

Pre-Trained Language Models for Image Generation

Miguel Domingo

midobal@prhlt.upv.es

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Tools and Applications of Artificial Intelligence

MIARFID, March 12, 2024

Outline

1. Introduction

2. Image generation

3. Video generation

4. Controversy

Outline

1. Introduction

2. Image generation

3. Video generation

4. Controversy

Goal

React to a text input (known as *prompt*) by generating new related images.

Example

Prompt: *Pink elephants on parade.*

Example

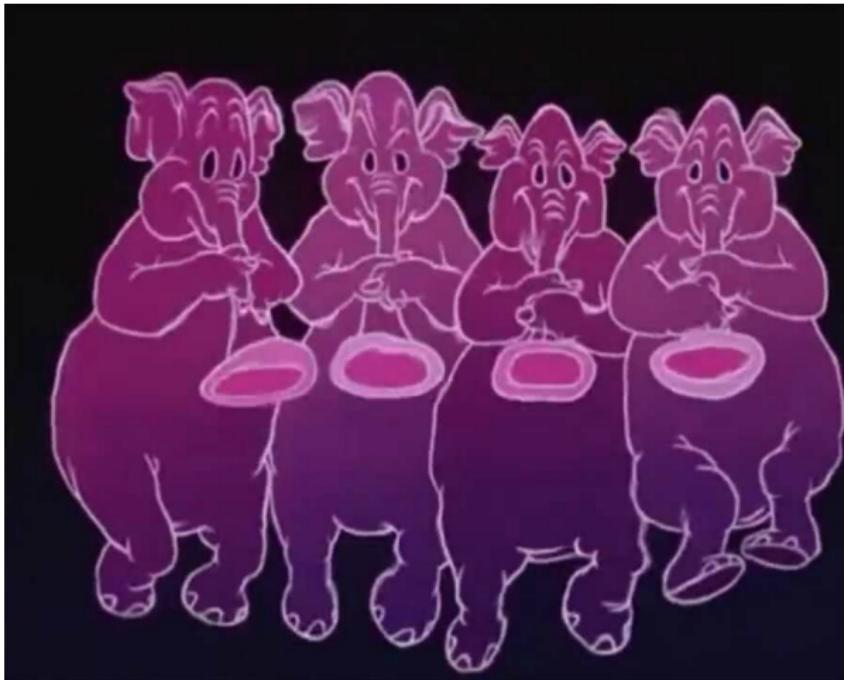


Image from Dumbo (1941).

Outline

1. Introduction

2. Image generation

- Dall-e
- Midjourney
- Leonardo AI
- DreamFusion
- Adobe Firefly
- Stable Diffusion

3. Video generation

Dall-e

- Developer: OpenAI (Ramesh et al., 2022).
- GPT-3 (Brown et al., 2020).
- Preventing harmful generation.
- Ensure content policy.
- Credit system.
- Beta no longer available.

Dall-e

Example



Pink elephants on parade.

Dall-e

Example

More examples: <https://openai.com/product/dall-e-2>.

Dall-e

Bing Image Creator

- Based on *Dall-e*.
- Integrated into *Microsoft Edge* and other *Bing* products.

Dall-e

Bing Image Creator

Bing

Can you create me an image of an astronaut walking through a galaxy of sunflowers?

Sure, I'll use Image Creator to draw that for you.



Made with Image Creator

Change the astronaut to a cat Change the sunflowers to roses Add a moon in the background

Type message

Midjourney

- Developer: independent research lab (Midjourney, Inc.).
- Discord-based.
- Subscription plan.
- Beta no longer available.

Examples: <https://www.midjourney.com/showcase/recent/>.

Leonardo AI

- Game assets generation.
- Artists tools.
- Use of pre-trained models.
- Train custom models.
- Content production platform.

Examples: <https://leonardo.ai/>.

DreamFusion

- Developer: Google Research (Poole et al, 2022).
- 3D assets generation.
- Text-to-3D using 2D Diffusion.

Examples: <https://dreamfusion3d.github.io/gallery.html>.

Adobe Firefly

“Generative AI made for creators”

- Developer: Adobe.
- Trained on *Adobe Stock* images:
 - ▶ Openly license content.
 - ▶ Public domain.
- Designed to generate content safe for commercial use.
- To be integrated into Adobe products.

