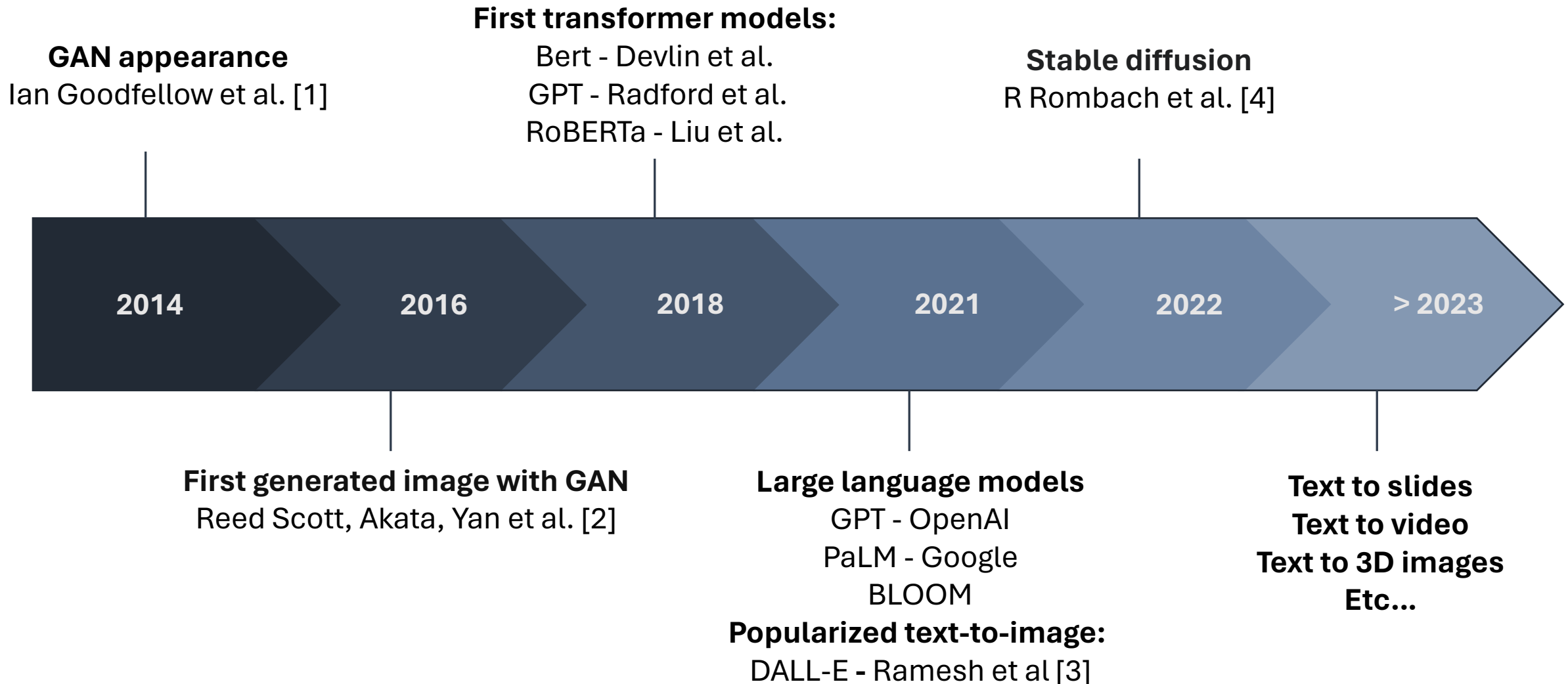


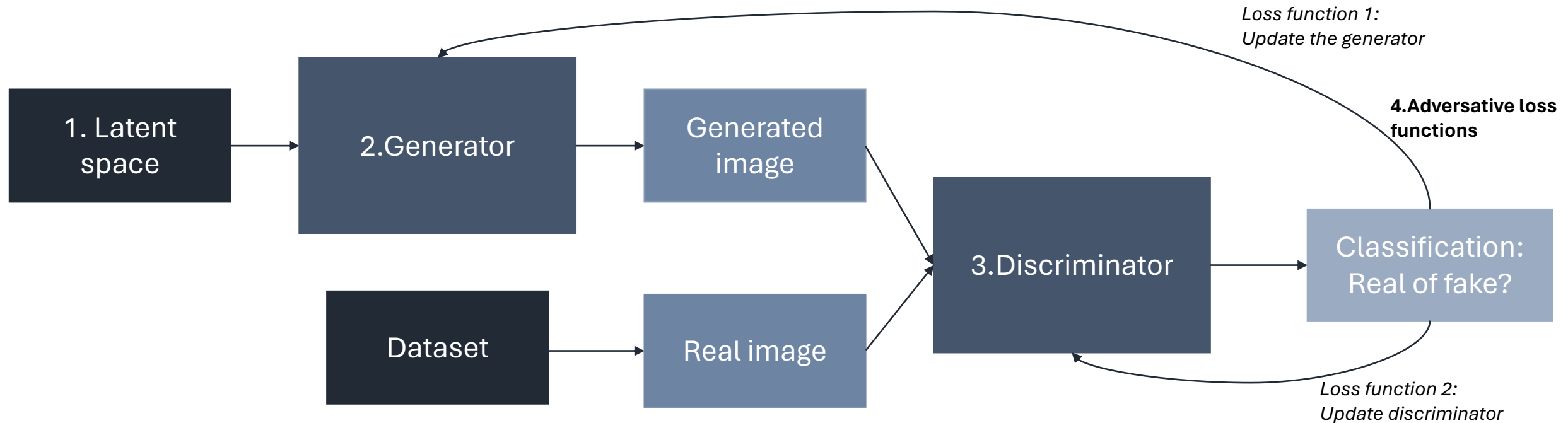


Image Generation from Texts and Multimodal Models

Apparition & Evolution of generative image from text



General idea of text-to-image network architecture



1. Create latent space

- Create dense vectors
- Employ LSTM, GRU, or Transformers (BERT, GPT) for contextual treatment (conditioning and attention mechanism)

2. Generator network:

- Generates images, the quality is progressively enhanced through the training.

3. Discriminator Network:

- Evaluates whether the images are authentic or not and how well they match the text.

4. Adversative loss functions:

- The first aims to minimize the loss by producing realistic images.
- The second aims to maximize the number of accurately classified images.

Key figures of generatives IA



Figure 1 : Number of parameters for several pre trained models (NVIDIA, 2021)

Financial valorisation M\$	
Open AI	90 000
Stable diffusion	1 000
Investment in IA M\$	
1. USA	250
2. China	90
3. United Kingdom	18
...	
6. France	6.5

Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

Classical models:
Use of **plain text**

PROMPT:

A photo of a Corgi dog riding a bike in Times square. It is wearing sunglasses and a beach hat.



Use of **Rich and expressive texts** :
Front, style, links, footnotes



Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

Font Color

PROMPT:

A beautiful girl with big eye, skin, and **long hair**, t-shirt, bursting with vivid colors.



Font Styles

PROMPT:

A male Lycan wolf wearing jet trench coat playing a guitar, high fantasy, highly detailed, digital painting, concept art



Font Color

PROMPT:

A beautiful girl with big eye, skin, and **long hair**, t-shirt, bursting with vivid colors.



Font Styles

PROMPT:

A male Lycan wolf wearing jet trench coat playing a **GUITAR**, high fantasy, highly detailed, digital painting, concept art



Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

FOOTNOTE

PROMPT:

A picture of a lego
brickheadz of a girl with
short hair and a dress



HYPERLINK

PROMPT:

A kid wearing a backpack
riding a bike in a street with
fallen leaves.



FOOTNOTE

PROMPT:

A picture of a lego
brickheadz of a girl with
short hair and a dress

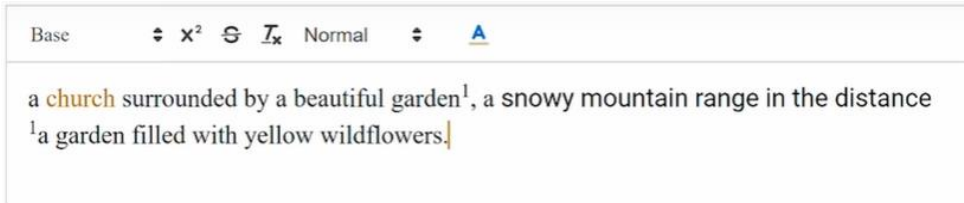


lego white wedding dress



Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

Rich Text editor

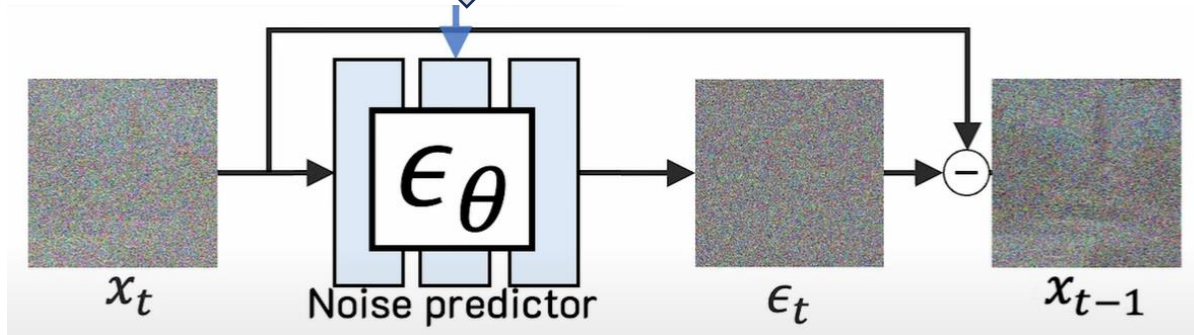


Raw text

Extraction

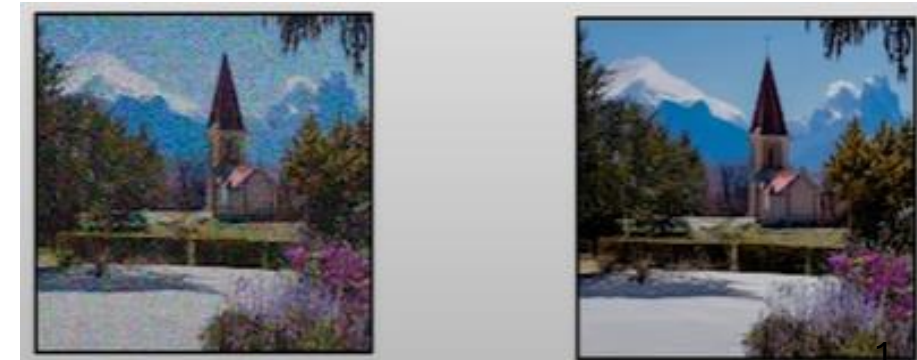
A church surrounded by a beautiful garden, a snowy mountain range in the distance

Denoising



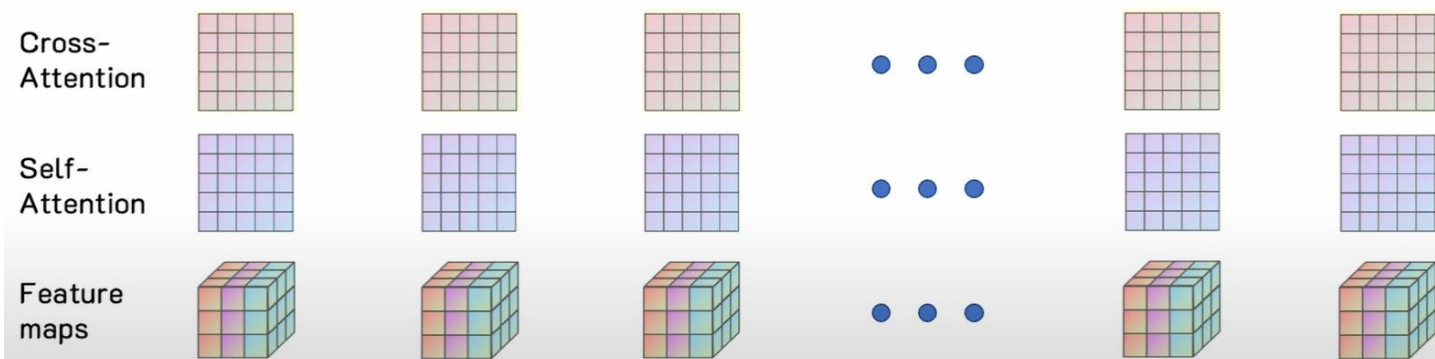
x1

x0

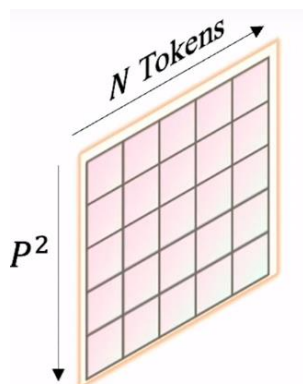


Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

In parallel, collect of "maps"

