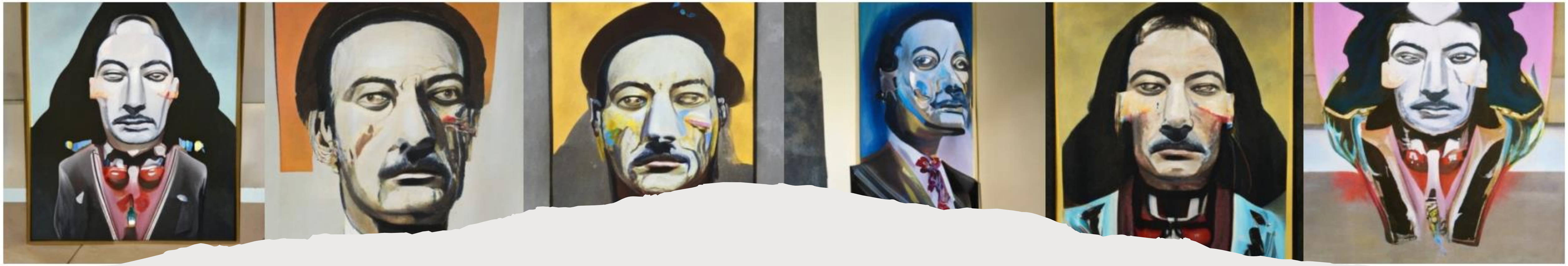
“Vibrant portrait painting of Salvador Dalí with a robotic half face.”

Ours (512px, 0.13s / img)

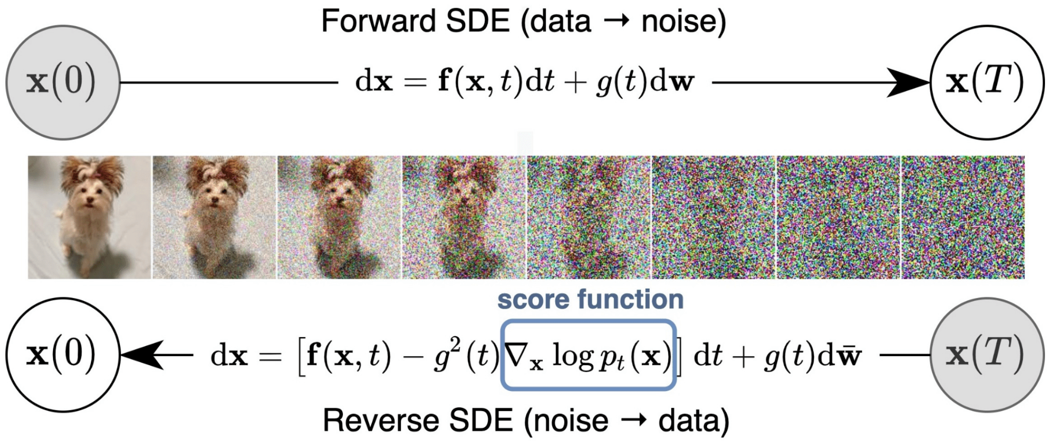


Scaling up GANs for Text-to-Image Synthesis (GigaGAN)

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, Taesung Park

GAN vs Diffussion Models

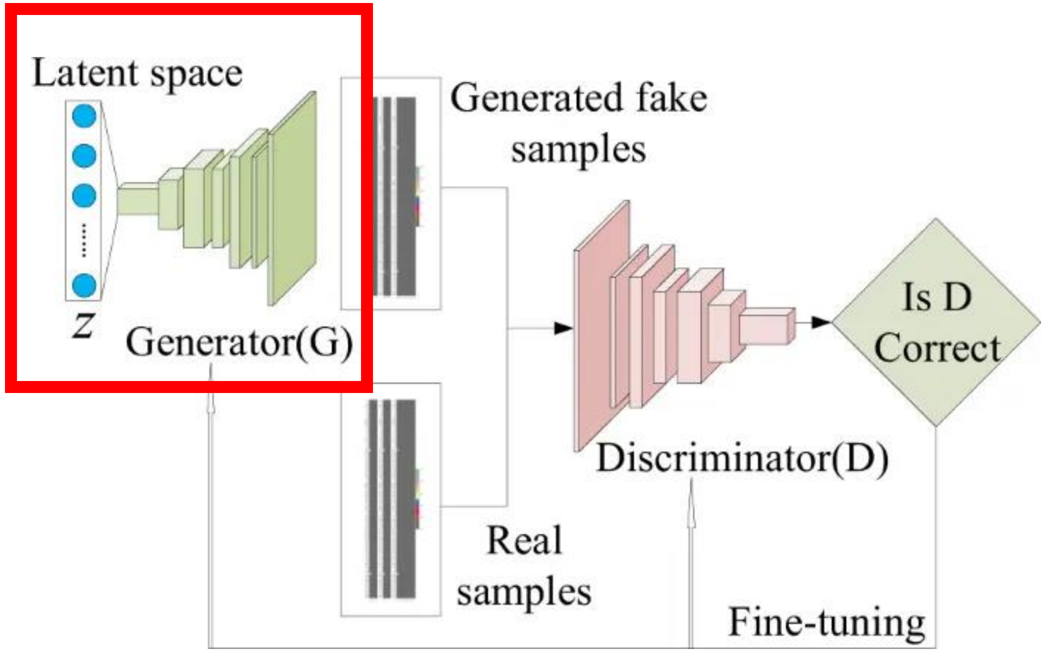
- Diffusion models are great.
- But... They rely on iterative inference.
  - Iterative methods enable stable training
  - And have a **high computational cost** during inference.



https://theaisummer.com/diffusion-models/

GAN vs Diffusion Models

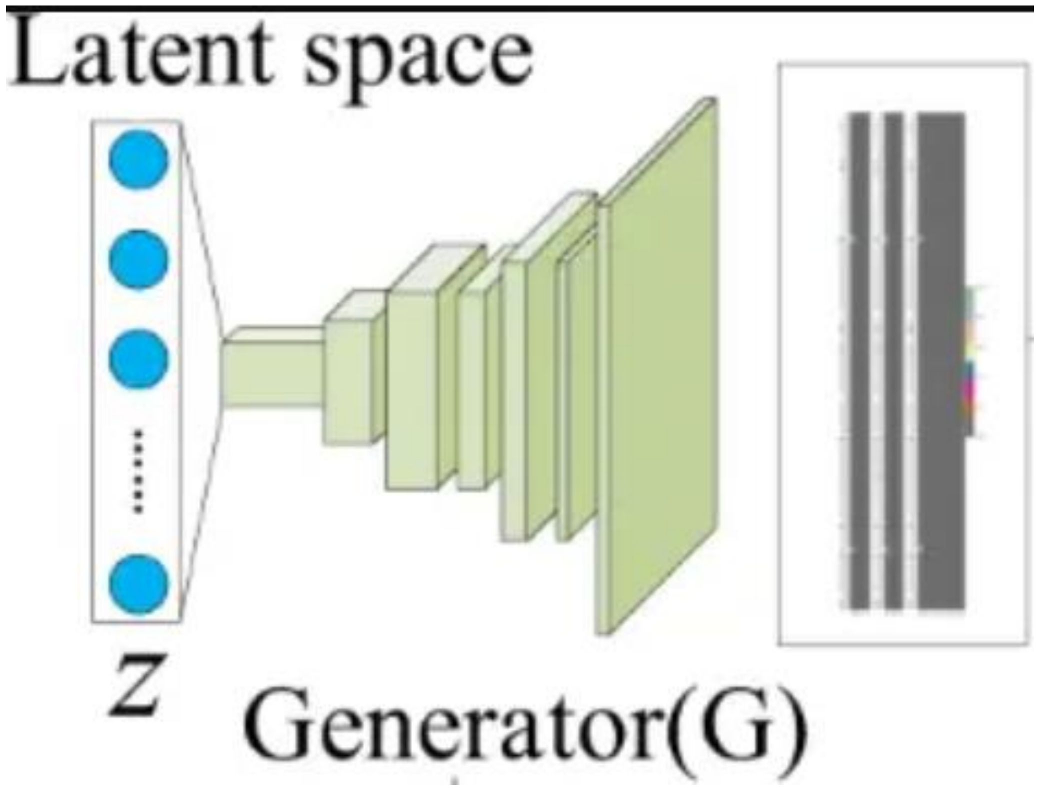
- GANs generate images through a single forward pass.
  - Not as visually accurate for diverse domains.
  - Great at modeling single (few) classes.



https://theaisummer.com/diffusion-models/

GigaGAN vs Diffusion Models

- Generating a 512px image in 0.13 seconds.
- It can synthesize ultra high-res images at **4k resolution** in 3.66 seconds.
- It still contains a "controllable", **latent vector space**.
    - Many of he know tricks to tailor image synthesis by modifying the latent space should work.



