ControlNet: Adding conditional control to text-to-image diffusion models

Zhang, L., Rao, A. & Agrawala, M.

Aprendizaje profundo para visión artificial. 2023

Juan Manuel Varela

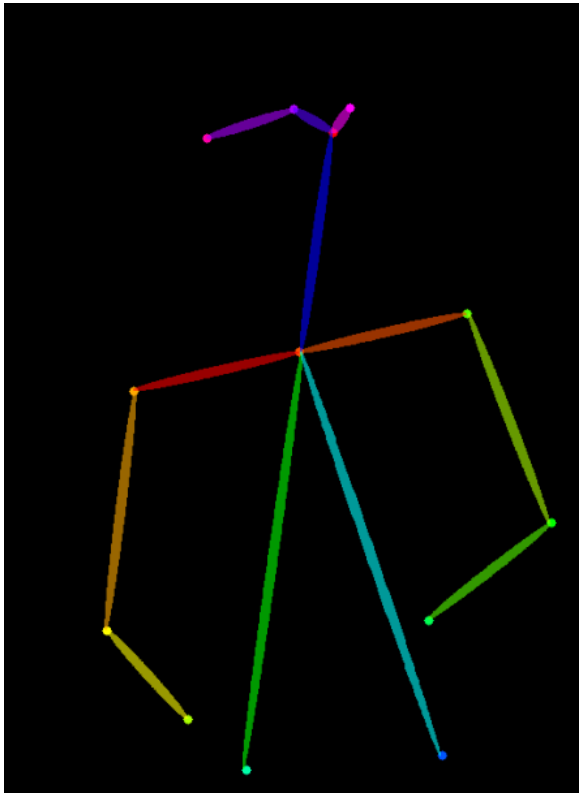
Agenda

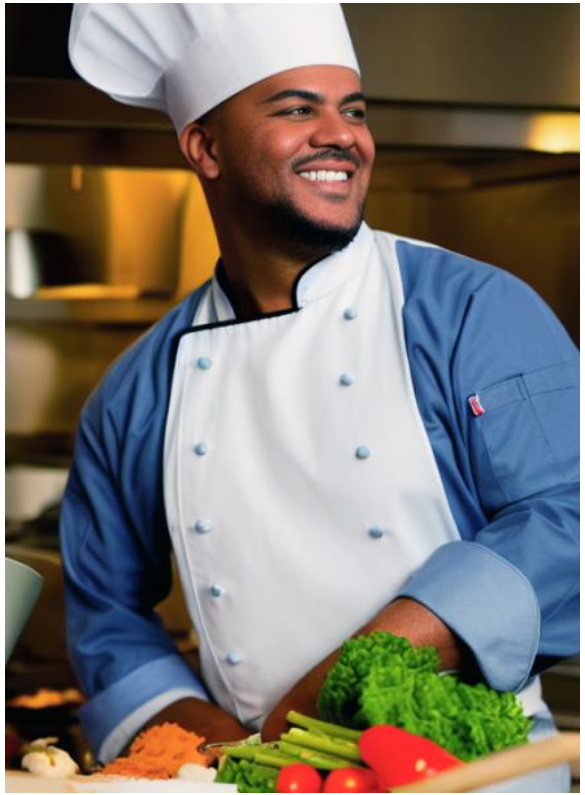
- Introducción
- Modelos de difusión
- Stable Diffusion
- ControlNet
- Entrenamiento
- Resultados
- Conclusiones

Introducción

Los modelos de difusión de texto a imágen permiten generar imágenes de alta calidad a partir de prompts, sin embargo se tiene poco control de la composición espacial de la imagen

