GigaGAN

Scaling up GANs for Text-to-Image Synthesis

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

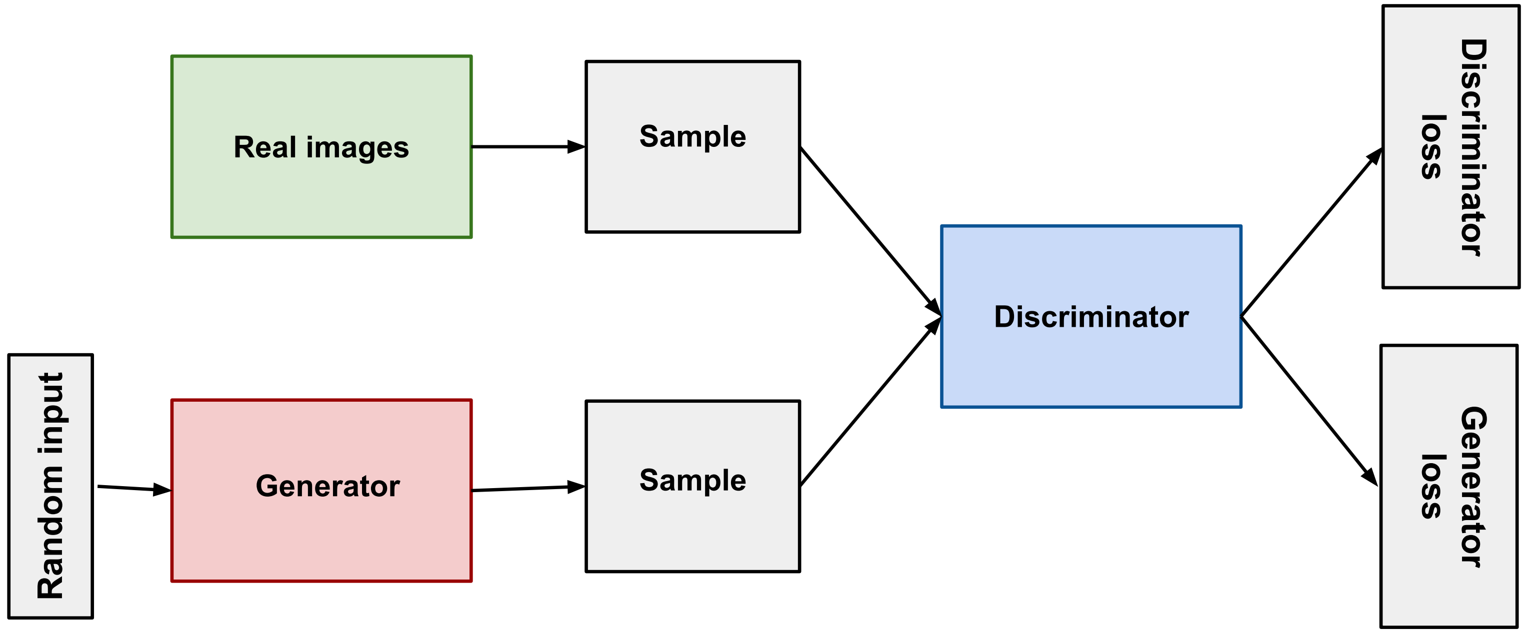
CVPR 2023

GigaGAN

- Text-to-image synthesis
- Contributions
  - Orders of magnitude faster than diffusion models
  - Ultra high-res images at 4k in a few seconds
  - Controllable latent vector space
  - Scalable GAN architecture