The "Giga" Part

- **Everything works better at scale.**

The "Giga" Part

- **Everything works better at scale.**
- Some of the larger GANs:
  - StyleGAN 70K-200K.
  - BigGAN 1M.
  - StyleGAN2 50M-100M.
- GigaGAN is **one billion parameter** GAN
- In context,1B parameter is still lower than:
  - Imagen (3.0B).
  - DALL-E 2 (5.5B).
  - Parti (20B).

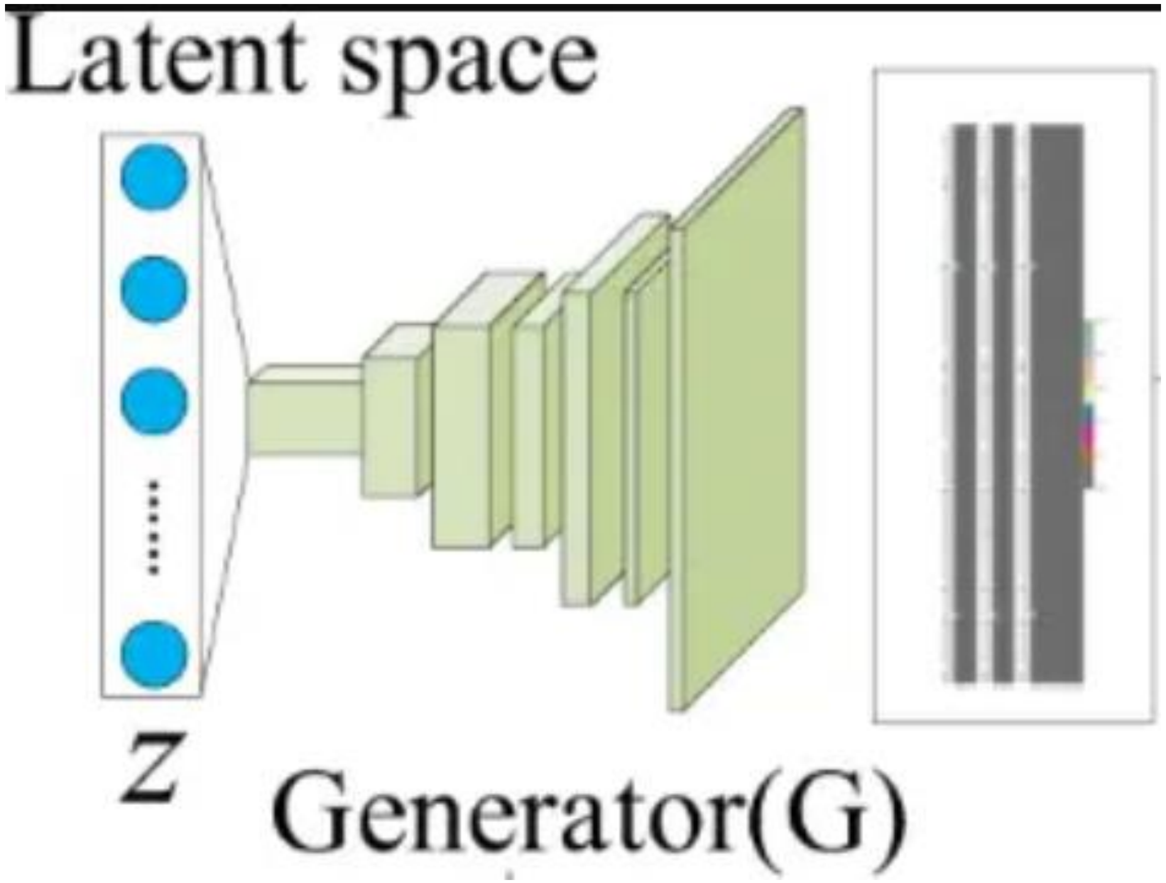
GigaGan - Training Data Size

- Trained on LAION5B-en.
  - **2.32 billion images.**
  - Each image is paired with text in the English language.
  - Also trained on ImageNet (upsampler).

Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

GigaGAN – Two Standalone Models

- GAN Synthesis Network up to 64x64.
  - First generates images at 64 × 64.
- Upsampler 512x512.



How does GigaGAN works?

Text Conditioning（CLIP）

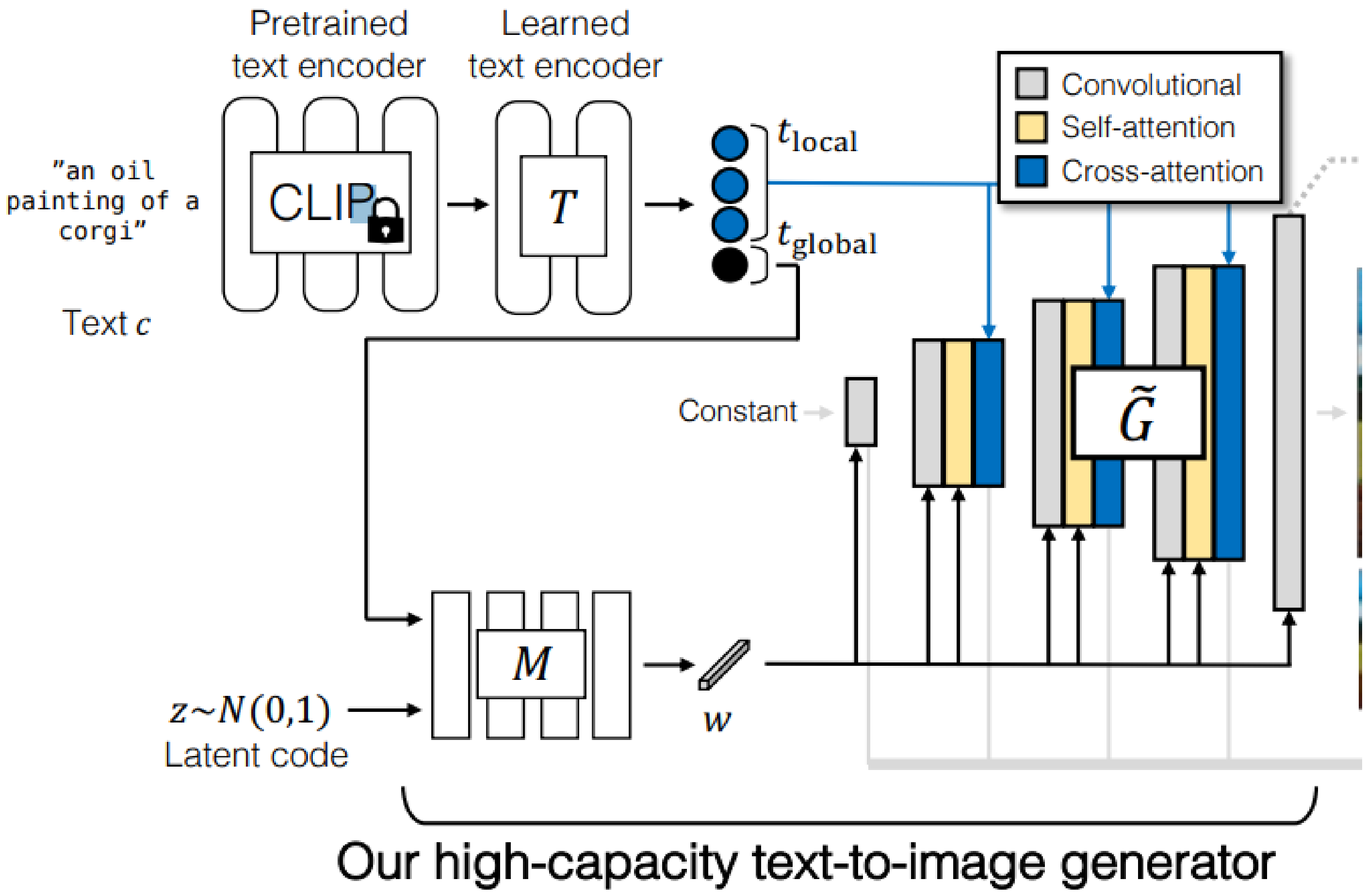
GAN **Synthesis**.

