

DEEP DIVE INTO AI IMAGE GENERATION WITH STABLE DIFFUSION

Arjun Singh



INTRODUCTION

Models like Stable Diffusion allow anyone to produce realistic and imaginative images from simple text prompts. This technology has found widespread use across various industries, including:

- Entertainment companies for concept art and storyboarding
- Product designers to quickly prototype and visualize different variations
- Content creators crafting unique visuals for blogs and social media
- Architects visualizing interior designs rapidly
- E-commerce platforms showcasing products in diverse settings

DIFFUSION PROCESS

- **Initial state:** Start with a noisy image.
- **Step by step denoising:** Iteratively reduces the noise, guided by the input prompts. The amount of noise to be removed at each step is learned during training. For example, to generate an image of a cat, the model is trained on many cat images and gradually adds noise to these images until they become pure noise. By learning how much noise to add at each step, the model essentially learns the reverse process—starting with a noisy image and progressively removing noise to arrive at the desired image.

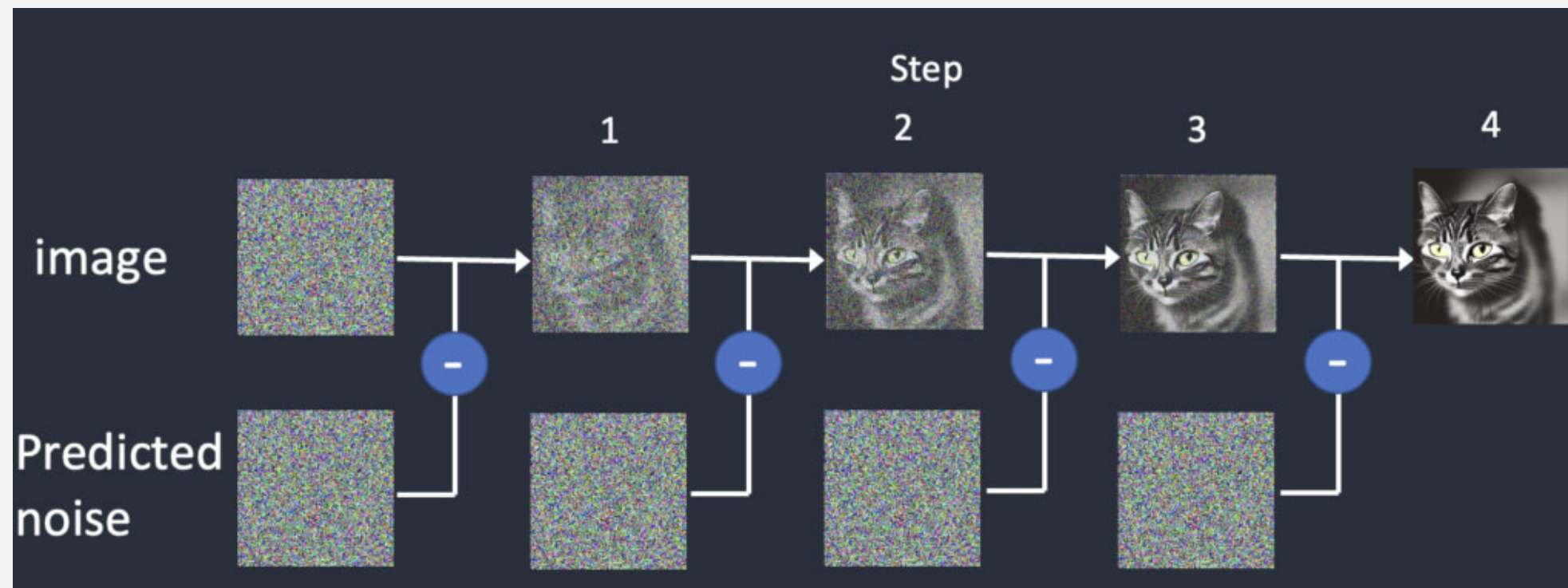


Image Source: <https://stable-diffusion-art.com/how-stable-diffusion-work/>

DIFFUSION PROCESS (CONTD.)

- **Latent space processing:** Processing images in pixel space is computationally costly due to the large amount of pixel data. diffusion models operate in a compressed "latent space," which allows for a more efficient way to capture the key features of an image. The latent space is 48 times smaller than the pixel space, hence model training and inference both occur in this lower dimension.

