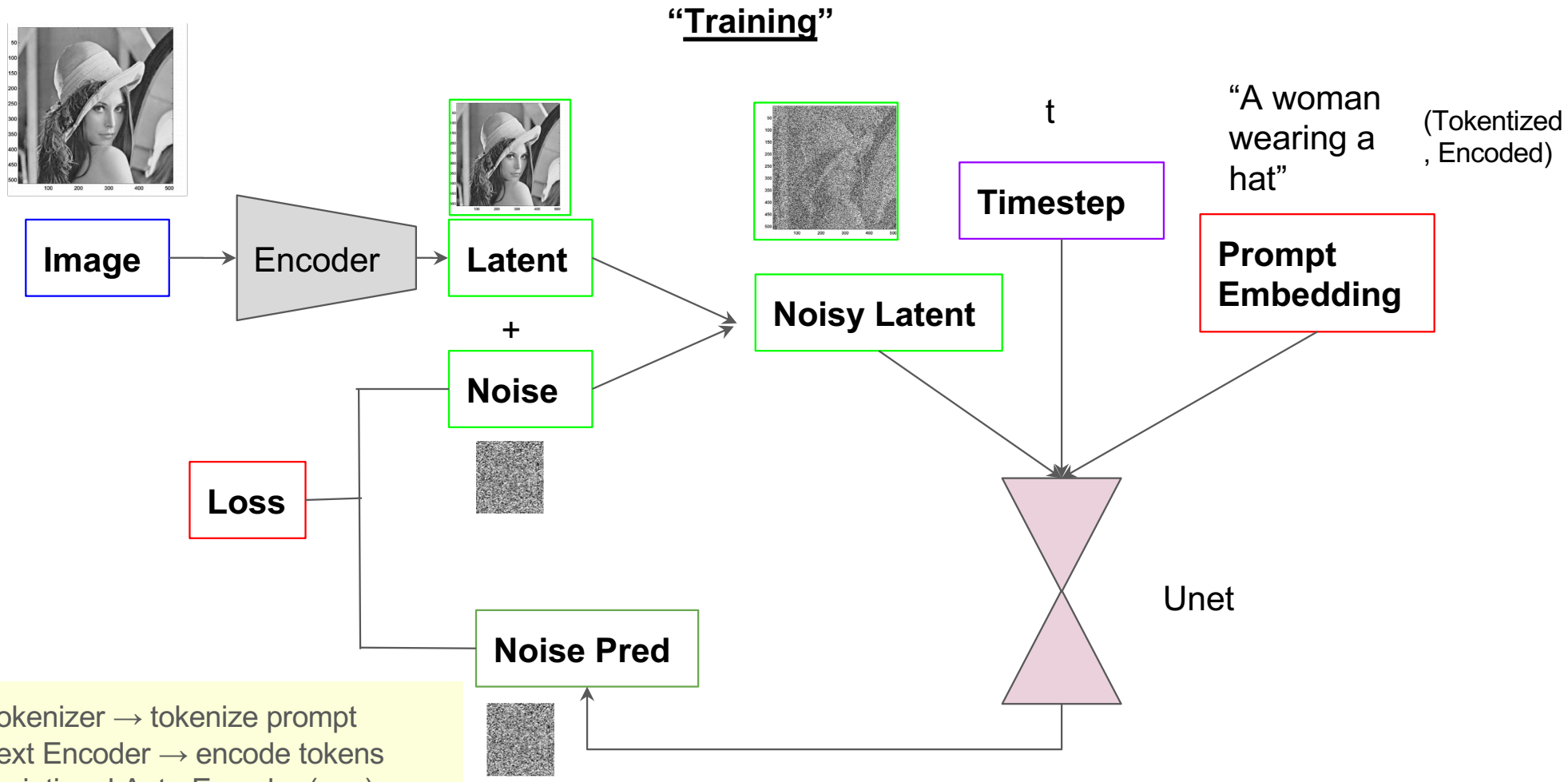
Outline

- How to train & generate image
- Training Techniques
- Code Demo

1) How to Train & Generate Image

A Latent Diffusion Model Comprises

1) Tokenizer → tokenize prompt
2) Text Encoder → encode tokens
3) Variational Auto-Encoder (vae) → map image to latent space
4) Noise Scheduler → generate noise
5) UNET → predict noise



1) Tokenizer → tokenize prompt
2) Text Encoder → encode tokens
3) Variational Auto-Encoder (vae)
   → map image to latent space
4) Noise Scheduler → generate noise
5) UNET → predict noise