Text Based GAN Discriminator.

Upsampler Module.

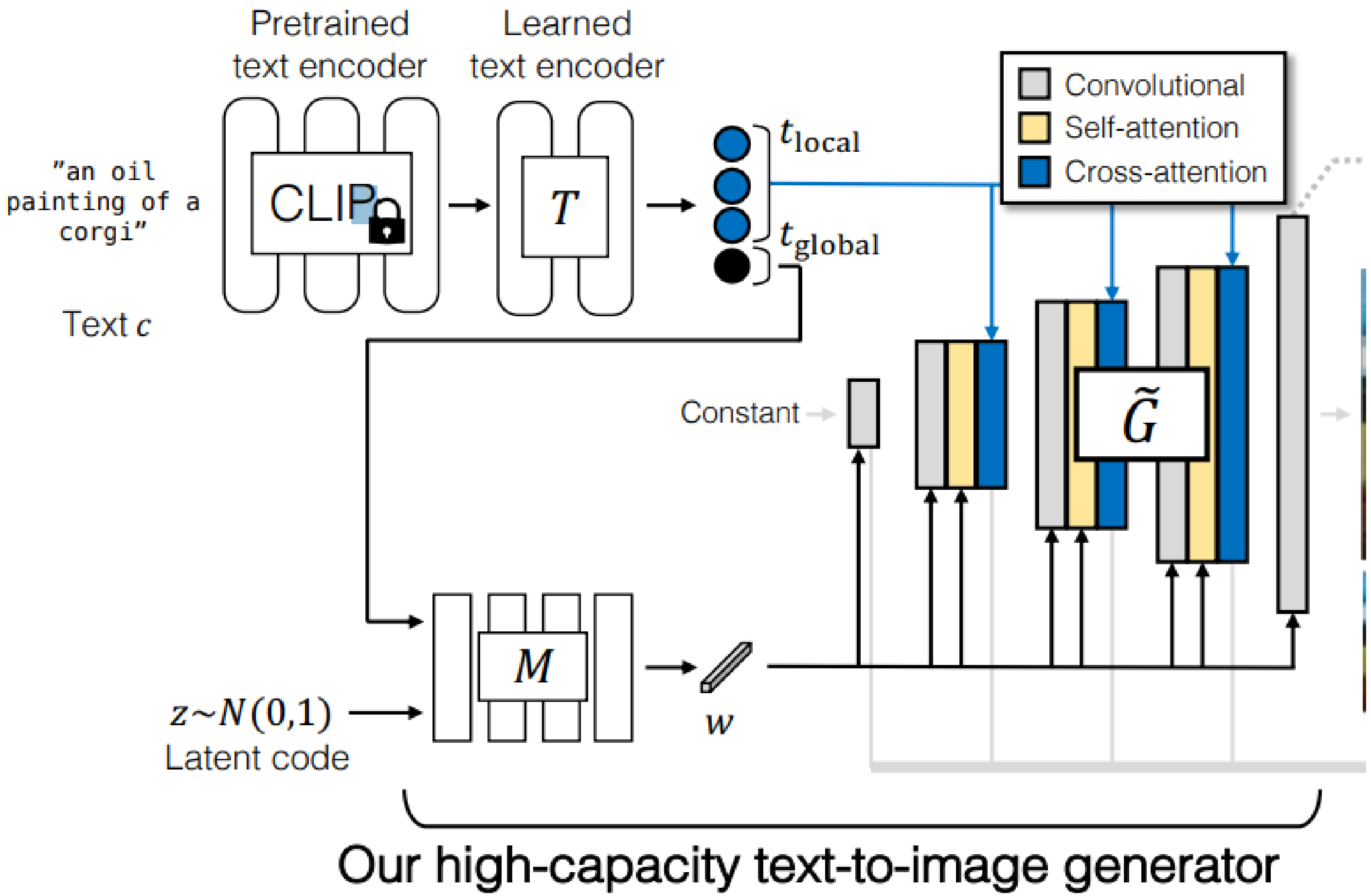
GigaGAN – Text Encoder

- Clip Text encoder
  - Original CLIP
  - And extra Transformer layers.

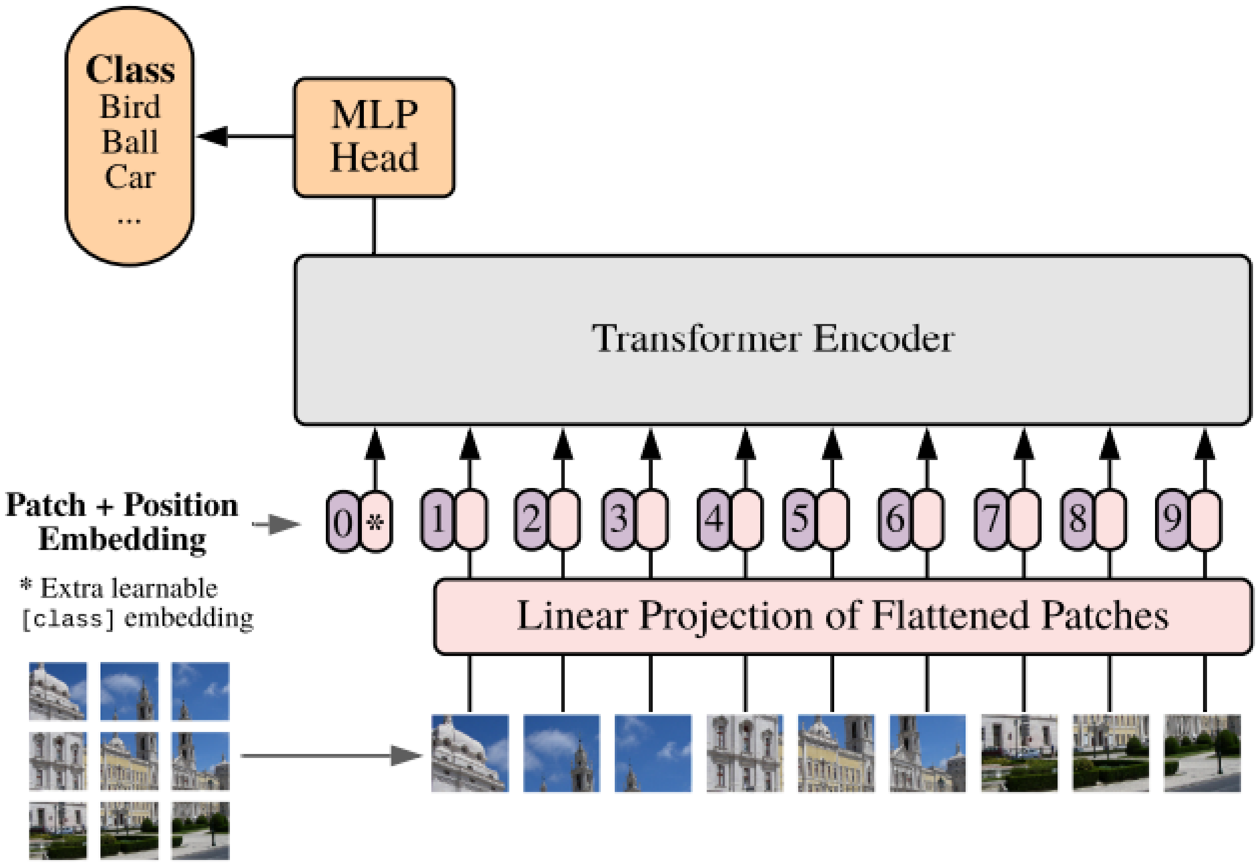


GigaGAN – Text Encoder

- Two outputs
  - Local Encoding
  - Global Encoding
- **Global encoding conditions L atent Vector.**
- **Local Encoder conditions the Synthesis network.**



Local and Global Features



Dosovitski et al.