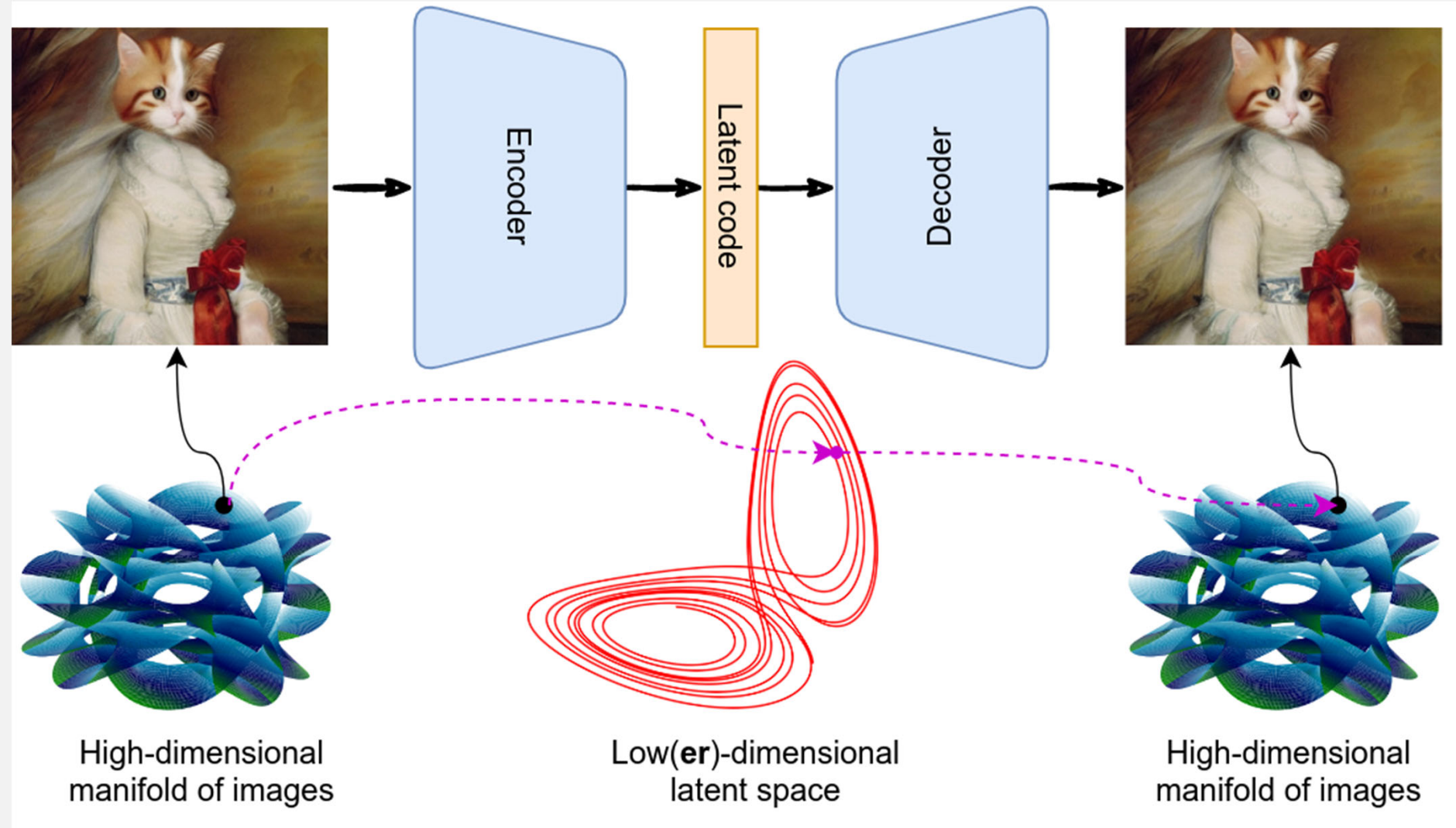


Image Source: <https://synthesis.ai/2023/03/21/generative-ai-ii-discrete-latent-spaces/>

DIFFUSION PROCESS (CONTD.)

- **Text Conditioning:** Without a text prompt, a user cannot control whether an image of a dog or cat gets generated. This is where text conditioning with prompting comes into play. The text prompt guides the generation process and determines how much noise to subtract at every step to arrive at the desired output.