Figure 4 Diffusion During Training

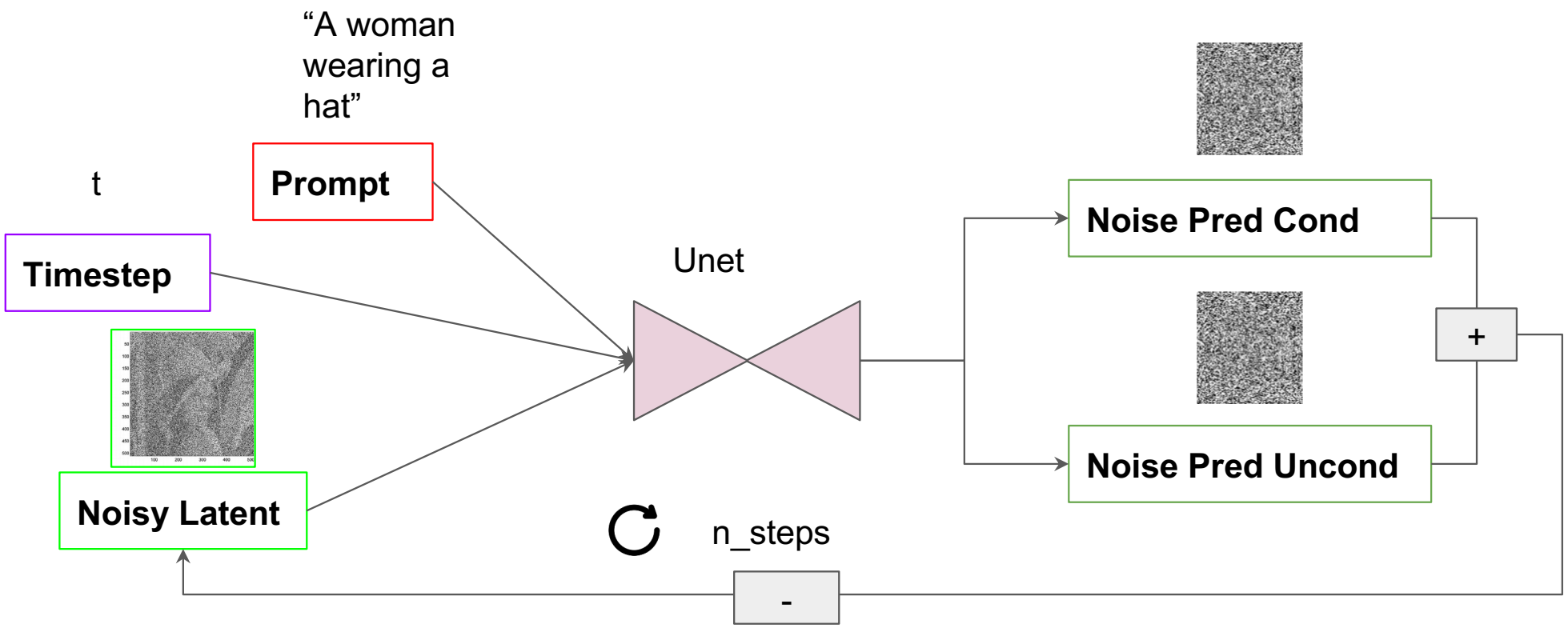


Figure 5 Diffusion Model During Inference

2) Training Techniques

Training Techniques

- Textual Inversion
- DreamBooth

Textual Inversion (2022)

- Use a small set of images (typically 3-5)
- The image depicts our target concept across multiple settings
- e.g. varied background or poses
- “We intervene in the embedding process and replace the vector associated with the tokenized string with a new, learning embedding”
- “In essence “injecting” the concept in to our vocabulary”

Cons of Textual Inversion (for X-ray image generation)

- This method only trains text encoder and not unet (image)
- It’s good when you want to give your “concept” a new style
- X-ray images are similar, changing text embedding is not enough

DreamBooth (currently using) (2022)