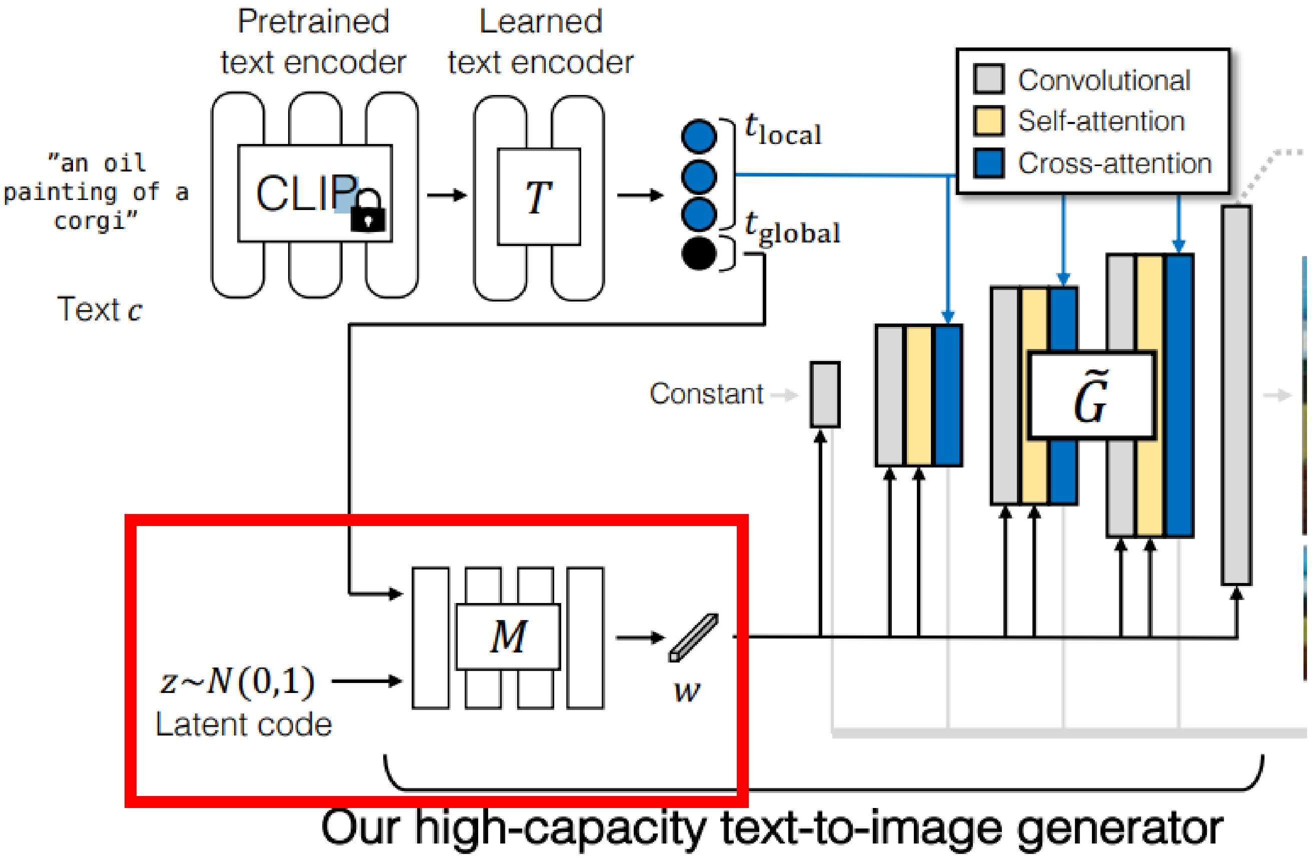
How does GigaGAN works?

- Text Conditioning (CLIP)
- **GAN Synthesis.**
- Text Based GAN Discriminator.
- Upsampler Module.

GigaGAN - Mapping Network

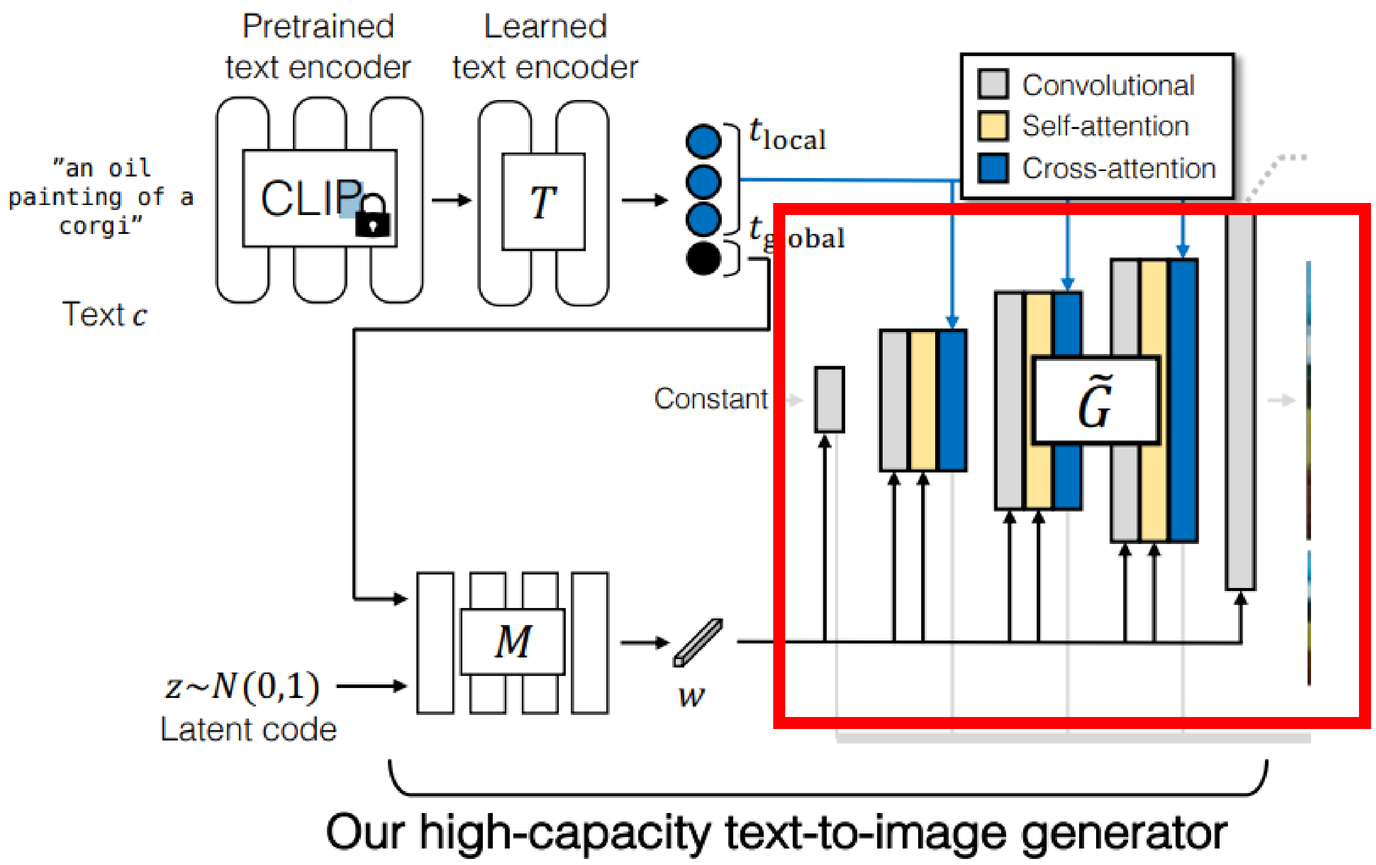
- Mapping Network
  - It mixes the latent code (z) and learned text encoding (t).