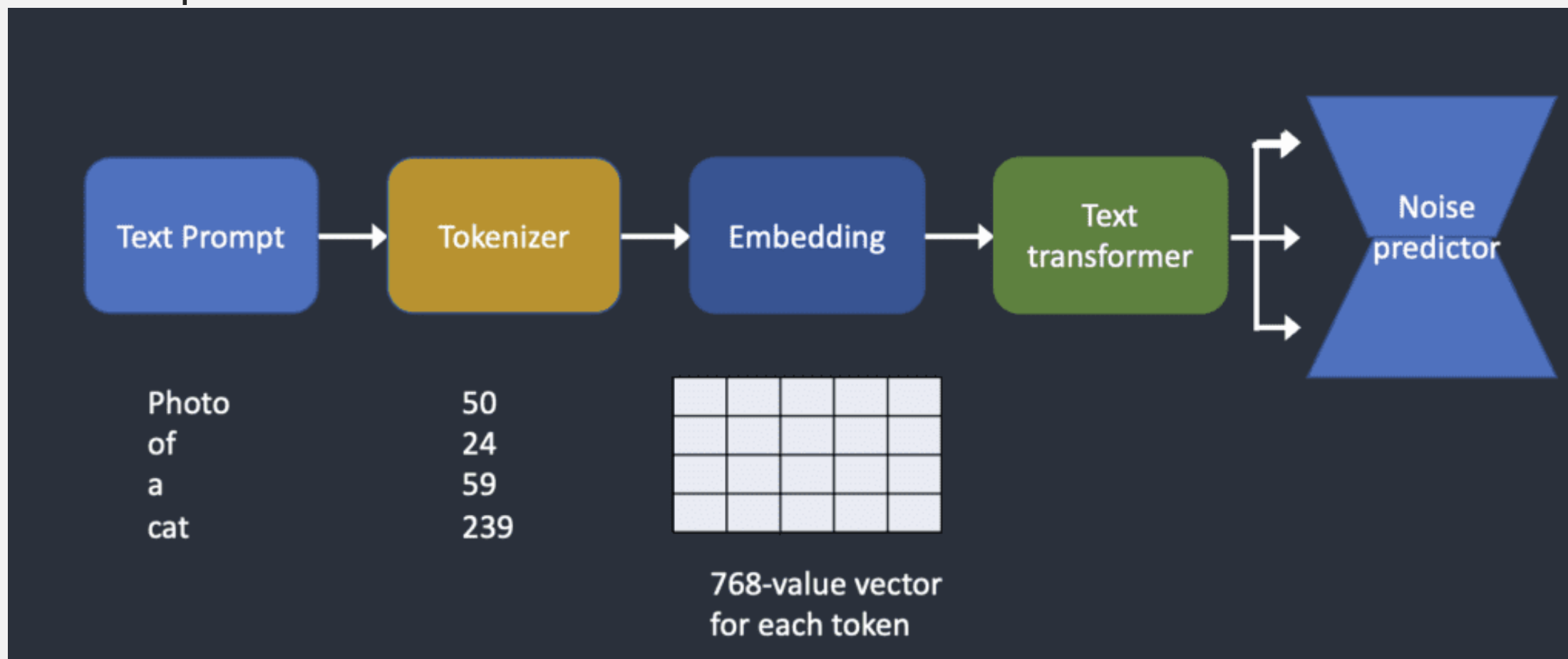


Image Source: <https://stable-diffusion-art.com/how-stable-diffusion-work/>

SAMPLERS

- **Sampling methods:** Samplers are responsible for carrying out the step-by-step denoising. This process is called sampling as it generates a new sample image in each step. Different algorithms (DDIM, Euler a, DPM++) control how the model performs sampling and is generally a trade-off between accuracy and speed.



Image Source: <https://stable-diffusion-art.com/how-stable-diffusion-work/>

GETTING SETUP WITH STABLE DIFFUSION

Deployment Type	Local Machine	Cloud Services
Web interfaces such as - <ul style="list-style-type: none">• AUTOMATIC1111 WebUI• Comfy UI	Download and install locally Requires compatible GPU Full control over installation No usage costs Limited by local hardware	Available through services like Think Diffusion, Jarvis Labs Pay-as-you-go or subscription pricing No local hardware requirements Can access high-end GPUs Cloud VMs with Python environment
Python with model files	Direct access to model files Maximum customization Requires Python programming knowledge Most flexible but highest technical barrier Limited by local hardware	Cloud notebooks like Google Colab, Kaggle, AWS Sagemaker Can utilize powerful cloud GPUs Some platforms offer free tiers Programming knowledge still required

PROMPTING FUNDAMENTALS

- AI image generation starts with crafting precise prompts – the skill of writing text descriptions that steer the model toward the desired result.
- A good prompt includes specific details about the
 - subject,
 - style,
 - lighting,
 - perspective, and
 - composition.
- Instead of saying "generate an image of a forest," a more detailed prompt could be "a serene forest with towering ancient trees, sunlight filtering through the leaves, soft mist rising from the forest floor, in a magical, fantasy-inspired style."

SUBJECT

- Who is the subject, and what are they doing?



