Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Panteleev Daniil

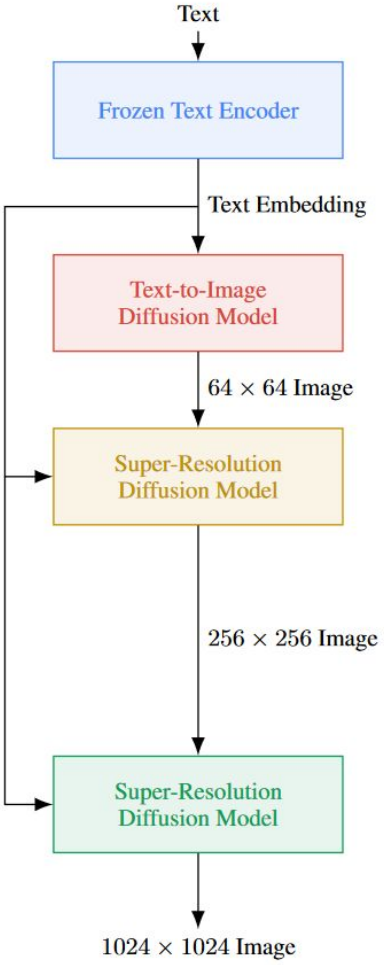
Imagen: photorealistic text-to-image diffusion model

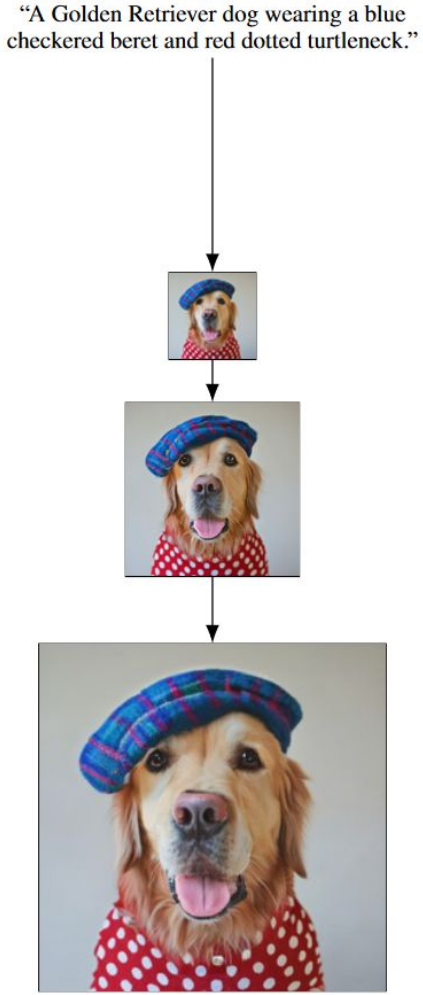
Sprouts in the shape of text ‘Imagen’ coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