Introducción



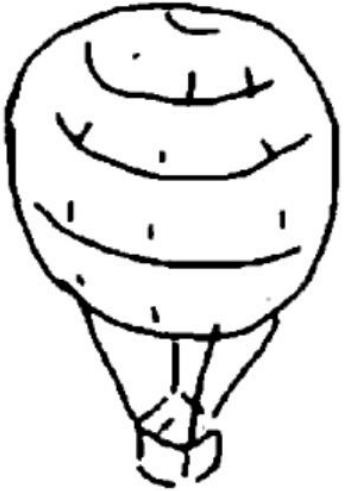


Texto: “Chef en la cocina”

Introducción

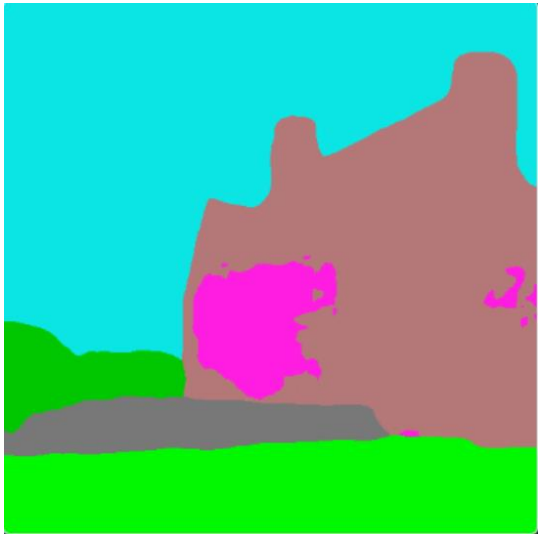
Bordes de Canny