- **Key idea**, now we have joint space (w) with the latent code and the text encoding.



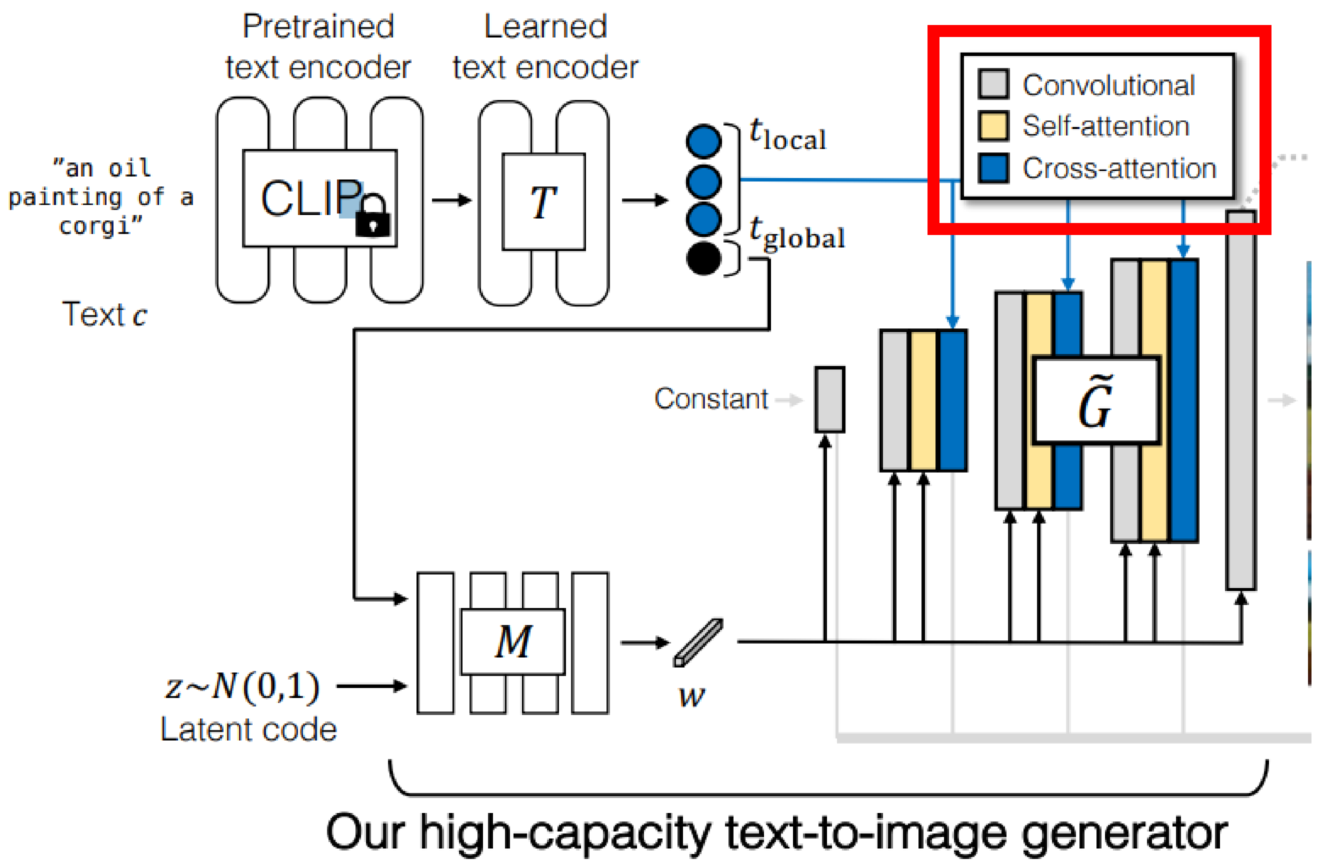
GigaGAN – Syntesis Network

- G maps a learned constant tensor to an output image x.
- Convolution **and attention** are the main tools to generate all output pixels.



GigaGAN – Syntesis Network

- Why would you even care about attention?
  - Trendy sure, but what could be the actual benefit?
  - We are already conditioning the latent vector.



GigaGAN – Training Issues

- There are issues in training very large GANs over a set of diverse data.
- Increasing data allows to increase the width of the convolution layers or the network depth, but becomes too computationally demanding.
- If we rely on convolutions: **The same operation is repeated across all locations.**

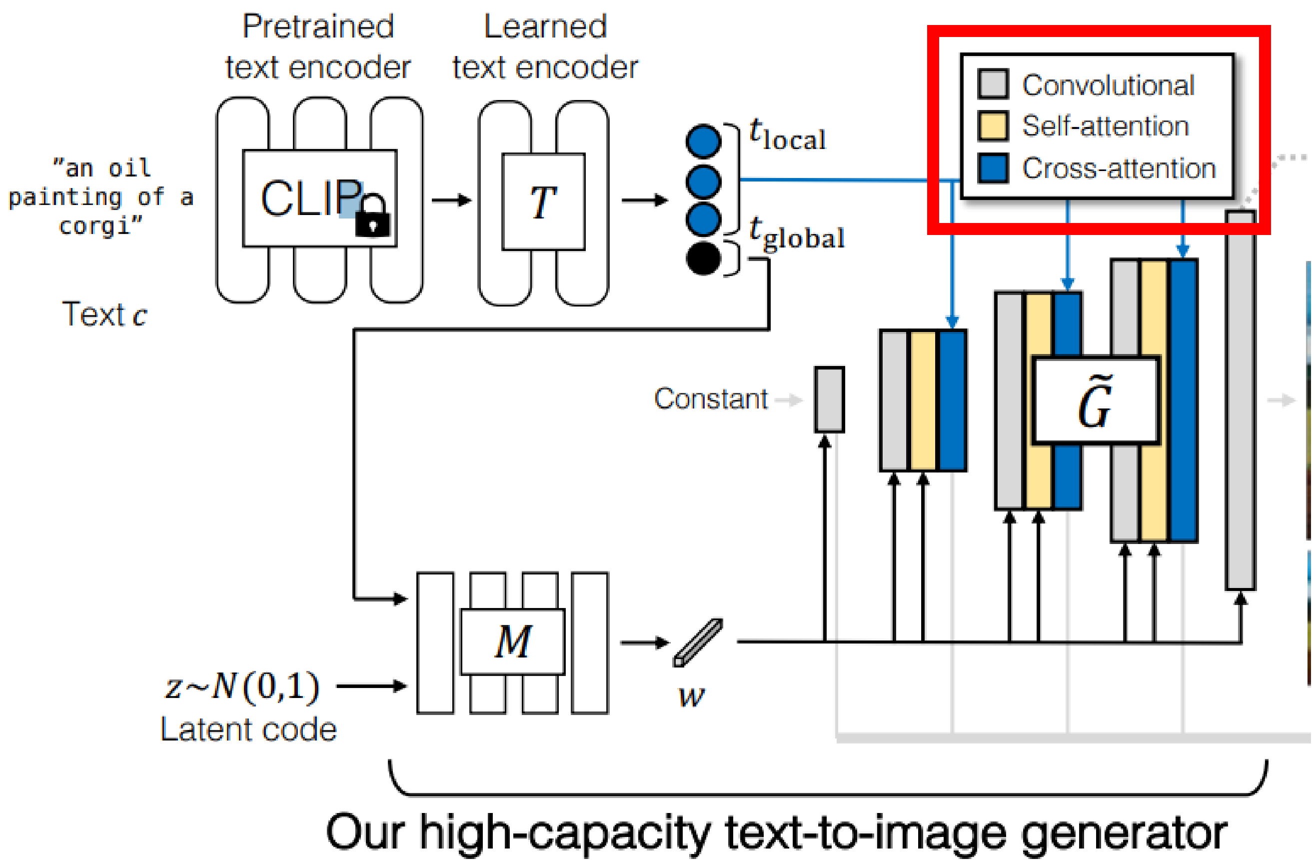
Conditioning the Sythesis Network

- We dont want to use the very same filter across all locations.
- We also dont want to use the very same filter for all the classes.
- Key Insight: The filter should be conditioned **spatially conditioned on the text**.