Adobe Firefly

Example



Adobe Firefly

Image tools

- Context-aware image generation.
- Vector, brushes and textures generations from few words and sketches.
- Template generation.
- 3D modeling.

Examples:

<https://www.adobe.com/sensei/generative-ai/firefly.html>.

Adobe Firefly

Video tools

- Text to color enhancements (e.g, “Make this scene feel warm and inviting”): change color schemes, time of day, or even the seasons.
- Advanced music and sound effects: generation of royalty-free custom sounds and music to reflect a certain feeling or scene.
- Stunning fonts, text effects, graphics, and logos: generation of subtitles, logos and title cards and custom contextual animations.
- Powerful script and B-roll capabilities: acceleration of production workflows to automatically create storyboards and pre-visualizations.
- Creative assistants and co-pilots: master new skills and accelerate processes from initial vision to creation and editing.

Examples:

<https://blog.adobe.com/en/publish/2023/04/17/reimagining-video-audio-adobe-firefly>

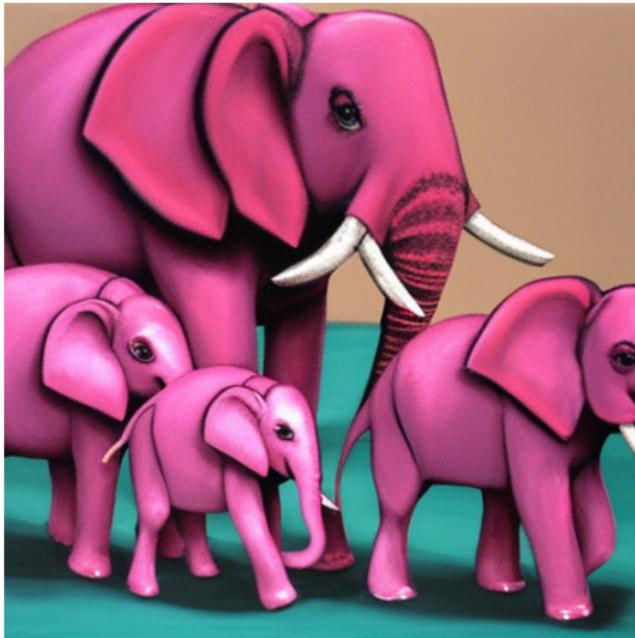
Stable Diffusion

Introduction

- Developer: Stability AI (Rombach et al., 2022).
- Trained on LAION-5B (Schuhmann et al., 2022).
- Open source.

Stable Diffusion

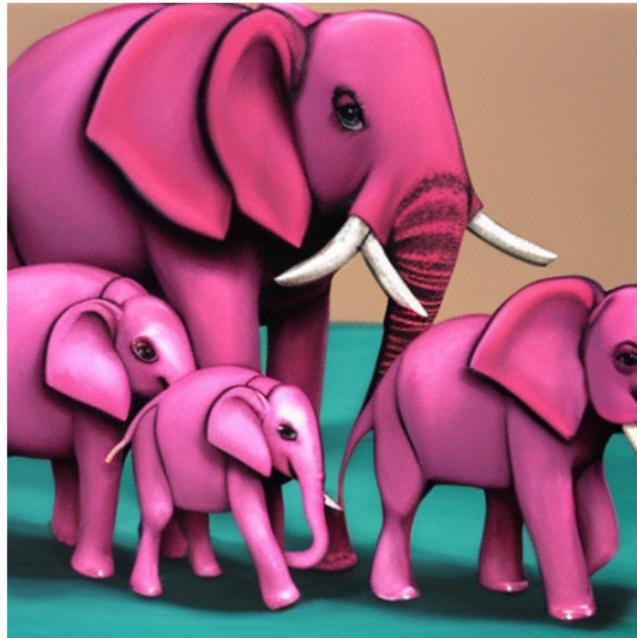
Example



Pink elephants on parade.