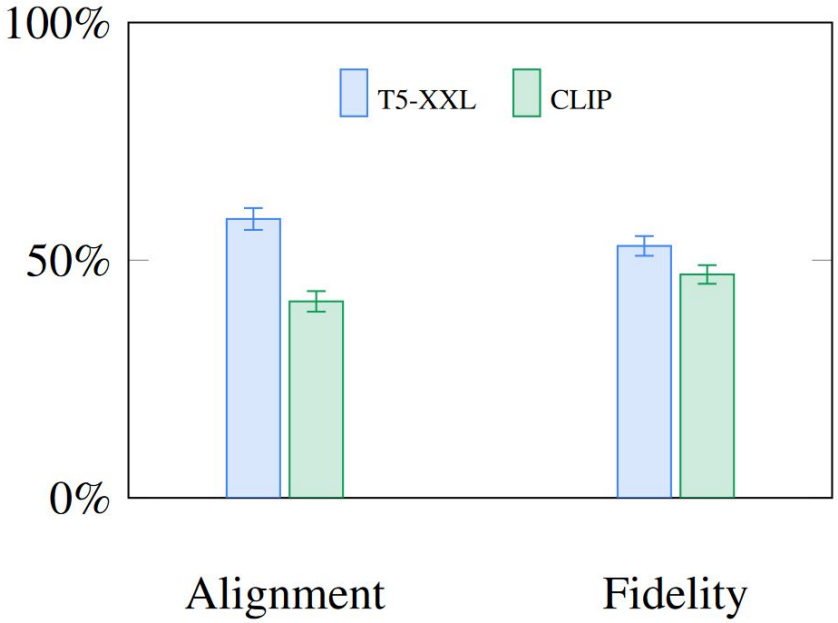
Imagen

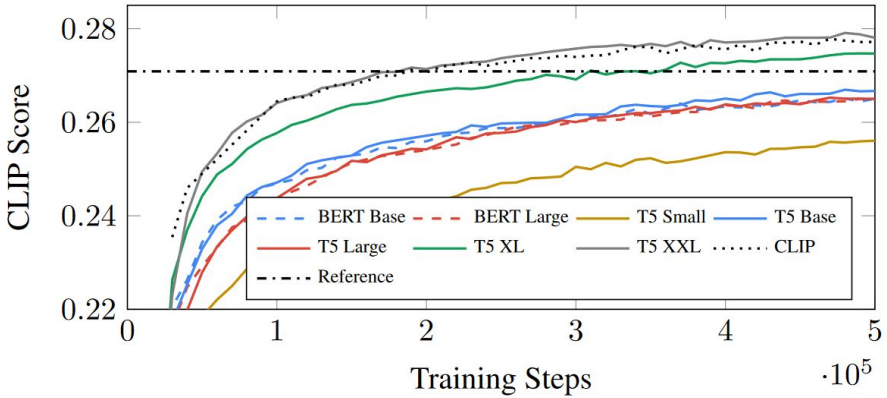

Text
Frozen Text Encoder
Text Embedding
Text-to-Image Diffusion Model
64 × 64 Image
Super-Resolution Diffusion Model
256 × 256 Image
Super-Resolution Diffusion Model
1024 × 1024 Image


"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck."

Text encoder

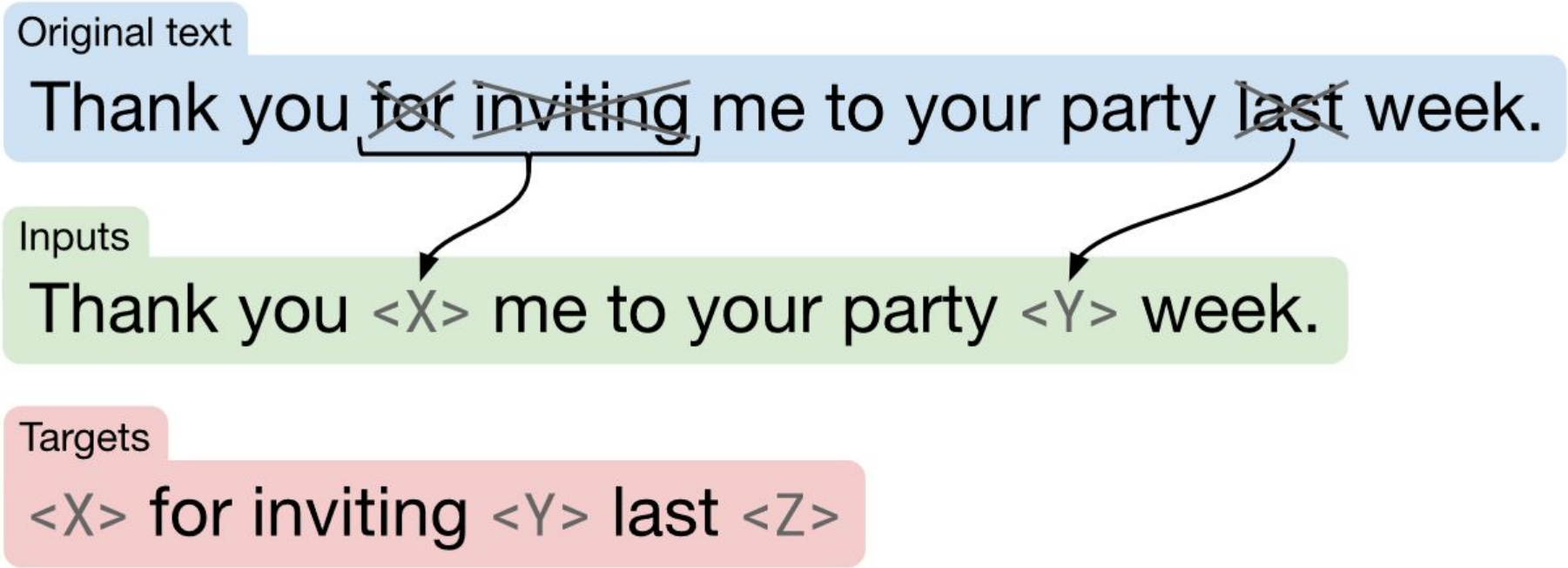
T5 vs CLIP vs BERT (frozen weights)