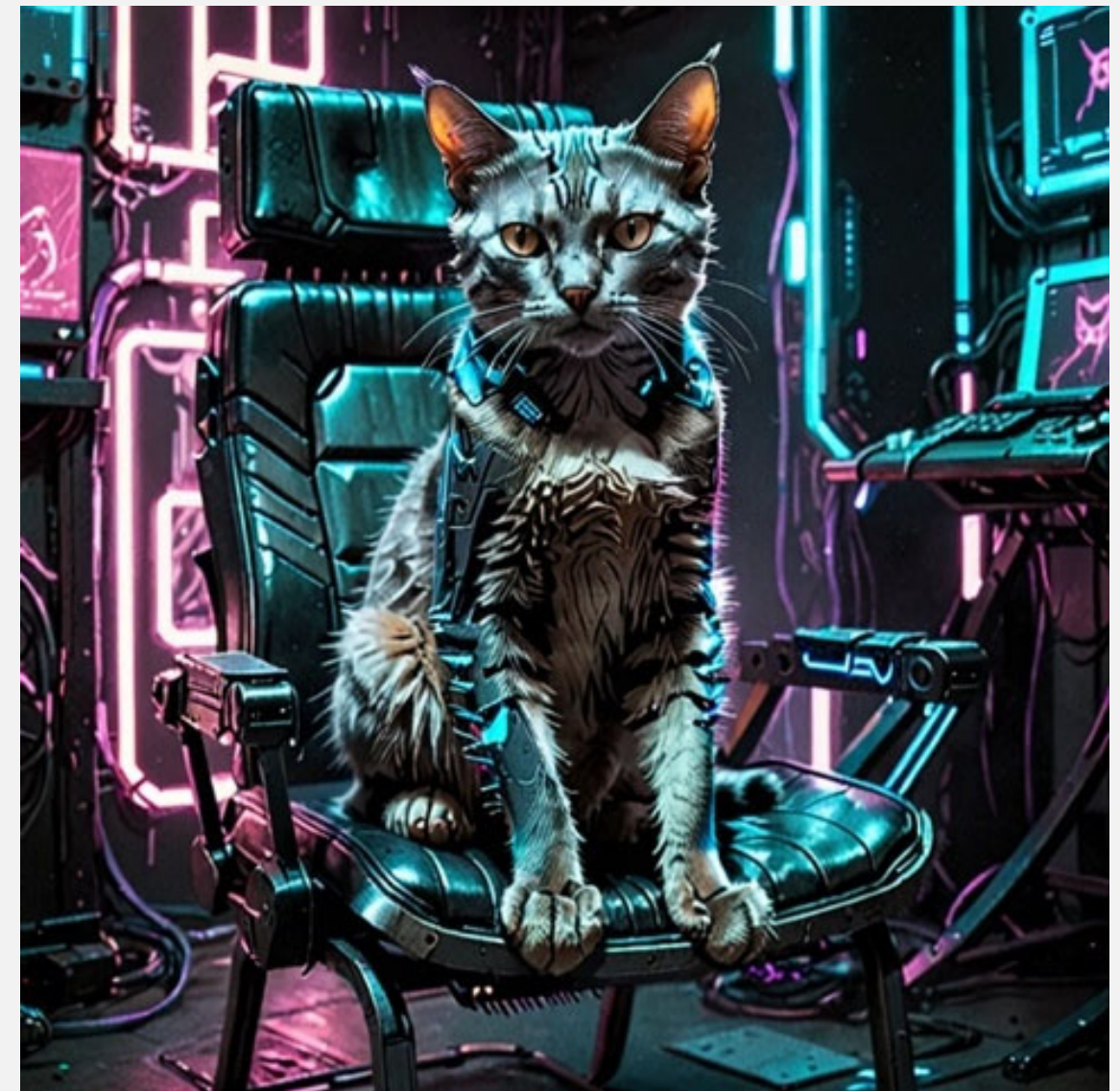
Prompt: A cat sitting on a chair

STYLE

- The visual style of the image (3D render, photorealistic, etc.)



Prompt: A cat sitting on a chair, oil painting style



Prompt: A cat sitting on a chair, cyberpunk style

COMPOSITION

- Composition of the image (foreground, background, perspective)



Prompt: A cat sitting on a chair, mountains in the background.



Prompt: A cat sitting on a chair, mountains in the background, low angle shot.

DEPTH AND LIGHTING



Prompt: A cat sitting on a chair, mountains in the background, soft lighting casting a shadow to the left.



Prompt: A cat sitting on a chair, mountains in the background, shallow depth of field, strong background blur.



A dog running, snowy forest in the background, anime style, sun casting shadow to the right

NEGATIVE PROMPT

- **Negative prompts** specify what to avoid. Common negative prompts include "blurry, distorted, low quality, poor anatomy, extra fingers" which help models avoid typical generation artifacts.