This is why:

- We need local features.
- We need extra layers over CLIP.
- We include attention layers on the synthesis network.

- Self-Attention vs Cross Attention.



A living room with a fireplace at a wood cabin. Interior design.

a blue Porsche 356 parked in front of a yellow brick wall.

Eiffel Tower, landscape photography

A painting of a majestic royal tall ship in Age of Discovery.

Isometric underwater Atlantis city with a Greek temple in a bubble.

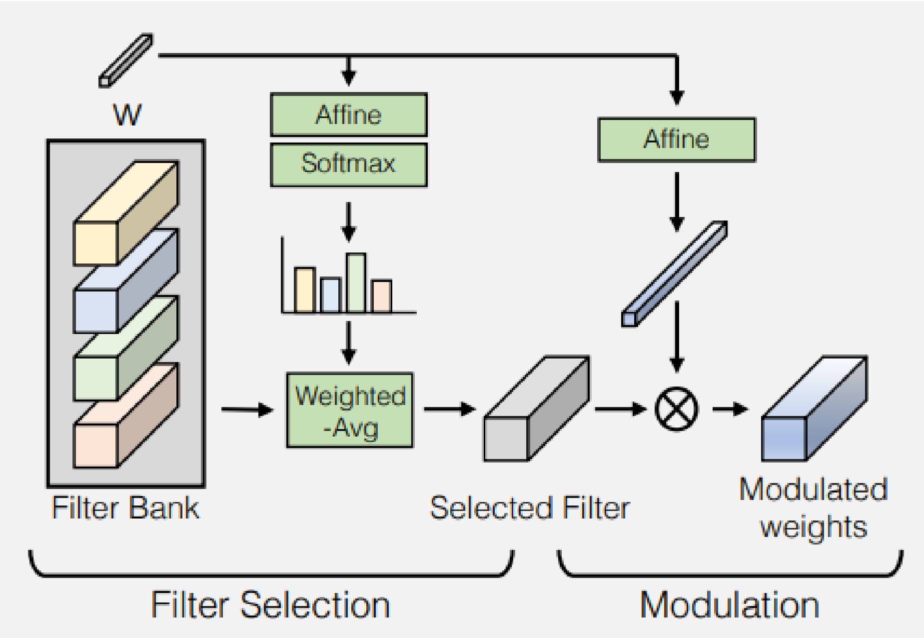
A hot air balloon in shape of a heart. Grand Canyon

low poly bunny with cute eyes

A cube made of denim on a wooden table

Sample-adaptive Kernel Selection.

- GigaGAN proposes to create kern els on-the-fly based on the text conditioning.
- The softmax-based weighting can be viewed as a differentiable filter selection process based on input conditioning.



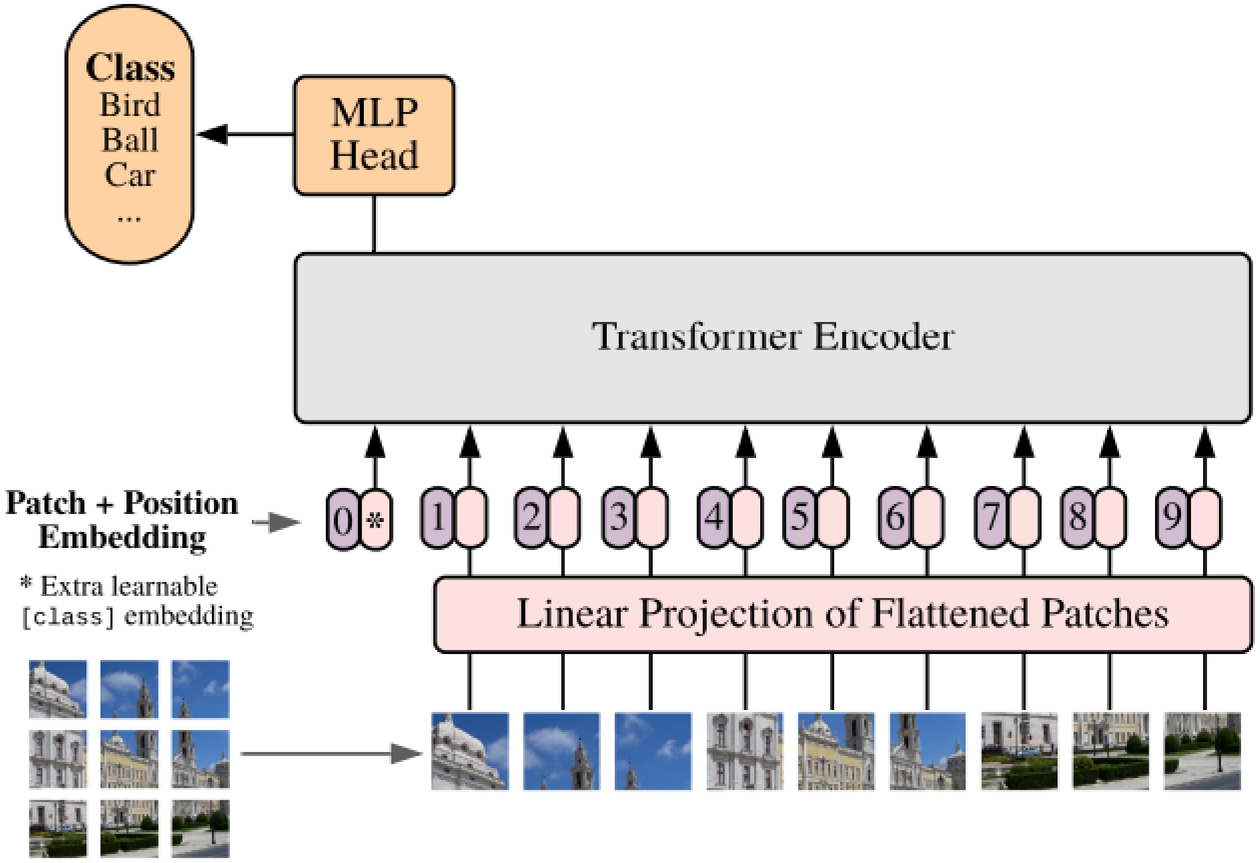
Sample-adaptive kernel selection

Some Extra Details on the Clip Text Head

- Apply additional attention layers T on top to process the word embeddings before passing them to the MLP-based mapping network.

- The **EOT** (“end of text”) component aggregates global information, and is called global.

Local and Global Features



Dosovitski et al.

How does GigaGAN works?