T5-XXL (11B params)

Trained on 20B Common Crawl excerptions - “C4”



T5-XXL (11B params)

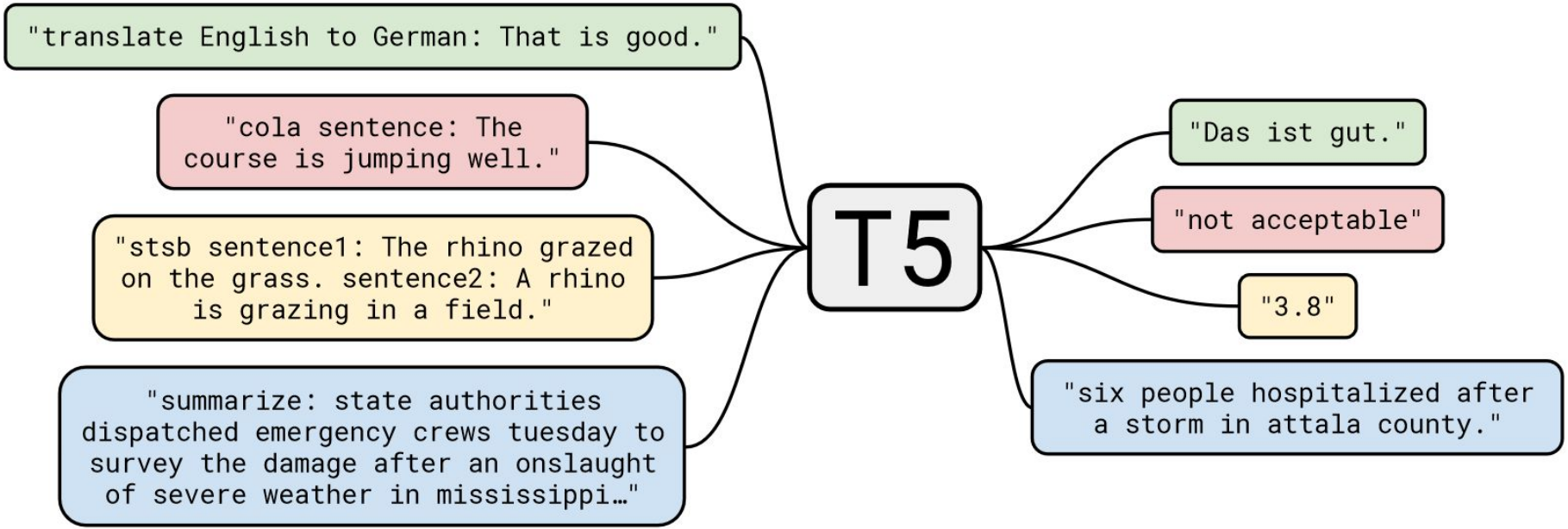


Imagen uses T5-XXL's **Frozen Encoder**

Diffusion models: classifier-free guidance

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}) = w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) + (1-w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$$