Prompt: autumn in paris, ornate, beautiful, atmosphere, vibe, mist, smoke, fire, chimney, rain, wet, pristine, puddles, melting, dripping, snow, creek, lush, ice, bridge, forest, roses, flowers, by stanley artgerm lau, greg rutkowski, thomas kindkade, alphonse mucha, loish, norman rockwell. **Negative Prompt:** People
(<https://stable-diffusion-art.com/how-to-use-negative-prompts/>)

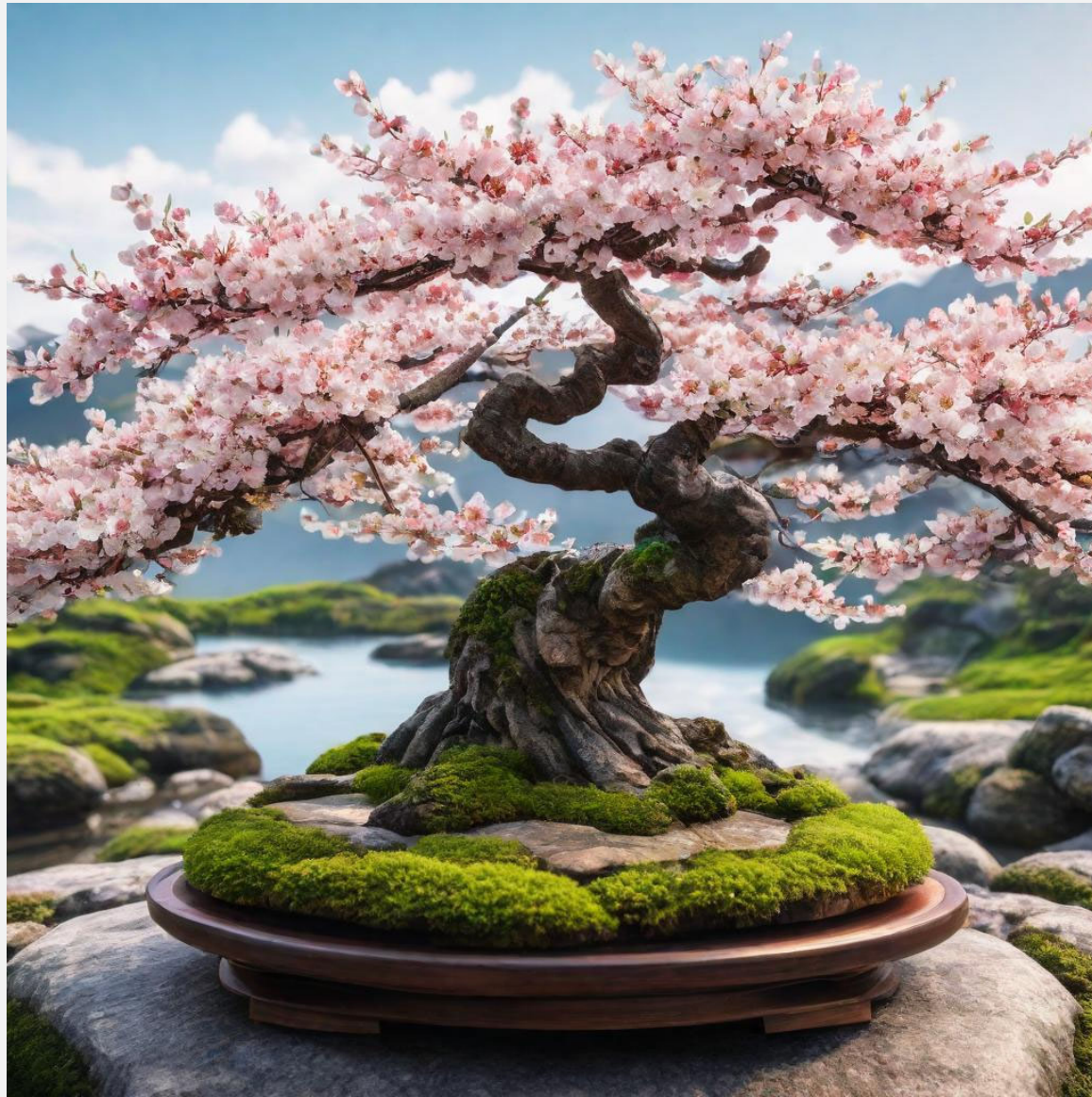
CLASSIFIER FREE GUIDANCE SCALE (CFG)

CFG Scale determines how closely the model follows the given prompt:

- **Low values (1-7):** Greater artistic freedom, resulting in visually appealing but less accurate interpretations.
- **Medium values (7-12):** A balanced setting, often recommended for general use.
- **High values (12-20+):** Strong adherence to the prompt, yielding precise outputs but sometimes reducing natural aesthetics and causing oversaturation.