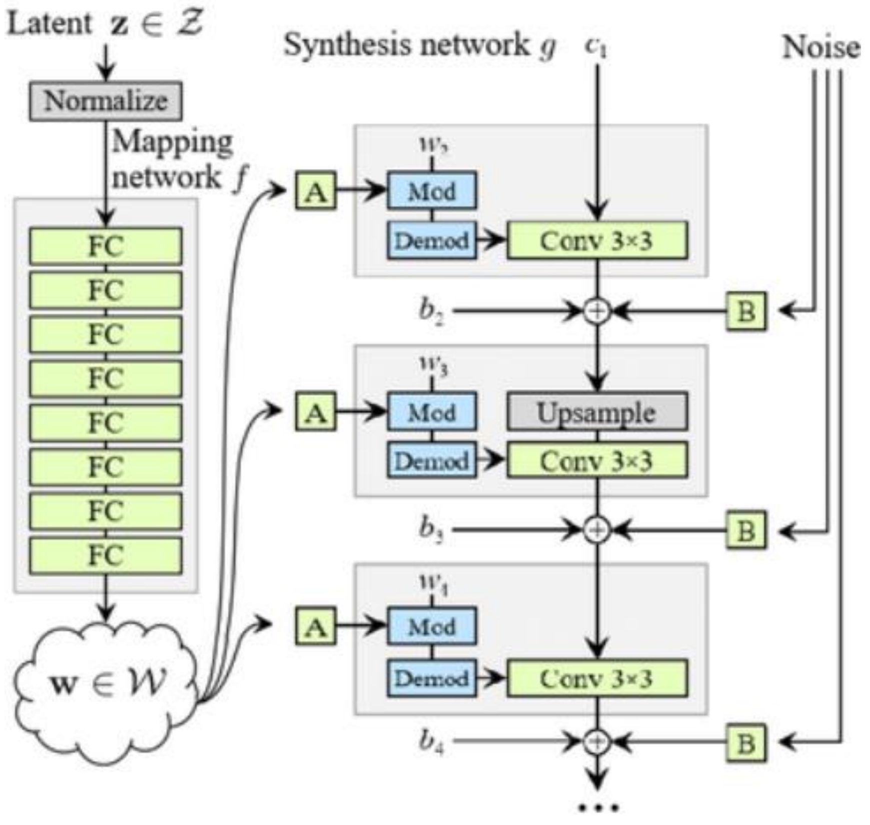
How GAN's work ?