Train conditional and unconditional models together,

dropping class label with a certain chance

c - class label, z_t - generated points,

w - chance of dropping c (may be greater than 1)

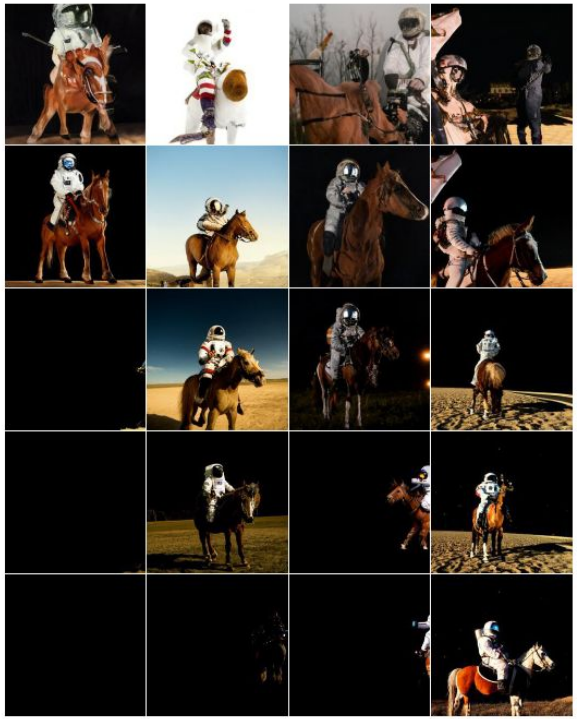
if $c = 1$ classifier-free guidance is turned off

Diffusion models: dynamic thresholding

Static thresholding: clip x to [-1, 1]
Result: images are still overly saturated

Dynamic thresholding: clip x to [-s, s], where s = some percentile of image absolute pixel value,
then x is divided by s
Result: success

Diffusion models: dynamic thresholding

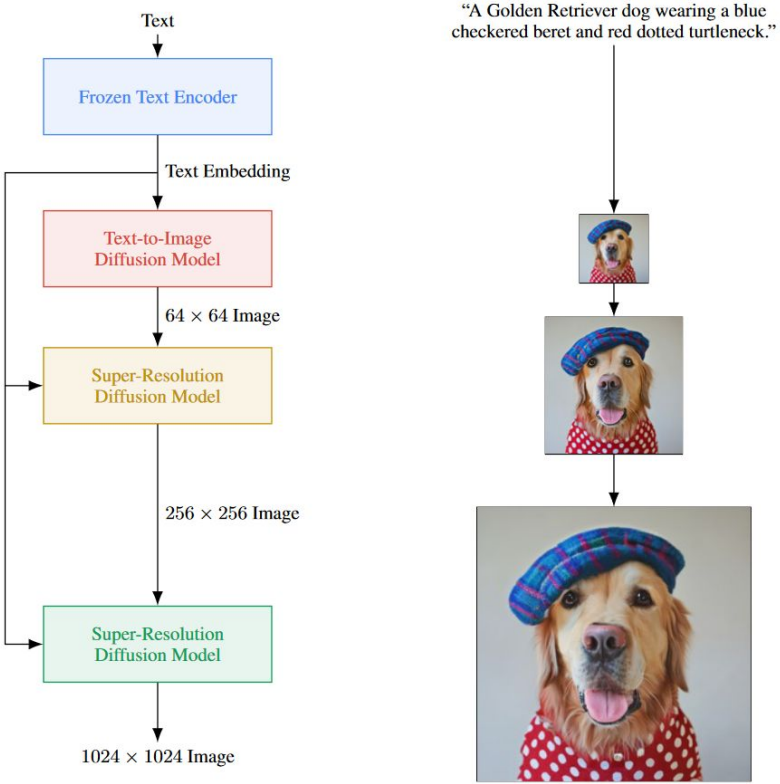


(a) No thresholding.