CFG SCALE (CONTD.)

Prompt: cherry blossoms, bonsai, Japanese style landscape, high resolution, 8k, lush greens in the background.



CFG 5: Lesser prompt adherence, fewer greens in background, no Japanese landscape.



CFG 13: Stricter prompt adherence, more lush greens and presence of Japanese architecture in background.

GENERATION STEPS

Steps in AI image generation represent the number of denoising iterations during the image creation process. Starting from pure noise, the model progressively refines the image through multiple passes.

- **Low steps (0-20):** Faster generation; Rough, less detailed images; Minimal computational requirements.
- **Medium steps (20-50):** Balanced detail and generation time; Most models' default range; Suitable for most general purposes.
- **High steps (50+):** More refined, detailed output; Longer generation time; Increased computational intensity.

GENERATION STEPS (CONTD.)



Iterating sampling steps from 2-40, image of a cat (Image Source: <https://stable-diffusion-art.com/samplers/>)

SEED VALUES

Every image generation starts with a **randomization seed**, a numerical value that sets the initial conditions:

- **Seeds are long integers** (e.g., 1234567890).
- **Using the same seed, prompt, and parameters** produces identical or nearly identical images.
- **Saving seeds** enables revisiting successful generations or tweaking variations of promising results.
- **Seed values don't affect quality**—they simply provide different starting points.

SEED VALUES (CONTD.)



Batman looking in the mirror, tweaking seed by 1.

REPRODUCIBILITY WITH SEED VALUES



A portrait of a girl smiling.



A portrait of a girl laughing.

ADVANCED FEATURES – CONTROL NET

- ControlNet extends diffusion models by allowing precise control over structural elements.
- It is a form of image conditioning where an additional condition is applied to the prompt that guides the generation process.
- With control nets, we can preserve certain structural components, such as poses, outlines or spatial features and recreate new images with similar structural integrity.
- Some common control nets:
 - Pose Control
 - Canny Edge Detection
 - Segmentation
 - Depth Maps

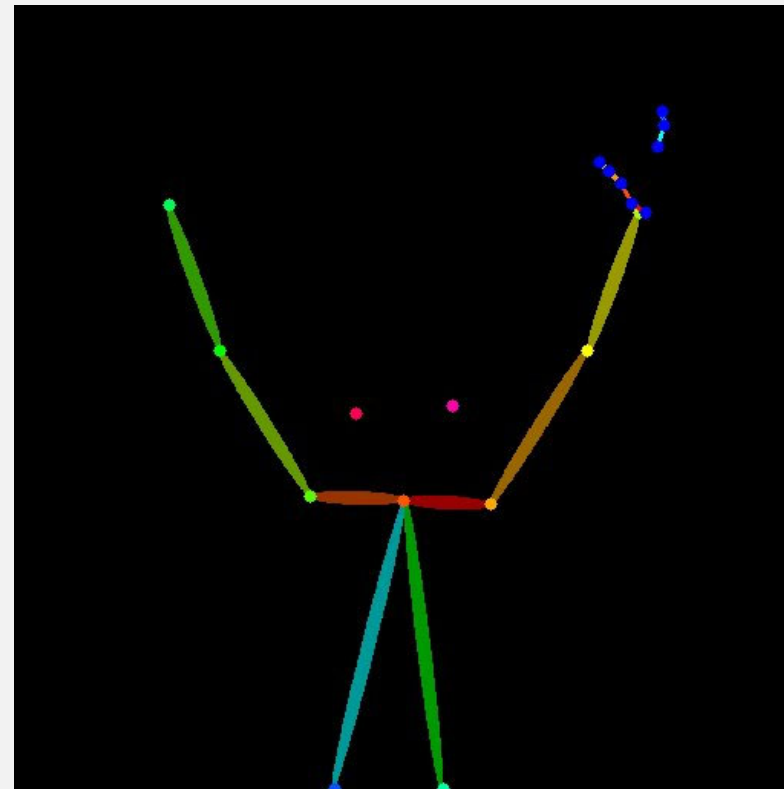
POSE CONTROL

Pose control: Generate images with figures matching specific poses. In the example below, the control net model maps the pose of the man in the input image (it can even interpret the direction the man is facing), and applies the same pose to the generated image with prompt – ‘A toy soldier’.

This is specifically useful when we want to maintain pose consistency across fashion models.



Input Image



Pose captured by model



A toy soldier

CANNY EDGE DETECTION

Canny edge detection: Preserve specific edges and boundaries during generation. This control net model captures the outline and boundaries of the input image and applies that as an additional condition to the generated image, thereby preserving the shape and physical features of the subject.