Lineas de Hough



Dibujos

Introducción

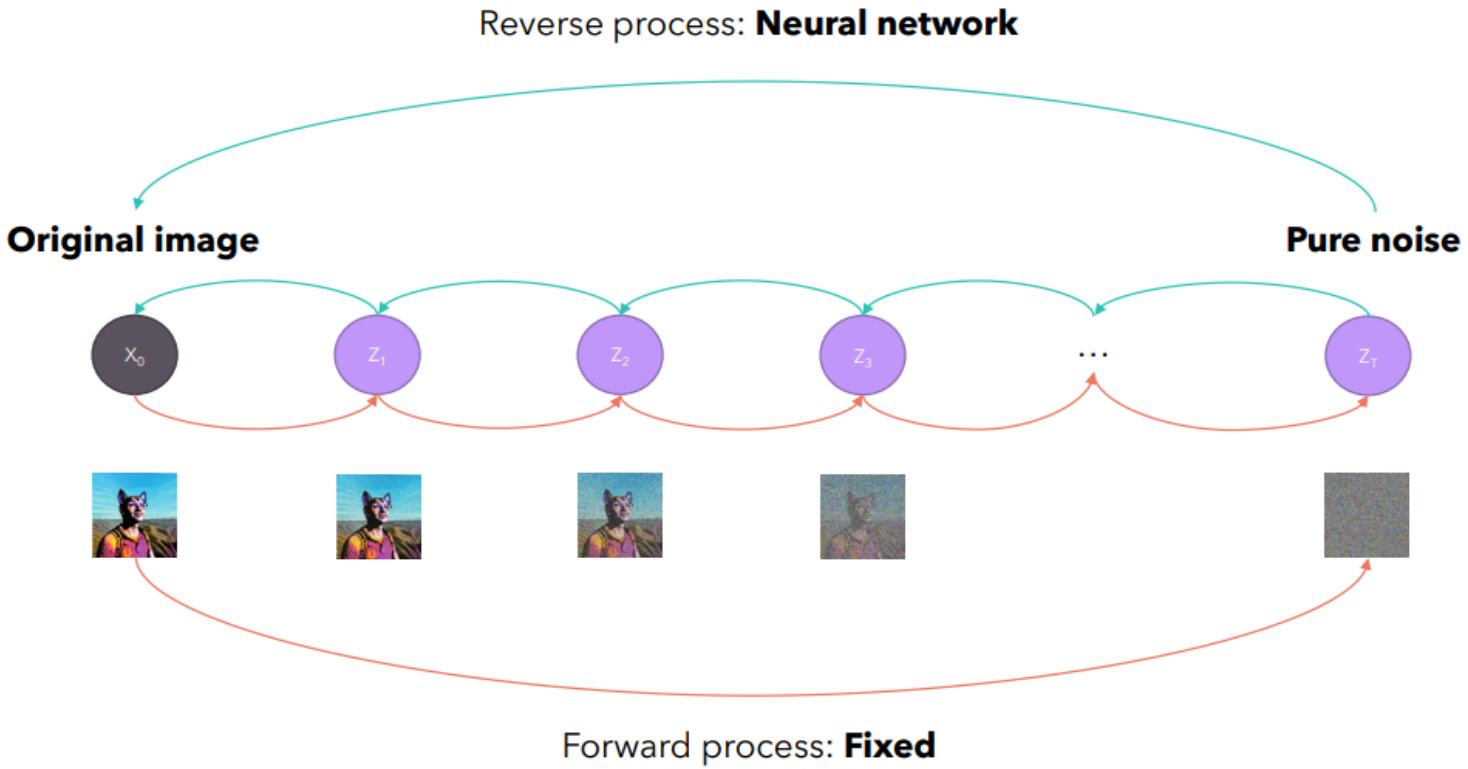


Mapas de segmentación semántica

Mapas de profundidad

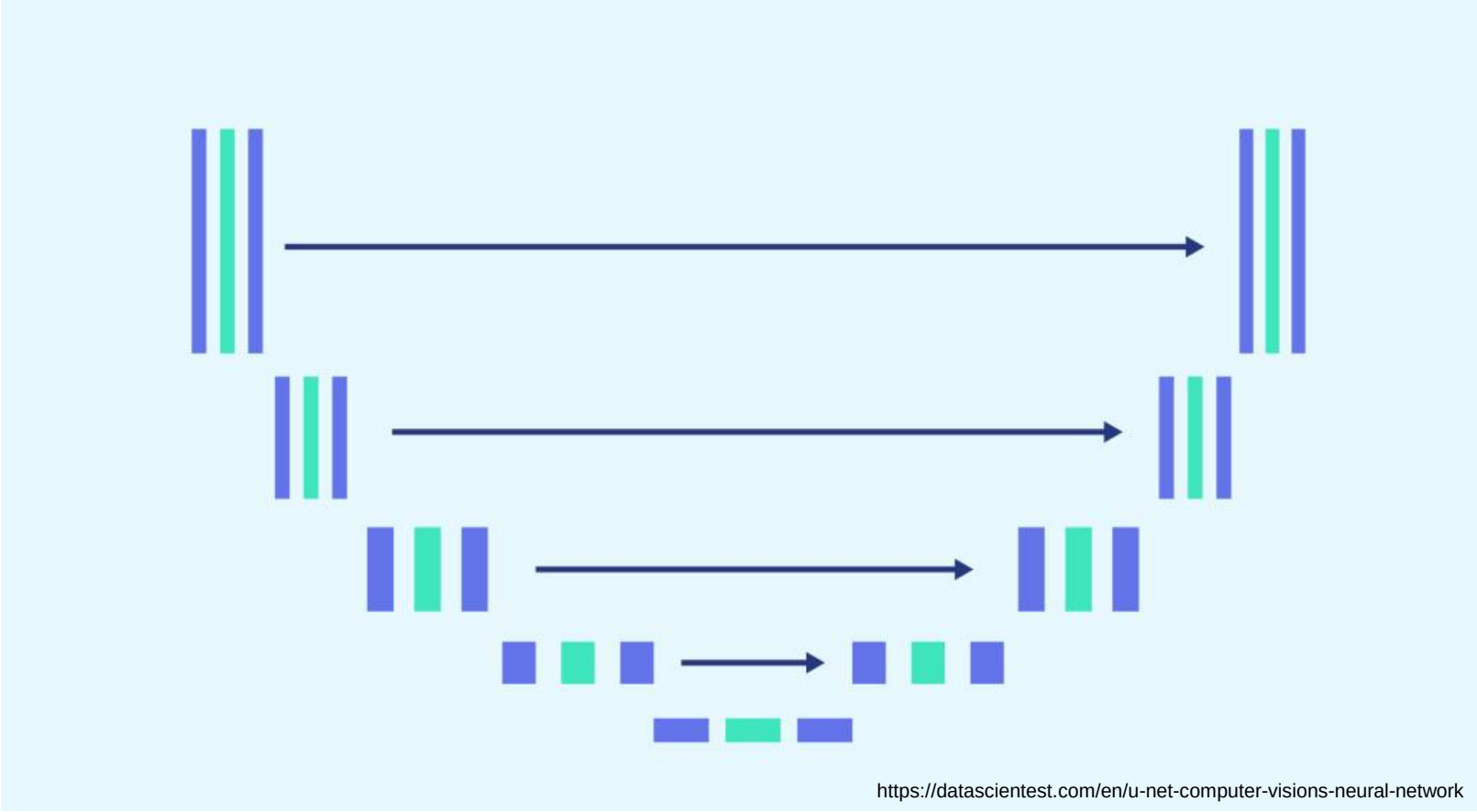