- trains unet (and text encoder if you want)
- only need 3-5 image per subject

Cons

- May overfit to training data
- Some subjects are easier to learn than others

Training Image

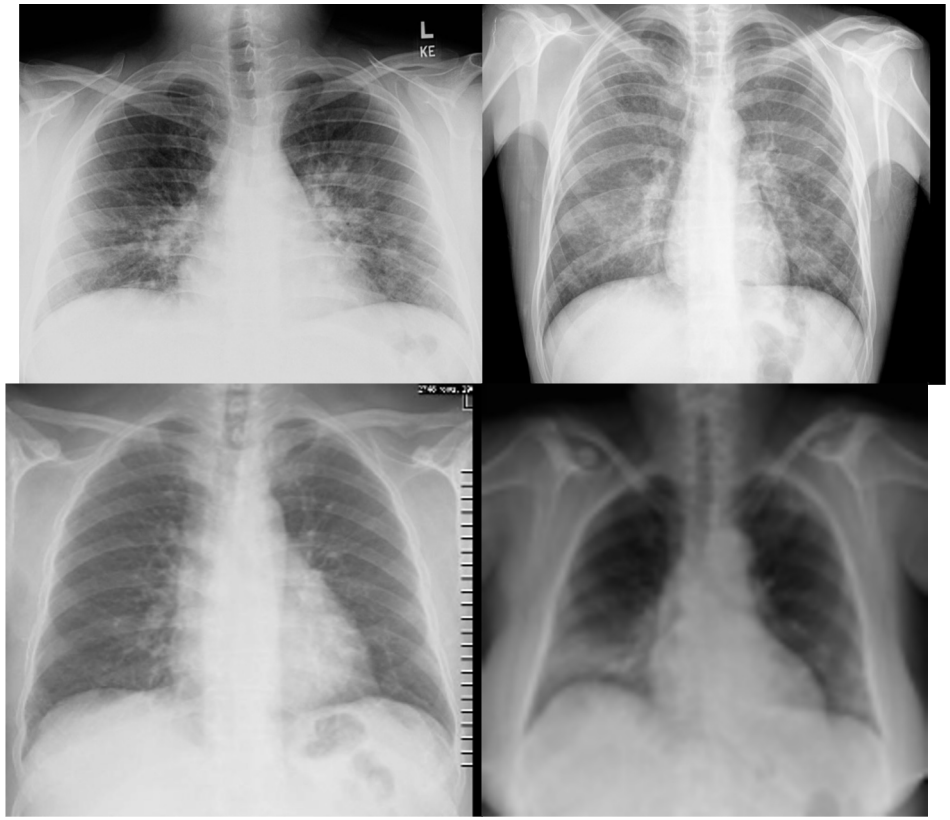


Figure 6 Training Image (COVIDx Dataset)