Input Image



Canny Edge Detection



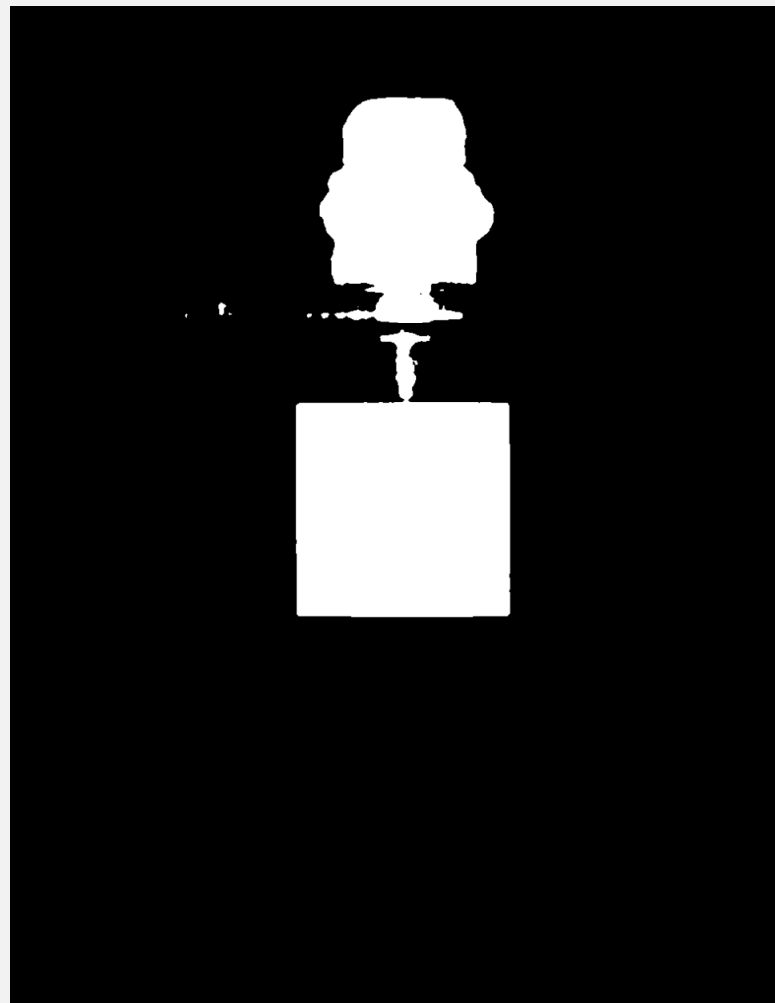
A perfume bottle placed in sand

SEGMENTATION

Segmentation: Define precise regions of the image to preserve to maintain accuracy. The technique below uses two control nets, first the input image is passed through a canny edge detector to get the contour of the bottle and then specific parts of the bottle are segmented to ensure they are preserved in the final image.



Input Image



Canny Edge Detection



A perfume bottle placed in sand

SEGMENTATION (CONTD.)



Snow



Water



Forest

DEPTH MAPS

Depth map control: Map spatial relationships within images to re-imagine spaces with prompts. The depth map of the room maps every element and its relative location to other elements. This can be used to generate multiple interior designs of the same room, saving hours and expenses for architecture firms.



Input Image



Depth map of room

DEPTH MAP (CONTD.)




A bohemian themed bedroom, photorealistic, 4k, high quality

STABLE DIFFUSION COMMUNITY


CIVITAI 🤖 >

Models ▾ Search Civitai / 🔍

Create ▾ 🖨️ 🔄 🔔 1 💬 👤 ⚡ 100




OpenDalle
DataVoid
↓ 5.5K 👁 87 🗲 18 ⚡ 2.4K 🍌 462




PixelWaveTurbo - Excellent images in 5 steps!
humblemikey
↓ 8.7K 👁 129 🗲 18 ⚡ 7.9K 🍌 707


Become a Member to turn off ads today.
[Do It ▶](#)