Mapas de normal

Modelos de difusión



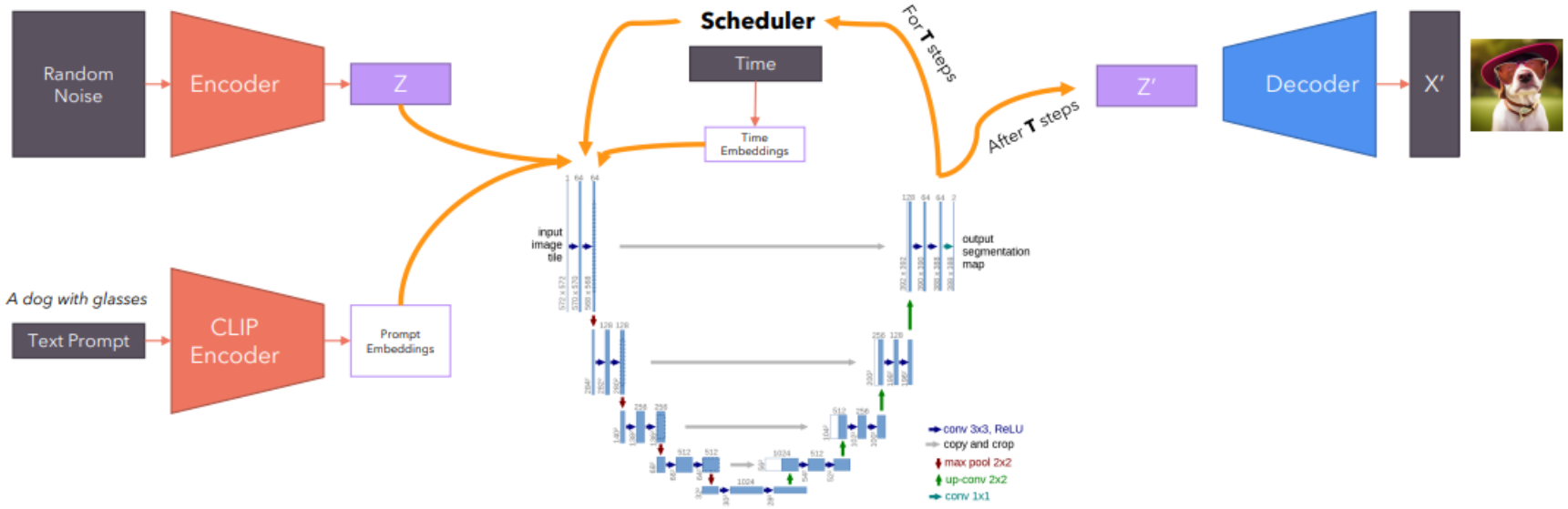
Umar Jamil - https://github.com/hkproj/pytorch-stable-diffusion