Stable Diffusion

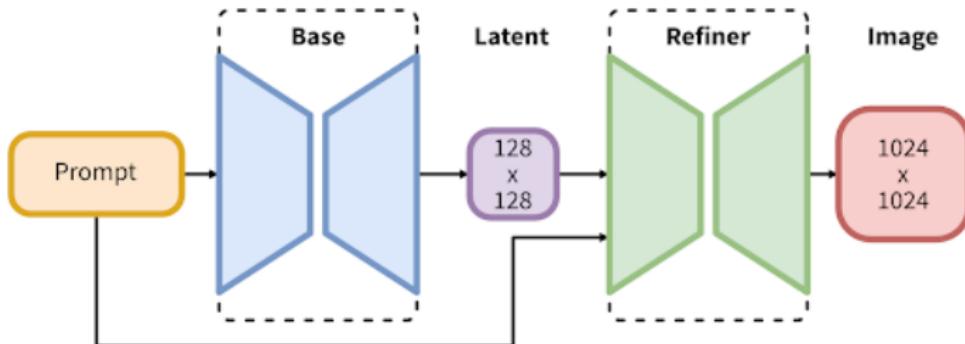
Example (SDXL)



Pink elephants on parade.

Stable Diffusion

SDXL



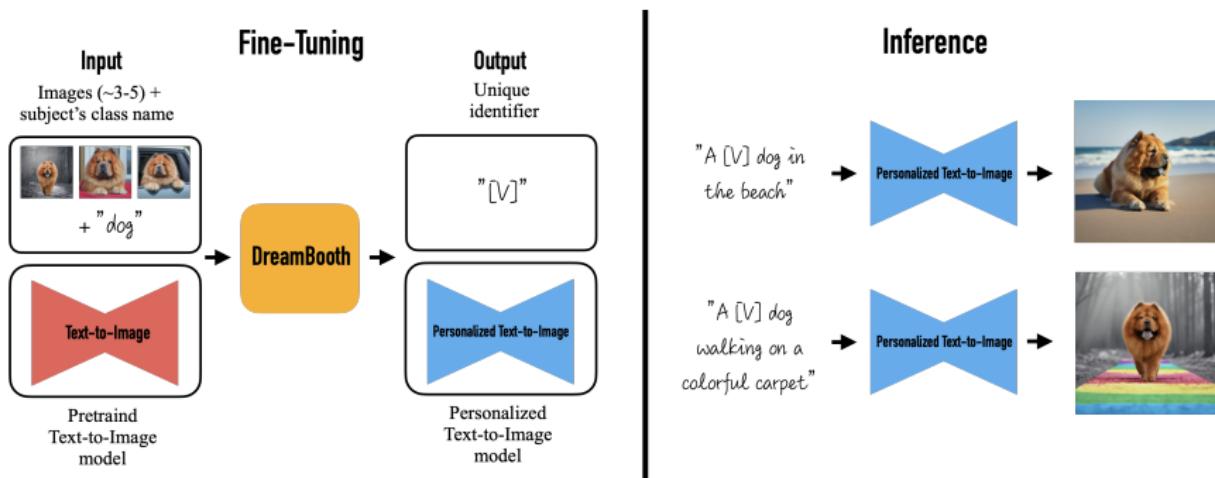
Stable Diffusion

DreamBooth

- A technique to fine-tune diffusion models by injecting a custom subject to the model (Ruiz et al., 2022).
- Training time: ~ 1 hour.
- Model size: Gigabytes.

Stable Diffusion

DreamBooth: Approach



Stable Diffusion

DreamBooth: Art rendition

Input images



Vincent Van Gogh



Michelangelo



Rembrandt



Johannes Vermeer



Pierre-Auguste Renoir



Leonardo da Vinci

Stable Diffusion

DreamBooth: Text-guided view synthesis

Input images



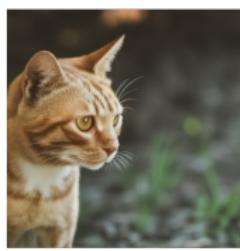
Top view ↑



Bottom view ↓



Side view →



Back view ↗



[V] cat seen from the top

[V] cat seen from the bottom

[V] cat seen from the side

[V] cat seen from the back

Stable Diffusion

DreamBooth: Property modification

Color modification ("A [color] [V] car")