- Proposed by Goodfellow et al. 2014
- Unsupervised learning



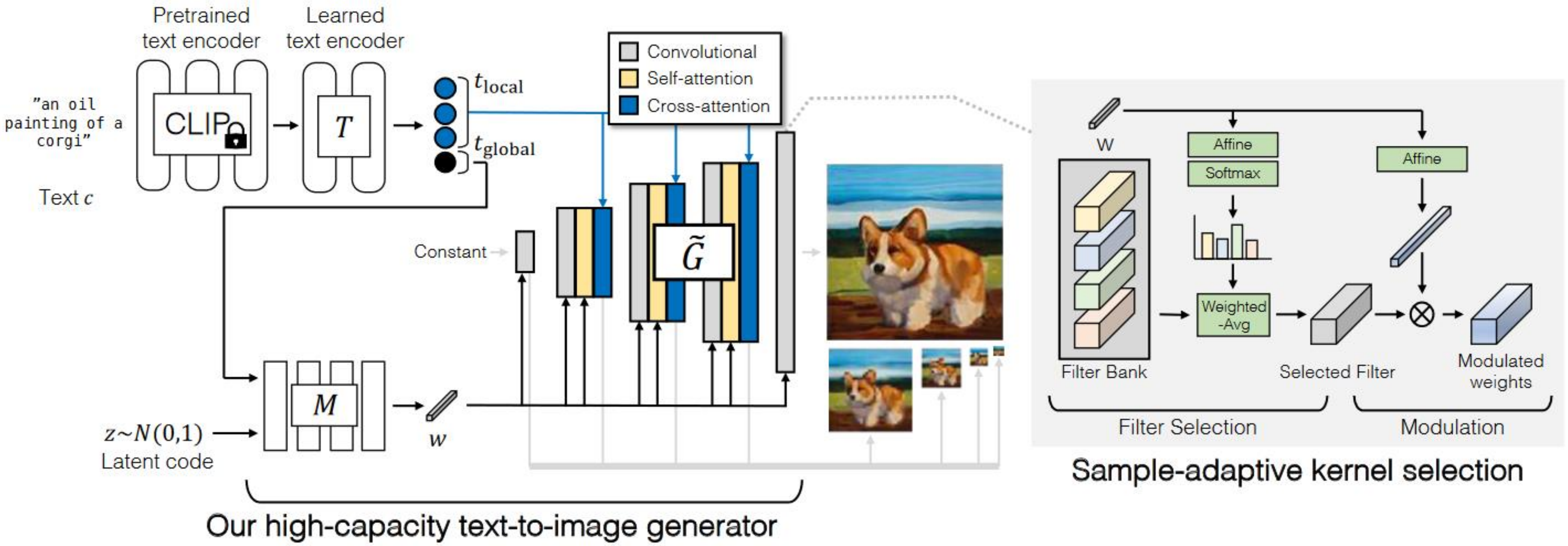
StyleGAN2

- GigaGAN is based on StyleGAN2



(c) StyleGAN2 generator

Architecture - Generator



Architecture - Discriminator

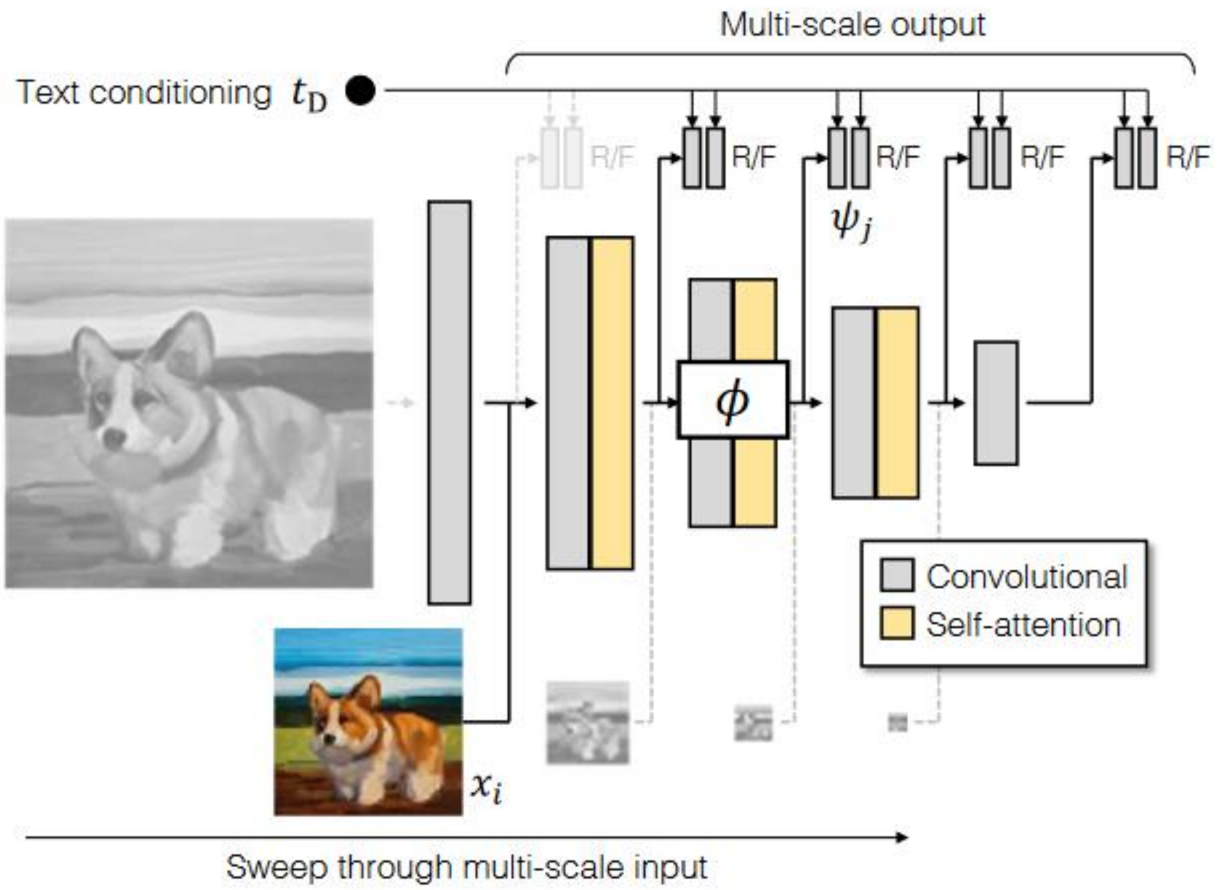
- Multiple discriminators

$$V(G,D) = V_{MS-I/O}(G,D) + L_{CLIP}(G) + L_{Vision}(G)$$

$$\mathcal{V}_{\text{MS-I/O}}(G,D) = \sum_{i=0}^{L-1}\sum_{j=1}^{L} \mathcal{V}_{\text{GAN}}(G_i, D_{ij}) + \mathcal{V}_{\text{match}}(G_i, D_{ij})$$

$$\mathcal{V}_{\text{match}} = \mathbb{E}_{\mathbf{x},\mathbf{c},\hat{\mathbf{c}}}\left[\log(1+\exp(D(\mathbf{x},\hat{\mathbf{c}}))) + \log(1+\exp(D(G(\mathbf{c}),\hat{\mathbf{c}}))\right]$$

$$\mathcal{L}_{\text{CLIP}} = \mathbb{E}_{\{\mathbf{c}_n\}}\left[-\log\frac{\exp(\mathcal{E}_{\text{img}}(G(\mathbf{c}_0))^\top\mathcal{E}_{\text{txt}}(\mathbf{c}_0))}{\sum_n \exp(\mathcal{E}_{\text{img}}(G(\mathbf{c}_0))^\top\mathcal{E}_{\text{txt}}(\mathbf{c}_n))})\right]$$



Results - Metrics

Model	FID-10k ↓	CLIP Score ↑	# Param.