(b) Static thresholding.

(c) Dynamic thresholding.

Diffusion models: architecture



Diffusion models: architecture (sequence to image)

U-Net (paper: "Denoising Diffusion Probabilistic models")

Input: Frozen Encoder Embeddings

Output: 64x64 image

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

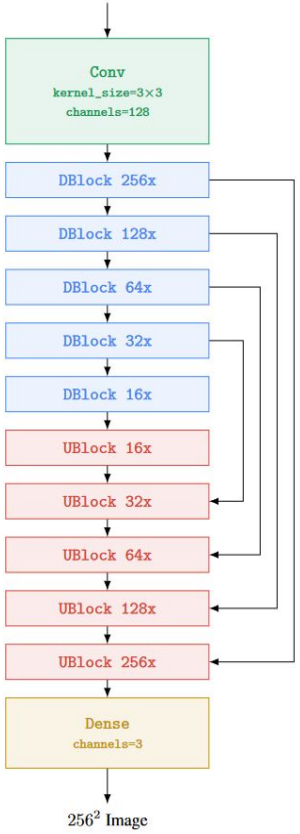
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Diffusion models: architecture (Super-Resolution)

Efficient U-Net

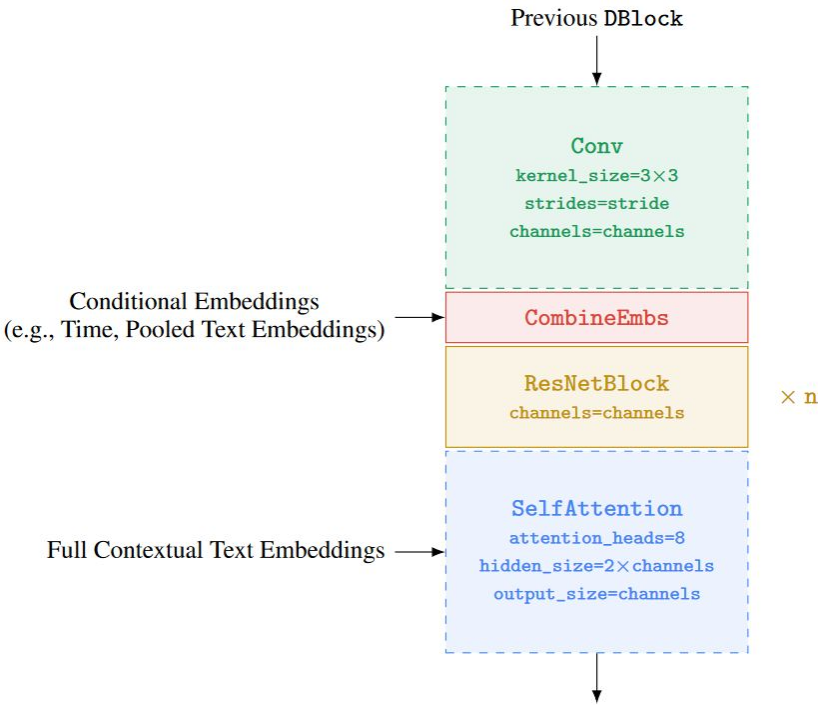
Input: Low-Res image

Output: 4xRes image

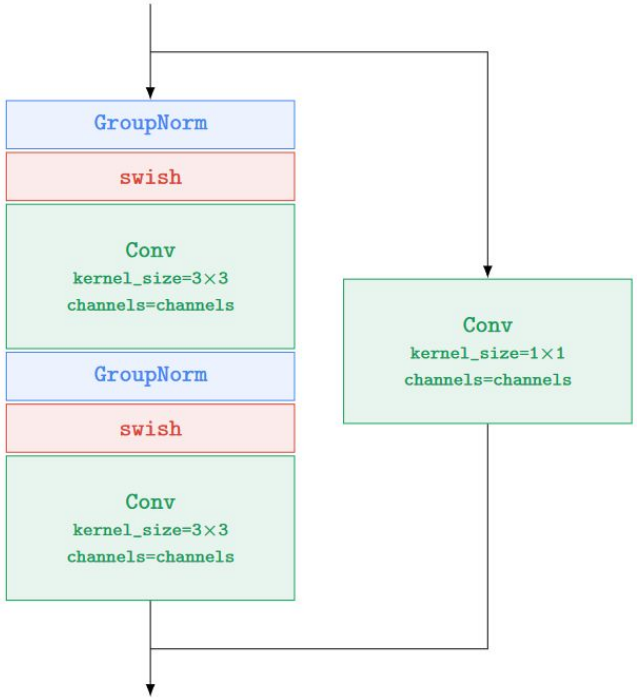


Diffusion models: architecture (Super-Resolution)

DBlock

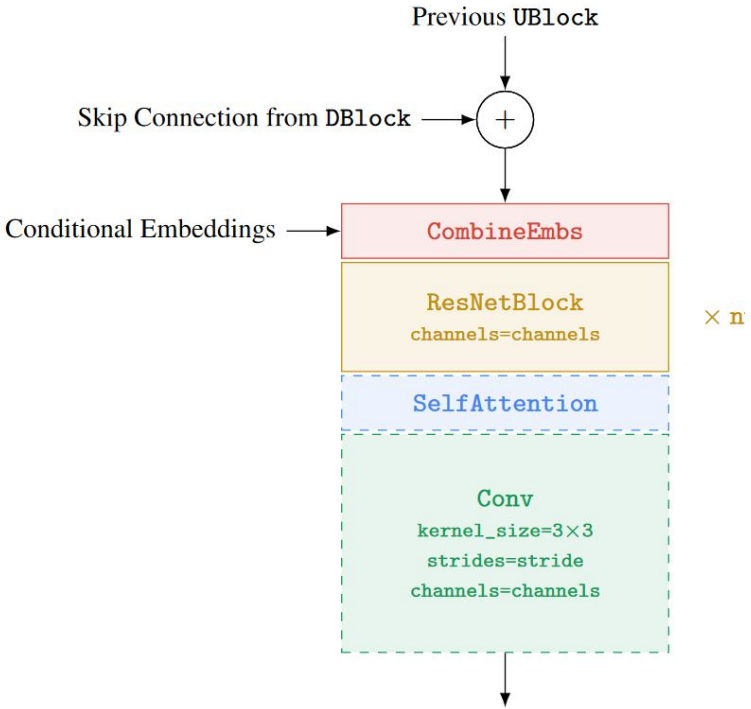


ResNetBlock



Diffusion models: architecture (Super-Resolution)