Input



purple

red

yellow

blue

pink

Hybrids ("A cross of a [V] dog and a [target species]")



Input



Bear

Panda

Koala

Lion

Hippo

Stable Diffusion

DreamBooth: Accessorization

Input images



Stable Diffusion

Low-Rank Adaptation (LoRA)

- Efficient adaptation strategy to fine-tune large language models (Hu et al., 2021).
- Freezes the weight of the pre-trained model.
- Fine-tunes the cross attention layers.
- The trick of LoRA is breaking a matrix into two smaller (low-rank) matrices.
- Training time: 25 minutes \sim 1 hour.
- Model size: Megabytes.
- Collection of models: <https://civitai.com/>.

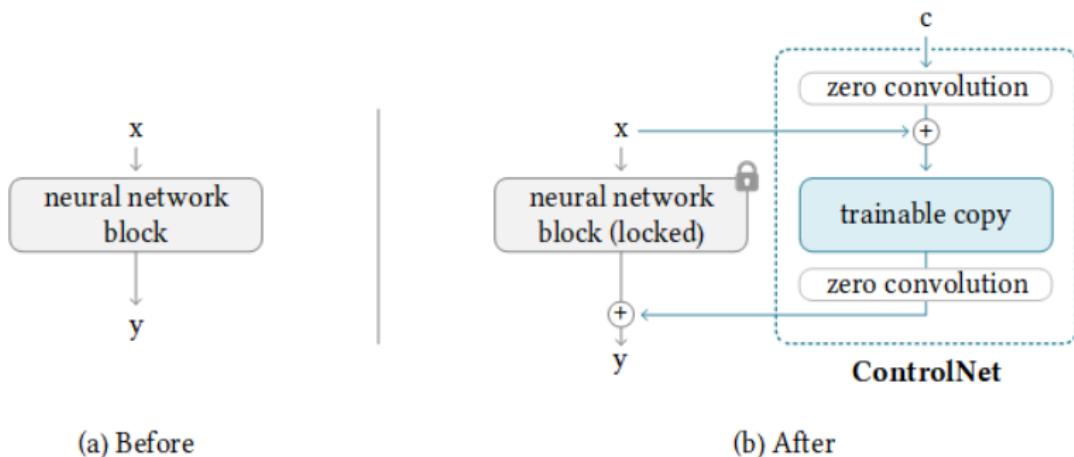
Stable Diffusion

ControlNet: Control human pose in Stable Diffusion

- ControlNet is a modified Stable Diffusion model (Zhang et al., 2023).
- It takes an additional input image and detects its outlines.
- This information is fed into the model as an additional conditioning.