StyleGAN2	29.91	0.222	27.8M
+ Larger (5.7×)	34.07	0.223	158.9M
+ Tuned	28.11	0.228	26.2M
+ Attention	23.87	0.235	59.0M
+ Matching-aware D	27.29	0.250	59.0M
+ Matching-aware G and D	21.66	0.254	59.0M
+ Adaptive convolution	19.97	0.261	80.2M
+ Deeper	19.18	0.263	161.9M
+ CLIP loss	14.88	0.280	161.9M
+ Multi-scale training	14.92	0.300	164.0M
+ Vision-aided GAN	13.67	0.287	164.0M
+ Scale-up (**GigaGAN**)	9.18	0.307	652.5M

	Model	Type	# Param.	# Images	FID-30k ↓	Inf. time
256	DALL·E [75]	Diff	12.0B	1.54B	27.50	-
	GLIDE [63]	Diff	5.0B	5.94B	12.24	15.0s
	LDM [79]	Diff	1.5B	0.27B	12.63	9.4s
	DALL·E 2 [74]	Diff	5.5B	5.63B	10.39	-
	Imagen [80]	Diff	3.0B	15.36B	7.27	9.1s
	eDiff-I [5]	Diff	9.1B	11.47B	6.95	32.0s
	Parti-750M [101]	AR	750M	3.69B	10.71	-
	Parti-3B [101]	AR	3.0B	3.69B	8.10	6.4s
	Parti-20B [101]	AR	20.0B	3.69B	7.23	-
	LAFITE [108]	GAN	75M	-	26.94	0.02s
512	SD-v1.5* [78]	Diff	0.9B	3.16B	9.62	2.9s
	Muse-3B [10]	AR	3.0B	0.51B	7.88	1.3s
	GigaGAN	GAN	1.0B	0.98B	9.09	0.13s

Results – Text-to-image

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.