Modelos de difusión



https://datascientest.com/en/u-net-computer-visions-neural-network

Stable Diffusion



Umar Jamil - https://github.com/hkproj/pytorch-stable-diffusion

Stable Diffusion

Inpainting

Modificación de imágenes

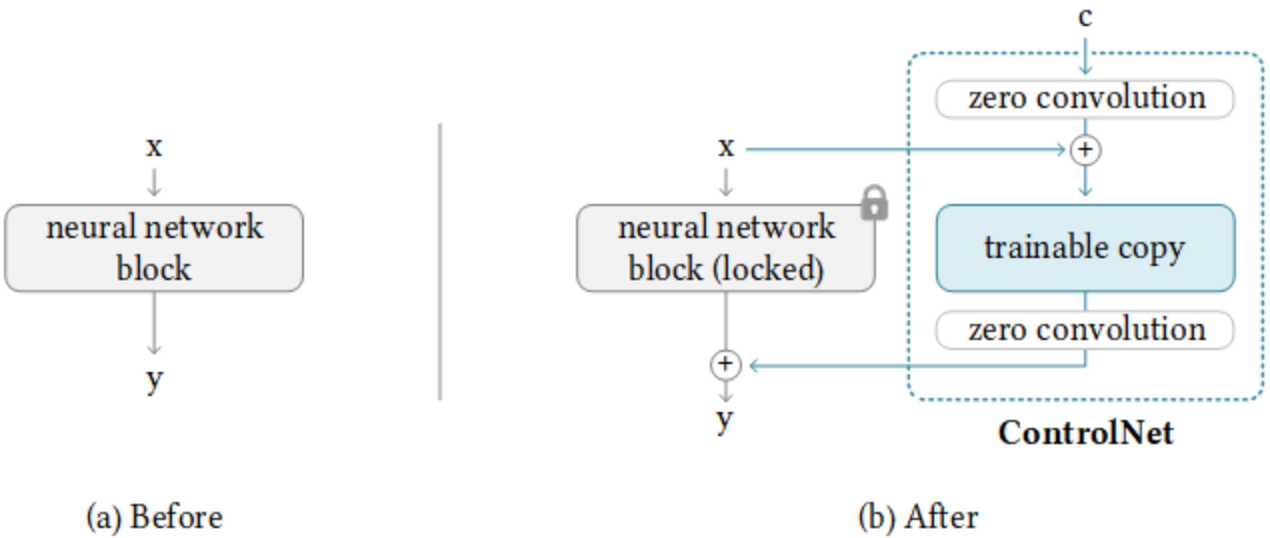
ControlNet

Desafíos:

- En general, se requieren soluciones end-to-end
- Conjuntos de datos de mucho menor tamaño que los originales.
- Hacer finetuning del modelo de difusión preentrenado, puede traer problemas

ControlNet

Solución:

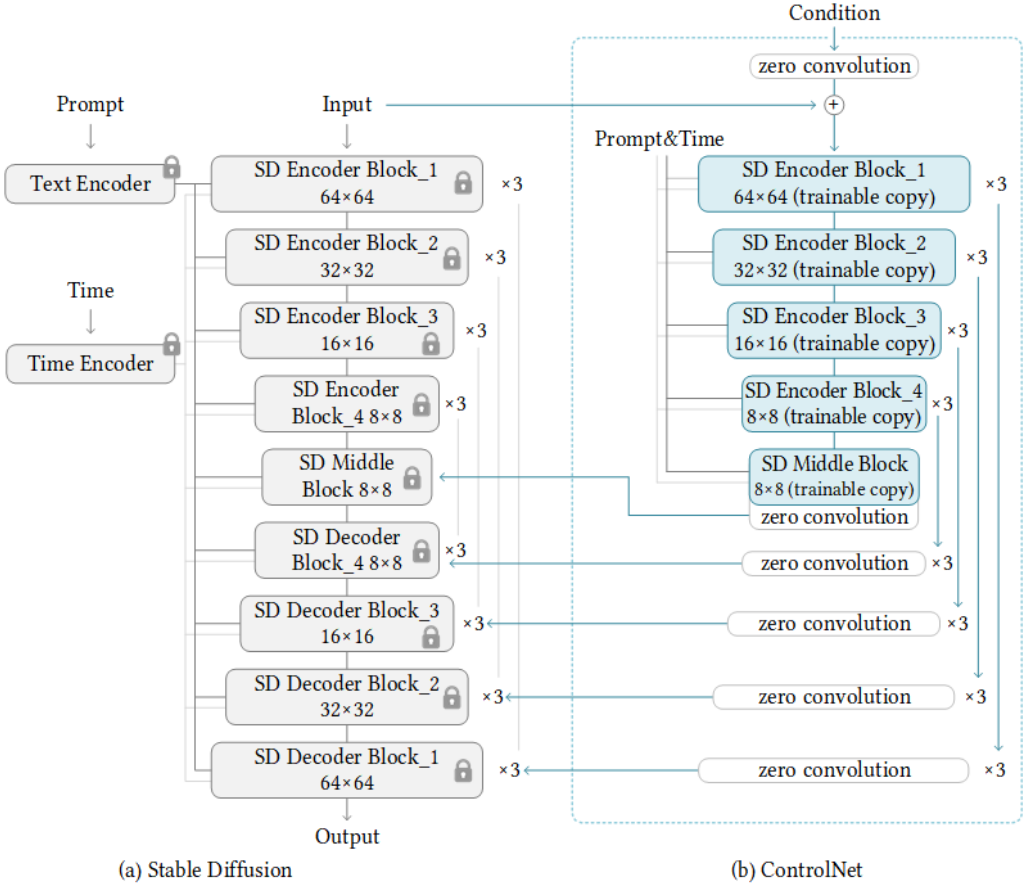


(a) Before

(b) After

ControlNet

Aplicada a Stable Diffusion:



Entrenamiento

Conjunto de datos:

“A cute dog”

Entrenamiento

- Tomar una instancia del conjunto de datos
- Muestrear un tiempo aleatorio
- Muestrear ruido
- Tomar paso de descenso por gradiente en base a la siguiente función de costo.

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}) \|_2^2 \right]$$

- Repetir hasta converger

ControlNet

Particularidades:

Input

“a high-quality and extremely detailed image”

ControlNet

Particularidades:

Test input | training step 100 | step 1000 | step 2000

step 6100 | **step 6133** | step 8000 | step 12000

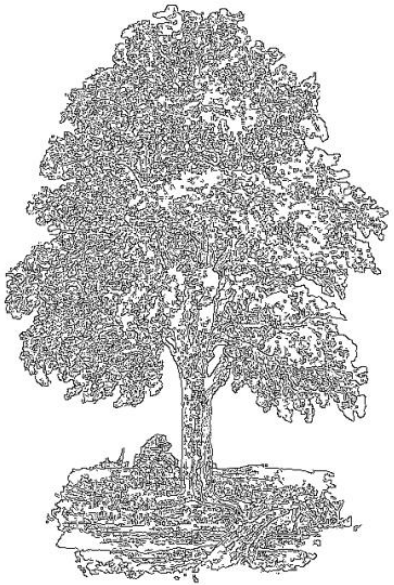
ControlNet

Particularidades:

"Lion" 1k images 50k images 3m images

Resultados

Bordes de canny



Palette

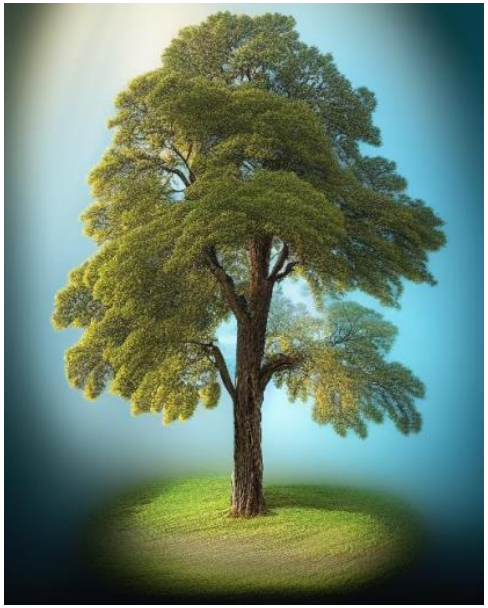
Taming Transformer

Resultados

LDM

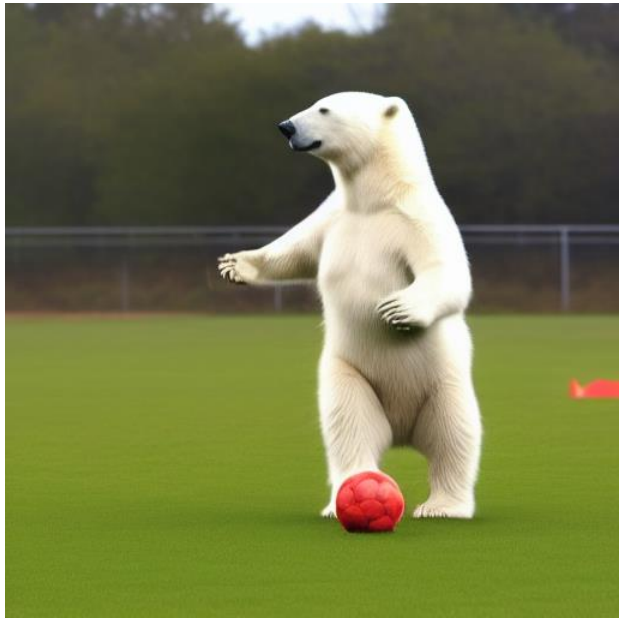
PITI

ControlNet+SD



Resultados

"Un oso polar jugando al futbol"



SD

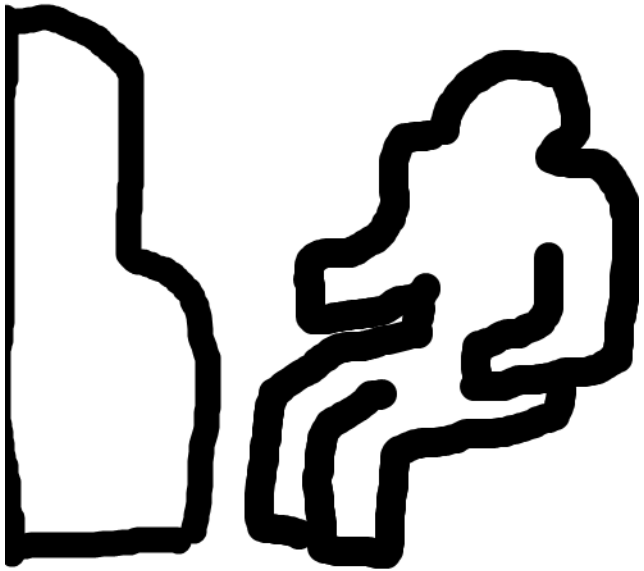
Dibujo

SD+ControlNet

Resultados

"Un astronauta tocando el piano"

SD



Dibujo

SD+ControlNet

Conclusiones

- Los autores lograron agregar con éxito un elemento extra de control a un gran modelo de difusión.
- Se podría implementar con otros modelos.
- Continuar explorando arquitecturas más simples.
- Disminuir tiempo de generación.
- La comparación con los otros modelos tiene matices.

¡Gracias!