Stable Diffusion

ControlNet: Architecture

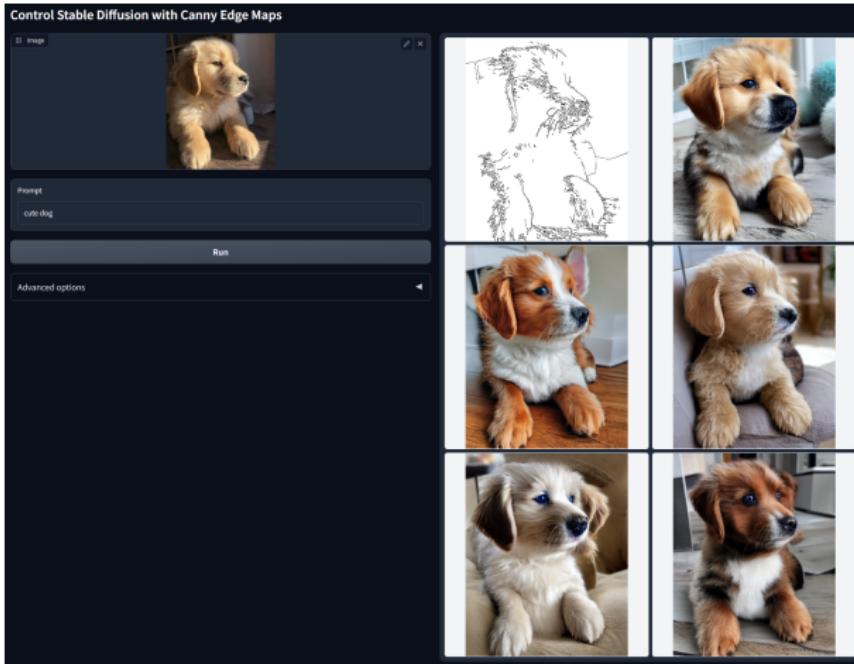


(a) Before

(b) After

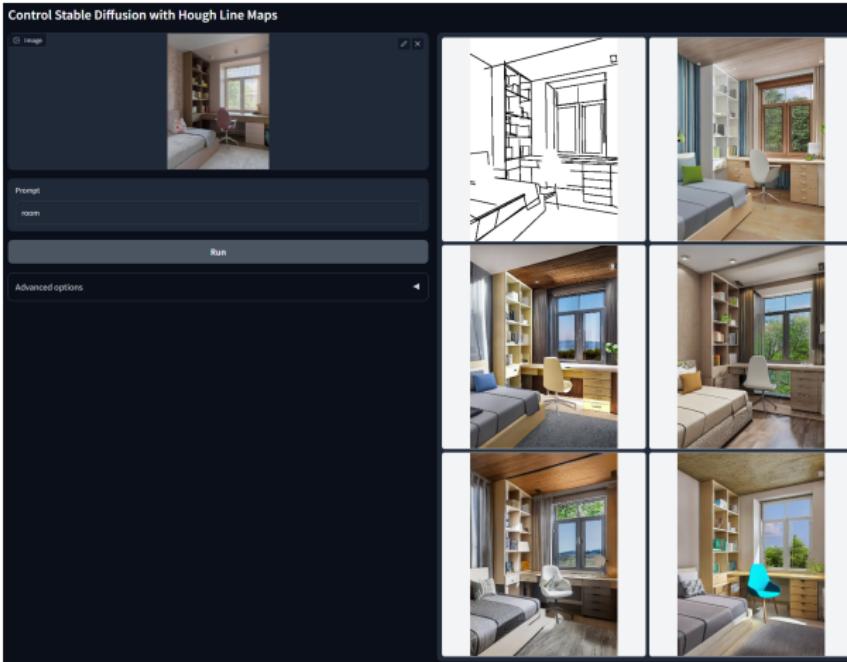
Stable Diffusion

ControlNet: Canny edge detection



Stable Diffusion

ControlNet: Hough line maps



Stable Diffusion

ControlNet: HED maps

Control Stable Diffusion with HED Maps

Image:

Prompt: oil painting of handsome old man, masterpiece

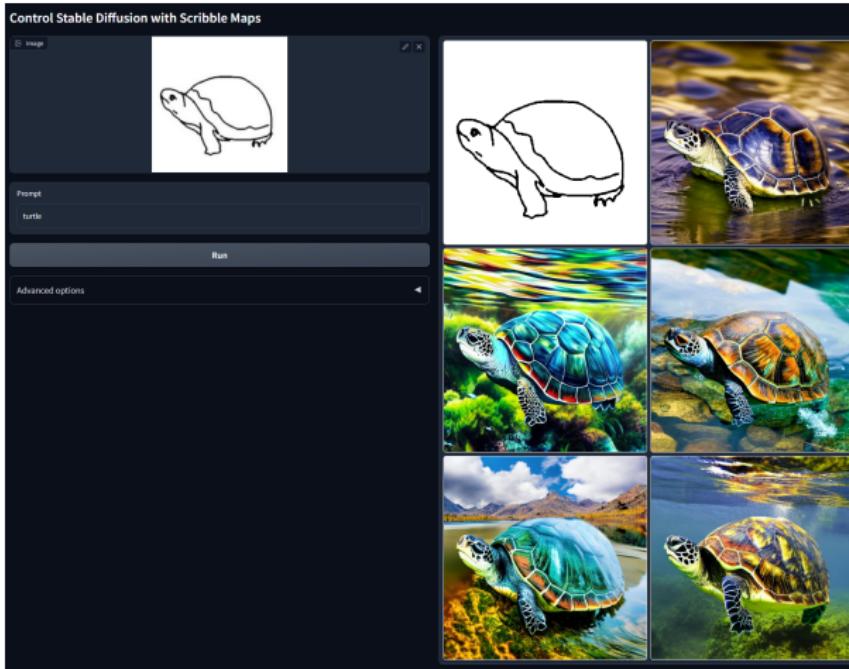
Ran

Advanced options:

The interface shows a 3x3 grid of generated images. The first column contains the input image and a black and white HED map. The second column contains two versions of the oil-painted result. The third column contains three versions of the oil-painted result, each with slightly different lighting and coloration.