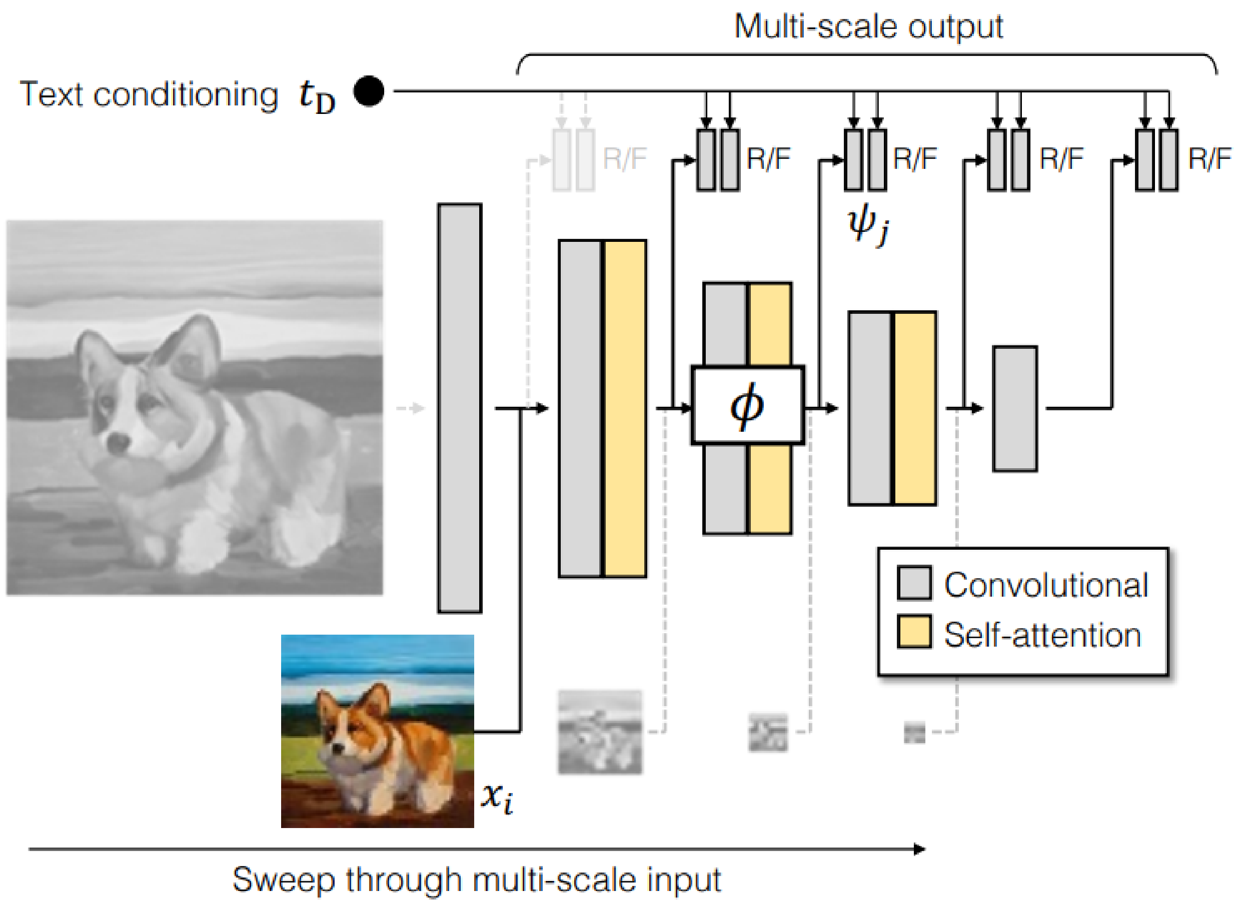
- Text Conditioning (CLIP)
- **GAN Synthesis.**
- **Text Based GAN Discriminator.**
- Upsampler Module.

Multi-scale Synthesis Network

- G generator outputs a **multi-scale image pyramid** with L = 5 levels, instead of a single image at the highest resolution.
- Spatial resolutions: {64, 32, 16, 8, 4}
- Each image of the pyramid is independently used to compute the GAN loss.

GigaGAN Discriminator

- Two branches for processing the image and the text conditioning t_D
- The text branch processes the text similar to the generator (**I'm not sure exactlyhow**).
- The image branch receives an image pyramid and makes independent predictions for each image scale.



How does GigaGAN works?

Text Conditioning (CLIP)

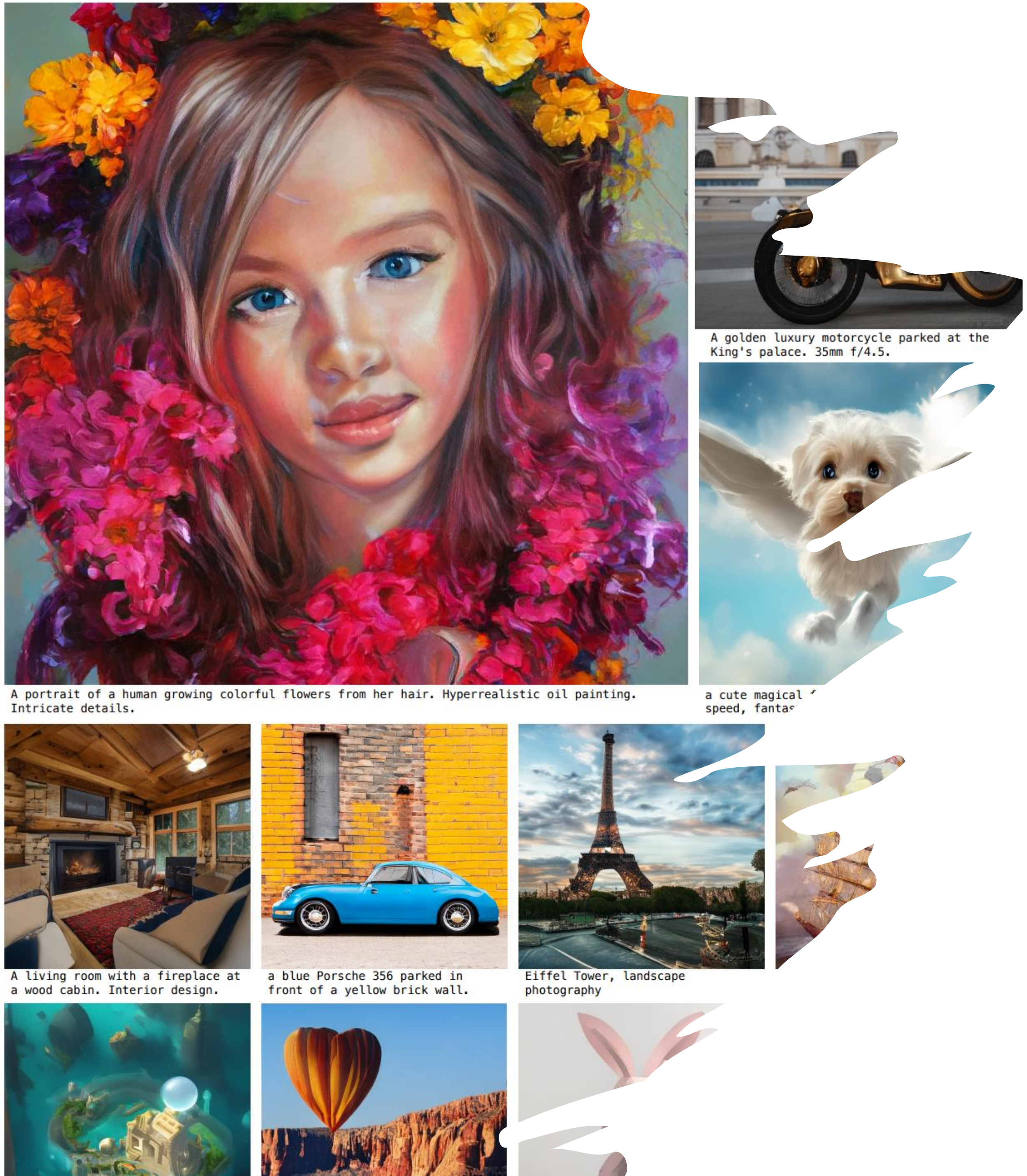
GAN Synthesis.