Textual Inversion Trained with 4 images

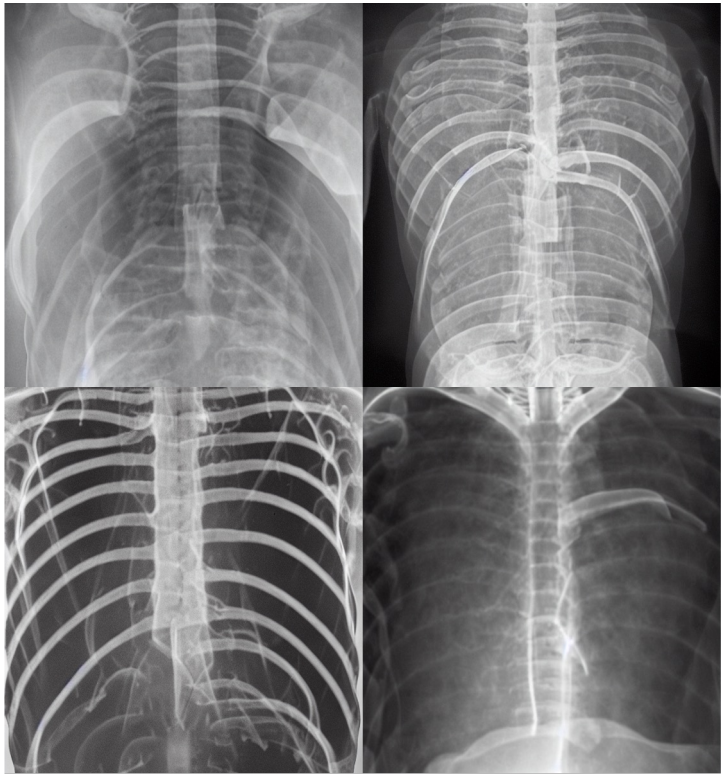


Figure 7 Generated Image, Trained by Textual Inversion

DreamBooth

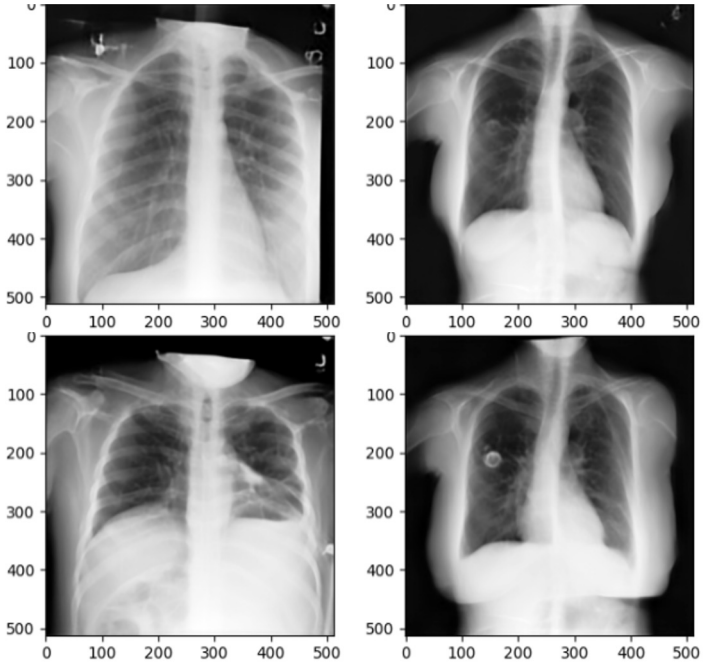


Figure 8 Generated Image, Trained by DreamBooth

Code Demo

Code Demo

- Link to Colab Notebook
- MONAI's code I use to train Diffusion

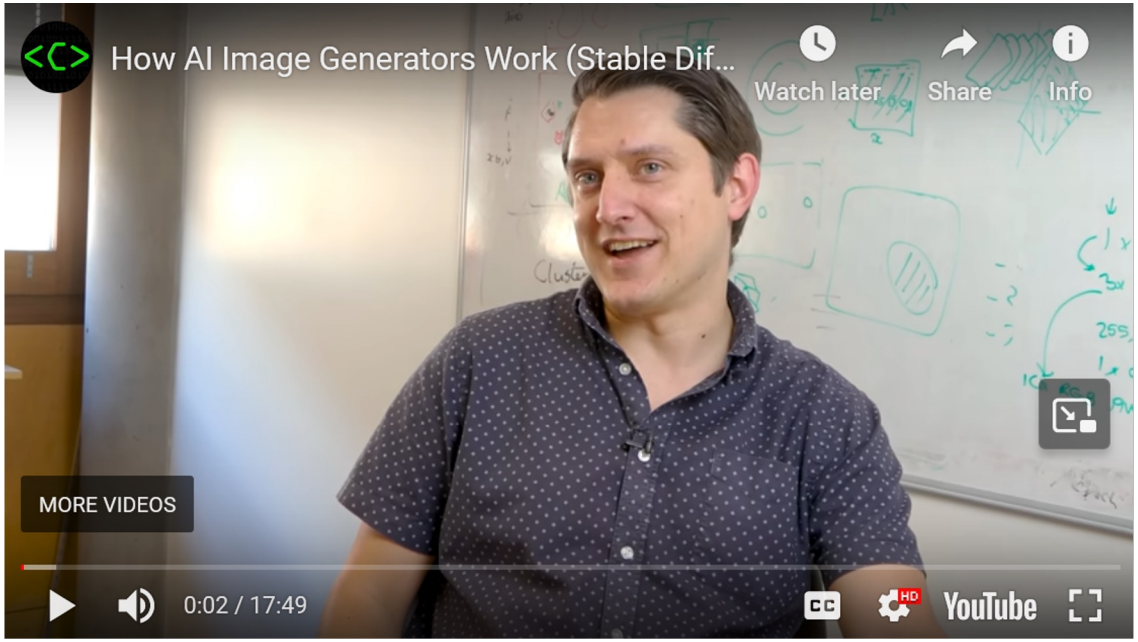
Van Goh Style, Man Playing Piano | Change Seed | Negative Prompt | Custom Model

Figure 9 Generated Image, Demo

Youtube Videos

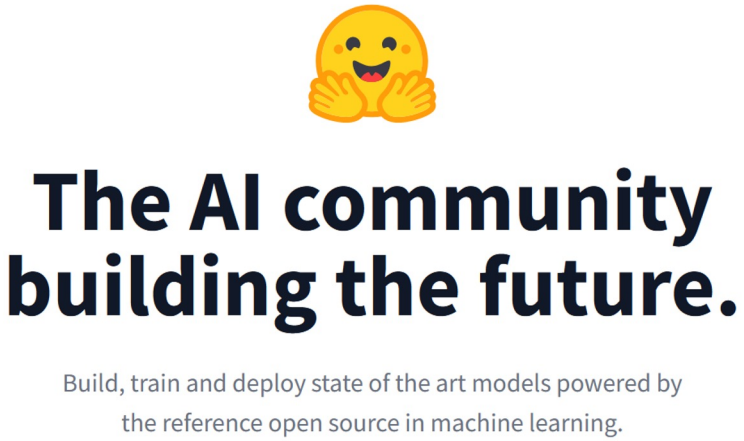
[How AI Image Generators Work]

[Stable Diffusion in Code]



Guide

[Huggingface's guide on dreambooth]



How I use stable diffusion

- There are 2 Jupyter notebooks → 1) Finetune and 2)GenPic
  - Finetune = load pretrained model, dataset → train → save finetuned model
  - GenPic = load finetuned model → Generate image using stable diffusion pipeline

FineTune Example

GenPic example