Stable Diffusion

ControlNet: Scribble maps



Stable Diffusion

ControlNet: Sketches

Control Stable Diffusion with Interactive Scribbles

Canvas Width: 512

Canvas Height: 512

Open drawing canvas

(Image)

Do not forget to change your brush width to make it thinner. [Gradio do not allow developers to set brush width so you need to do it manually.] Just click on the small pencil icon in the upper right corner of the above block.

Prompt: dog in a room

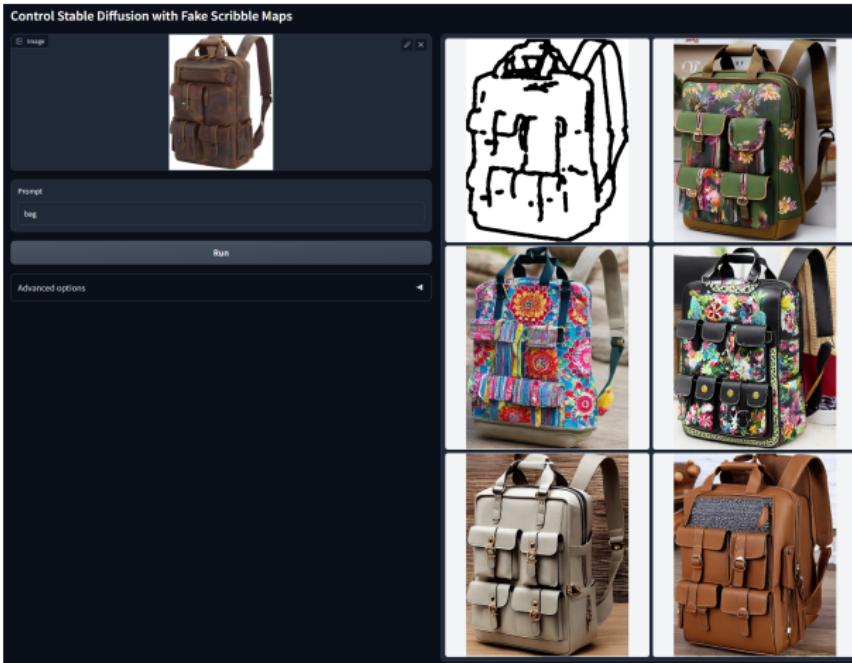
Run

Advanced options

The interface shows a drawing canvas where a user has drawn a simple black outline of a dog's head and shoulders. To the right of the canvas are five generated images of dogs, each showing a different variation of the dog's face based on the scribble. The generated images include a white puppy, a brown and white puppy, a brown and white puppy sitting, a small brown and white dog lying down, and a small white dog with brown ears lying down.

Stable Diffusion

ControlNet: Fake scribbles



Stable Diffusion

ControlNet: Human poses

Control Stable Diffusion with Human Pose

Image:

Prompt: Chef in the kitchen

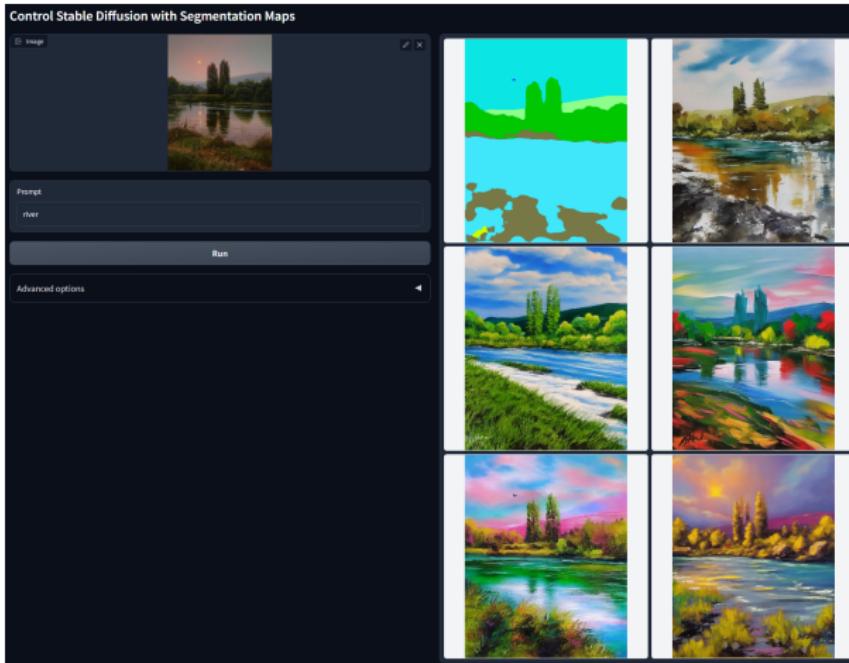
Run

Advanced options

The interface allows users to upload an image, provide a text prompt, and use a control sketch to generate images where the subject's pose matches the sketch. The generated images are displayed in a 3x3 grid.

Stable Diffusion

ControlNet: Segmentation maps



Stable Diffusion

Prompt Generation

Anatomy of a good prompt

- **Subject:** what you want to see in the image.
- **Medium:** the material used to make artwork.
- **Style:** artistic style of the image.
- **Artist:** to use a particular artist as a reference.
- **Website:** graphic websites such as *Artstation* and *Deviant Art*.
- **Resolution:** how sharp and detailed the image is.
- **Additional details:** such as sci-fi, stunningly beautiful, etc.
- **Color:** color keywords to control the overall color of the image.
- **Lighting:** lighting keywords can have a huge effect on how the image looks.

Examples: <https://stable-diffusion-art.com/prompt-guide/>.

Outline

1. Introduction

2. Image generation

3. Video generation

4. Controversy

Deforum

- Deforum is a tool to create animation videos with Stable Diffusion.
- Example and tutorial:
<https://stable-diffusion-art.com/deforum/>.

Lumiere

- Developer: Google.
- Features: Generates short videos from a prompt.
- Examples: <https://lumiere-video.github.io/>.

Sora

The world simulator

- Developer: OpenAI.
- Features:
 - ▶ Prompting with images and videos.
 - ▶ Extending generated videos.
 - ▶ Video-to-video editing.
 - ▶ Connecting videos.
 - ▶ Simulating digital worlds.
- Examples: [https://openai.com/research/
video-generation-models-as-world-simulators](https://openai.com/research/video-generation-models-as-world-simulators).

Outline

1. Introduction

2. Image generation

3. Video generation

4. Controversy

Controversy

- Fake news.
- Most models are trained with images scraped from the web, without paying attention to copyright.
- Some developers are creating the models for lucrative purposes.
- Some users are using the generated images for commercial and lucrative purposes.
- This is specially cumbersome for artists, whose personal styles are being “replicated” by the models.
- Overall, this is a delicate matter that needs to be addressed and legislated carefully.

Bibliography

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). **Language models are few-shot learners.** In *Advances in neural information processing systems*, 33, 1877–1901.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). **Lora: Low-rank adaptation of large language models.** *arXiv preprint arXiv:2106.09685*.
- Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). **Dreamfusion: Text-to-3d using 2d diffusion.** *arXiv preprint arXiv:2209.14988*.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). **Hierarchical text-conditional image generation with clip latents.** *arXiv preprint arXiv:2204.06125*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). **High-resolution image synthesis with latent diffusion models.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Bibliography

- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2022). **DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation.** *arXiv preprint arxiv:2208.12242*.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... & Jitsev, J. (2022). **Laion-5b: An open large-scale dataset for training next generation image-text models.** *arXiv preprint arXiv:2210.08402*.
- Stable Diffusion Art. <https://stable-diffusion-art.com/>.
- Zhang, L., & Agrawala, M. (2023). **Adding conditional control to text-to-image diffusion models.** *arXiv preprint arXiv:2302.05543*.