UBlock



ResNetBlock

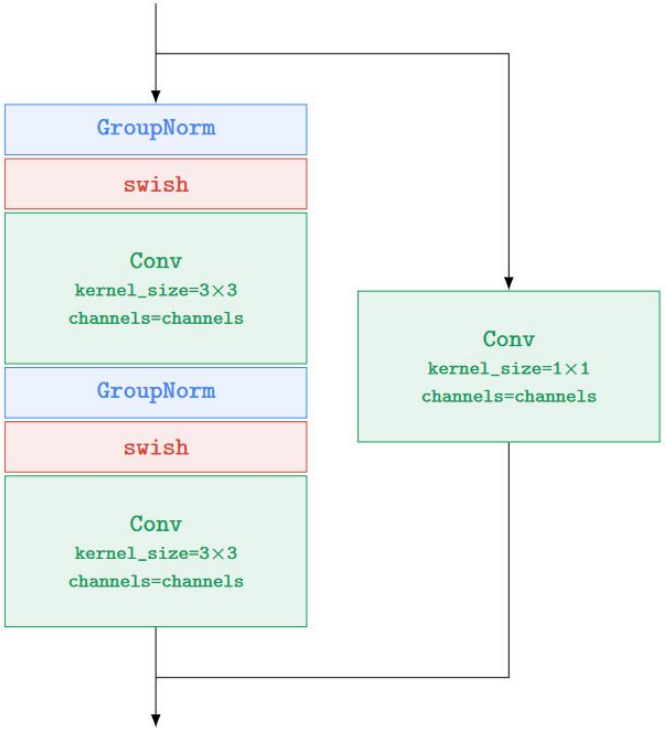
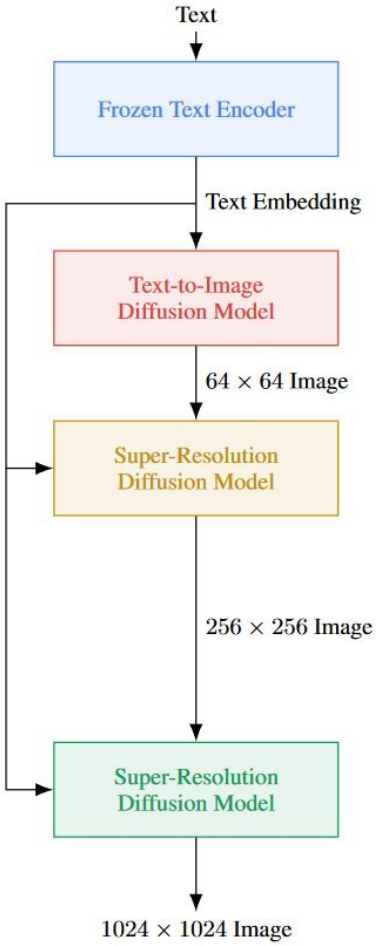


Imagen: recap



T5-XXL encoder

U-Net x64

Efficient U-Net x256

Efficient U-Net x1024 (no self-attention, cross-attention instead)

Evaluation

Model evaluation: FID score (fidelity), CLIP score (image-text alignment)

Experimental evaluation:

A brown bird and a blue bear.

One cat and two dogs sitting on the grass.

A sign that says 'NeurIPS'.

A small blue book sitting on a large red book.

A blue coloured pizza.

A wine glass on top of a dog.

1. Preference rate (how many times generated image has been chosen)
2. Alignment (does the text represent what's on the image)

CoCo dataset (25GB)

Analysis of Imagen

1. Scaling text encoder size is extremely effective
2. Scaling text encoder size is more important than U-Net size
3. Dynamic thresholding is critical
4. Human raters prefer T5-XXL over CLIP on DrawBench
5. Text conditioning method is critical
6. Efficient U-Net is critical

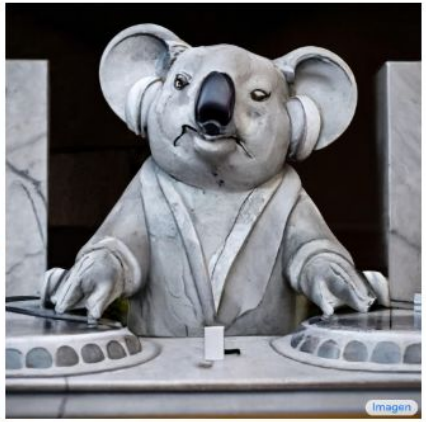
A relaxed garlic with a blindfold reading a newspaper while floating in a pool of tomato soup.

A photo of a corgi dog wearing a wizard hat playing guitar on the top of a mountain.

A single beam of light enter the room from the ceiling. The beam of light is illuminating an easel. On the easel there is a Rembrandt painting of a raccoon.

A squirrel is inside a giant bright shiny crystal ball in on the surface of blue ocean. There are few clouds in the sky.

A bald eagle made of chocolate powder, mango, and whipped cream.



A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala has wearing large marble headphones.