Text Based GAN Discriminator.

Upsampler Module.

GAN Upsampler

- GigaGAN can be extended to train a text-conditioned super-resolution model.
- Upsampling the outputs of the base GigaGAN generator to obtain high-resolution images at 512px or 2k resolution.



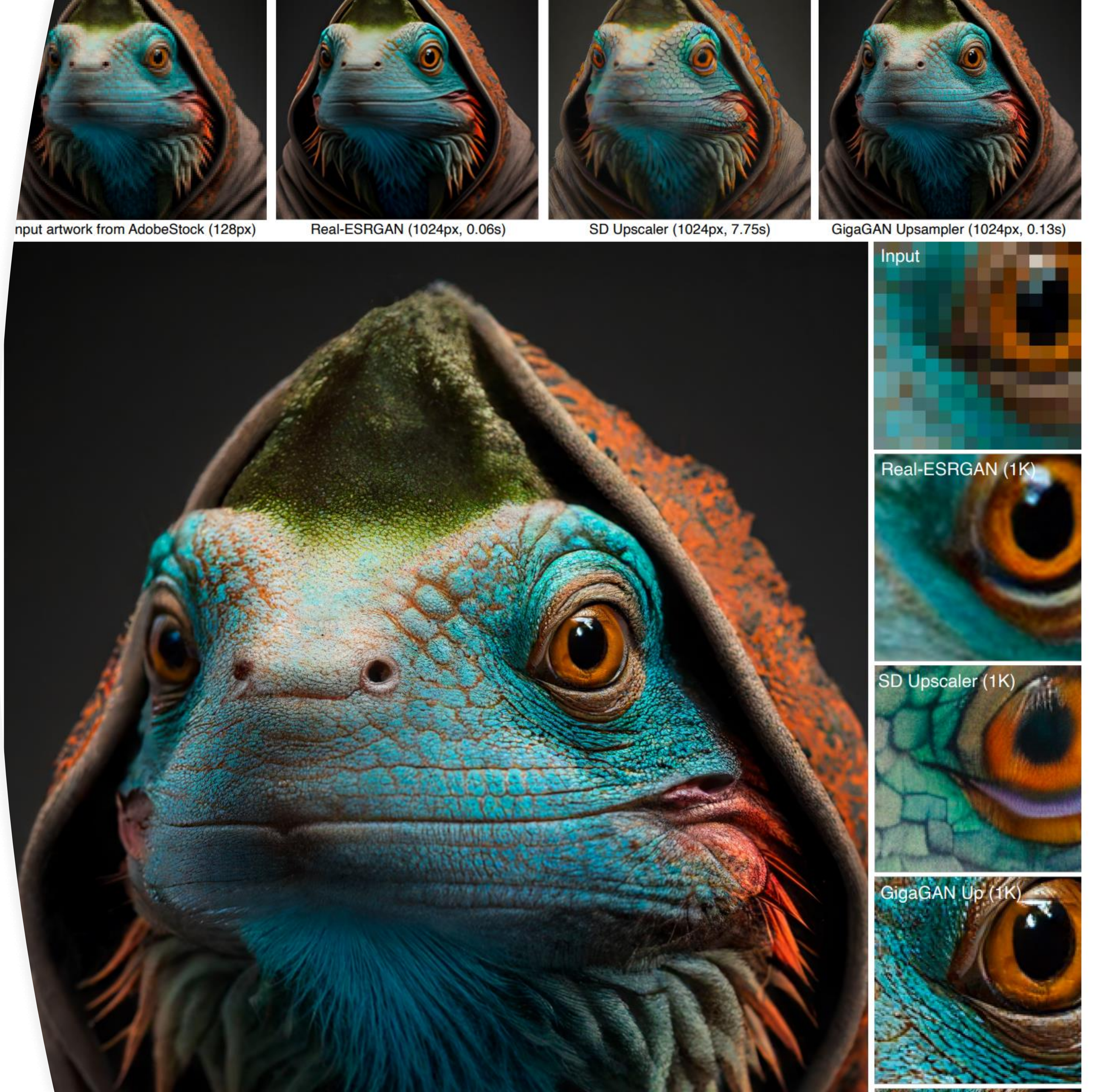
Training Data

- For text-to-image synthesis, we train our models on the union of LAION2B-en and COYO-700M [8] datasets,
- The 128-to1024 upsampler model trained on Adobe's internal Stock images & imagenet.
- Use CLIP ViT-L/14 [71] for the pre-trained text encoder

Latent Space Interpolation

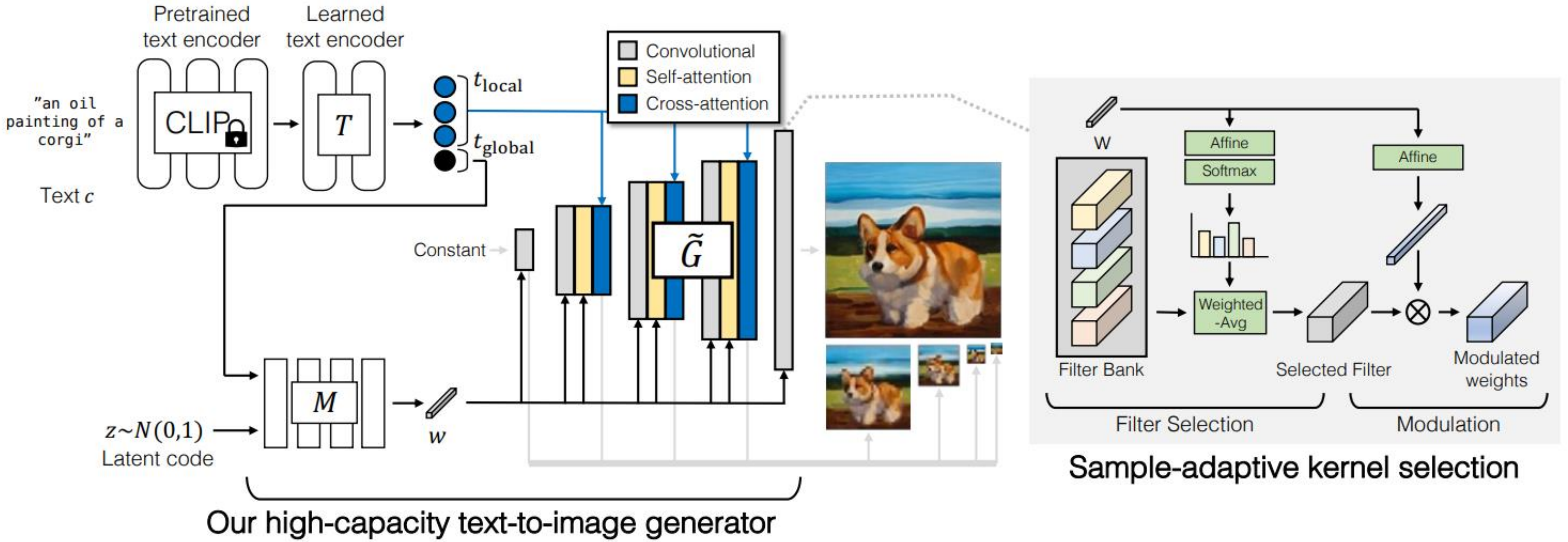
Coarse and Fine Style Control

Upsampler Results



Upsampler Results

Questions?


Pretrained text encoder
Learned text encoder
"an oil painting of a corgi"
Text c
CLIP
T
t_{local}
t_{global}
Convolutional
Self-attention
Cross-attention
Constant
$\tilde{G}$
M
w
$z \sim N(0,1)$
Latent code
Our high-capacity text-to-image generator
W
Affine
Softmax
Affine
Filter Bank
Weighted -Avg
Selected Filter
Modulated weights
Filter Selection
Modulation
Sample-adaptive kernel selection