CIVITAI
Please support Civitai and Creators by disabling ad block




Ultrium
DSchroeder
V9.333.NSFW+SFW.SDXL.VAE
↓ 19.1K 👁 327 🗲 33 ⚡ 8.2K 🍌 1.2K




RealMix XL
waterdrinker
↓ 7.5K 👁 146 🗲 5 ⚡ 40 🍌 643




Samaritan 3d Cartoon
PR PromptSharingSamaritan
↓ 59.3K 👁 724 🗲 43 ⚡ 847 🍌 6K




ICBINP XL
unidentified_author
↓ 16.5K 👁 310 🗲 31 ⚡ 11.2K 🍌 1.1K



Psy Animated XL
RelicVisuals
↓ 3.4K 👁 81 🗲 3 ⚡ 0 🍌 476



GalaxyTimeMachine's GTM "XLPlus"
galaxytimemachine
↓ 7K 👁 160 🗲 18 ⚡ 36.5K 🍌 588

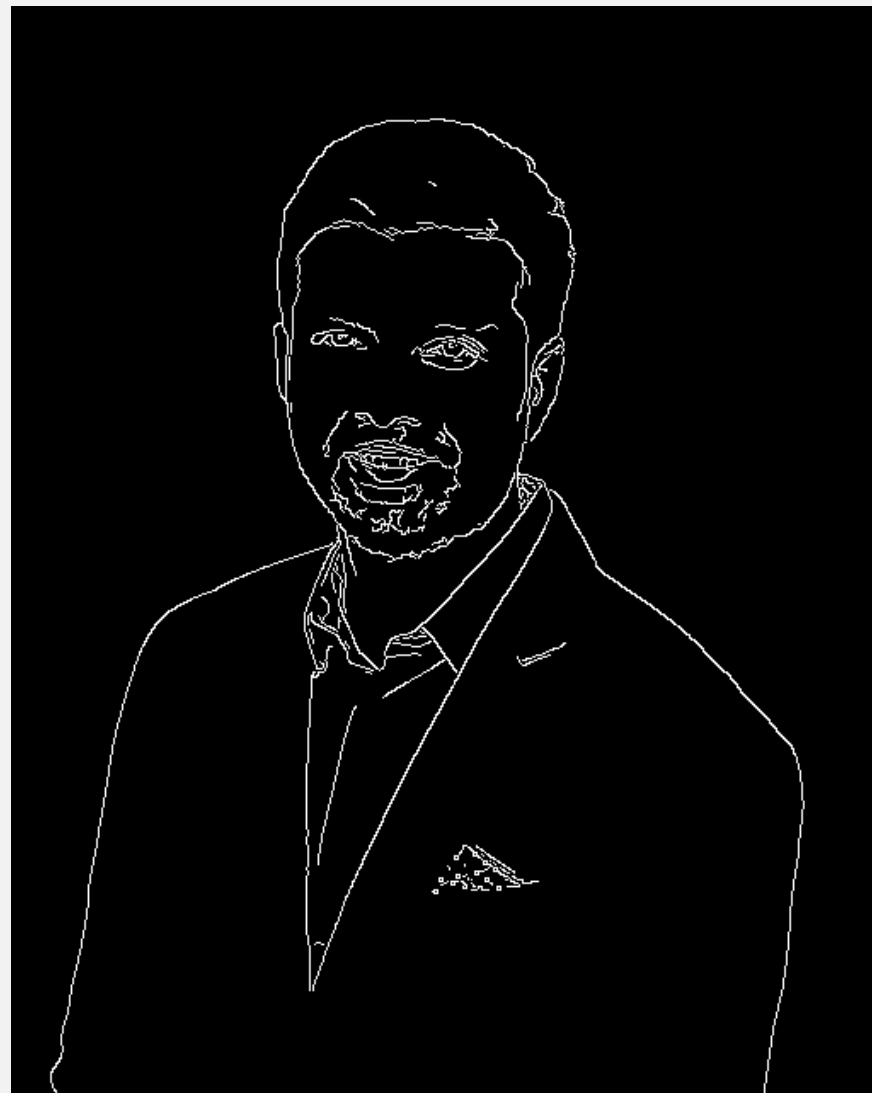


Moxie Diffusion XL
moxie1776
↓ 17.3K 👁 372 🗲 24 ⚡ 24.9K 🍌 993

GETTING THE GHIBLI STYLE IMAGE



Input Image



Prompt: A man in studio ghibli style, gentle smile, light stubble, black hair, wearing a white shirt and navy suit



WHAT'S NEXT? - FINE TUNING

- Fine tuning takes a model which has been trained on a diverse dataset, and then trains it a bit more on the dataset you are specifically interested in.
- This is achieved by providing additional images of a custom subject, which can be a person, object or even a style (Ghibli, for example).
- We define the custom subject using an instance prompt – this is a token that the model has not been trained on, and going forward it identifies this token as the additional images injected into the training process.
- For example, we can feed a model with multiple images of a person and give this individual an instance prompt '**awhd**' or any word that the model is not familiar with.
- After training, if we want to generate an image of that person, we simply add this instance prompt – 'An **awhd** man wearing a white shirt walking in a hallway'.



THANK YOU!



LinkedIn: <https://www.linkedin.com/in/arjsngh/>

Email: I9arjun89@gmail.com