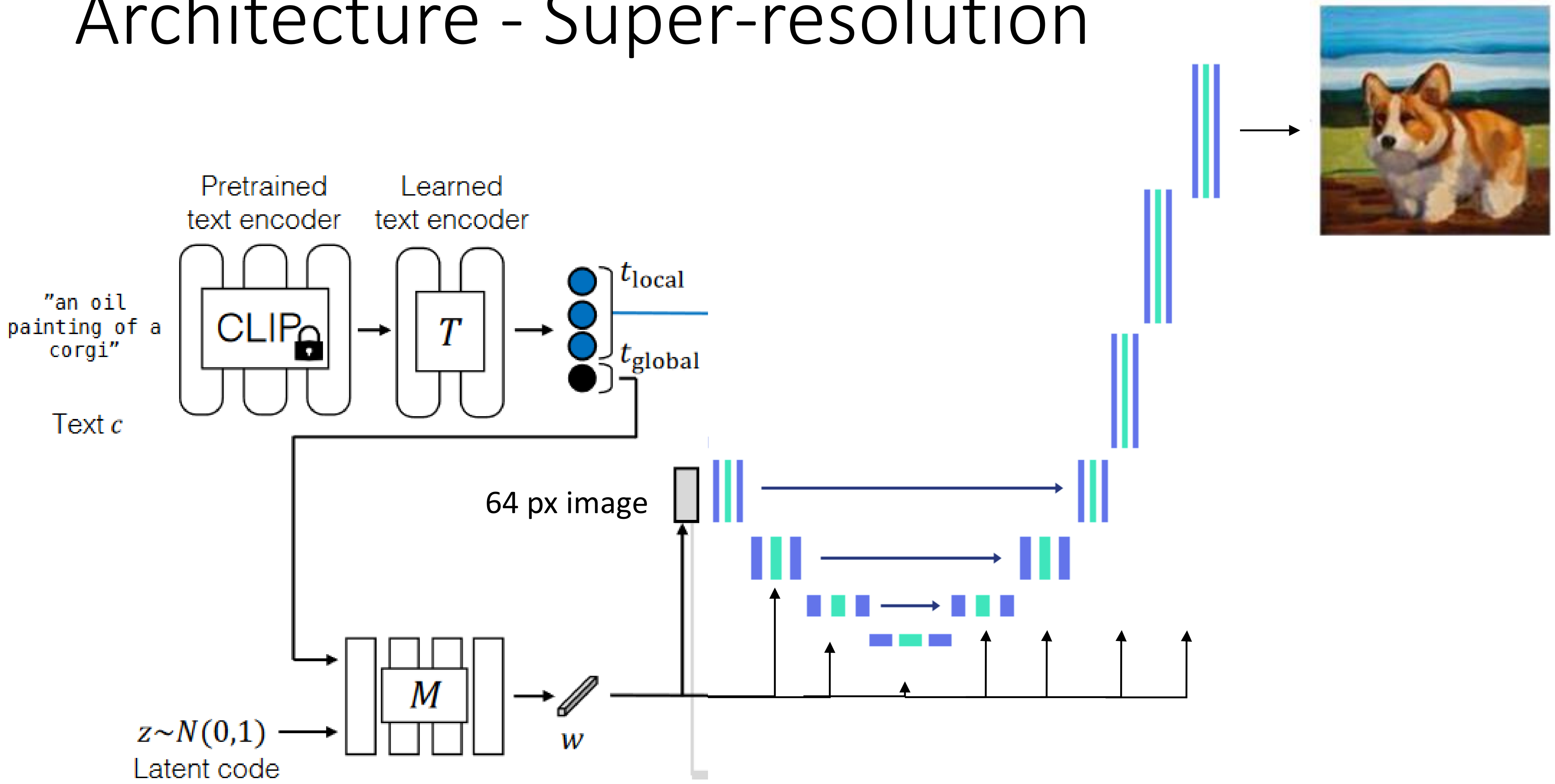
A hot air balloon in shape of a heart. Grand Canyon

low poly bunny with cute eyes

A cube made of denim on a wooden table

Results - Controls

Architecture - Super-resolution



Results – Super-resolution

"An elephant spraying water with its trunk".

Comparison

“A teddy bear on a skateboard in times square.”

Ours (512px, 0.13s / img)

Stable Diffusion v1.5 (512px, 2.9s / img, 50 steps, guidance=7.5)

DALL·E 2 (1024px)

Conclusion

- Lower synthesis quality…
- … but better metrics
- Controllable latent space
- Faster inference
- GANs are still a **viable option** for text-to-image synthesis

- Project page