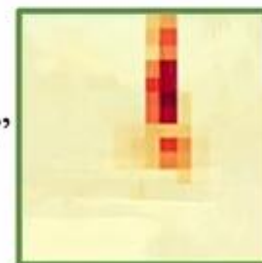


In particular : Cross-Attention



Info of token + location in image

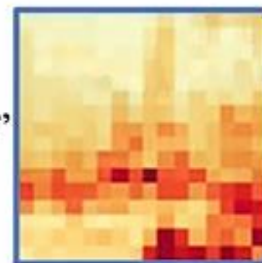
"church"



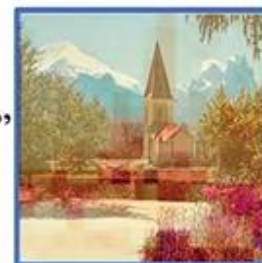
"church"



"garden"



"garden"



Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

To reinforce the detection : self-attention



Cross + Self map



labelling



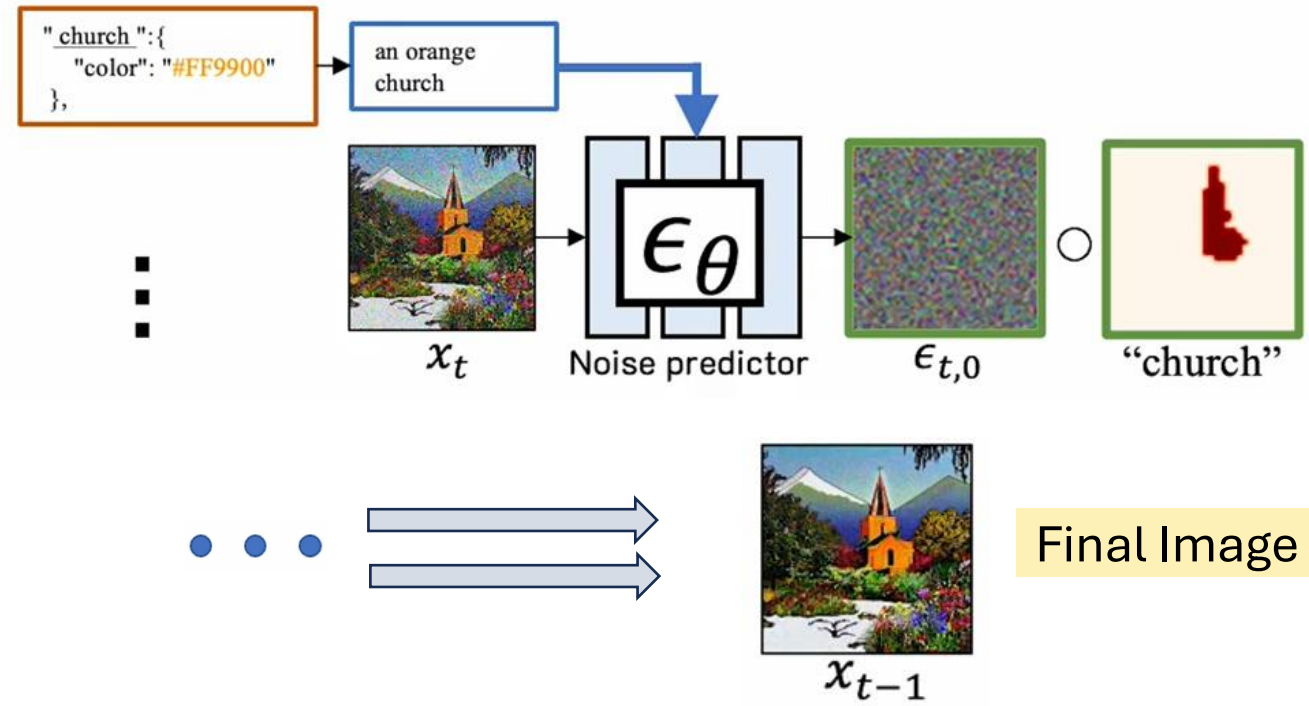
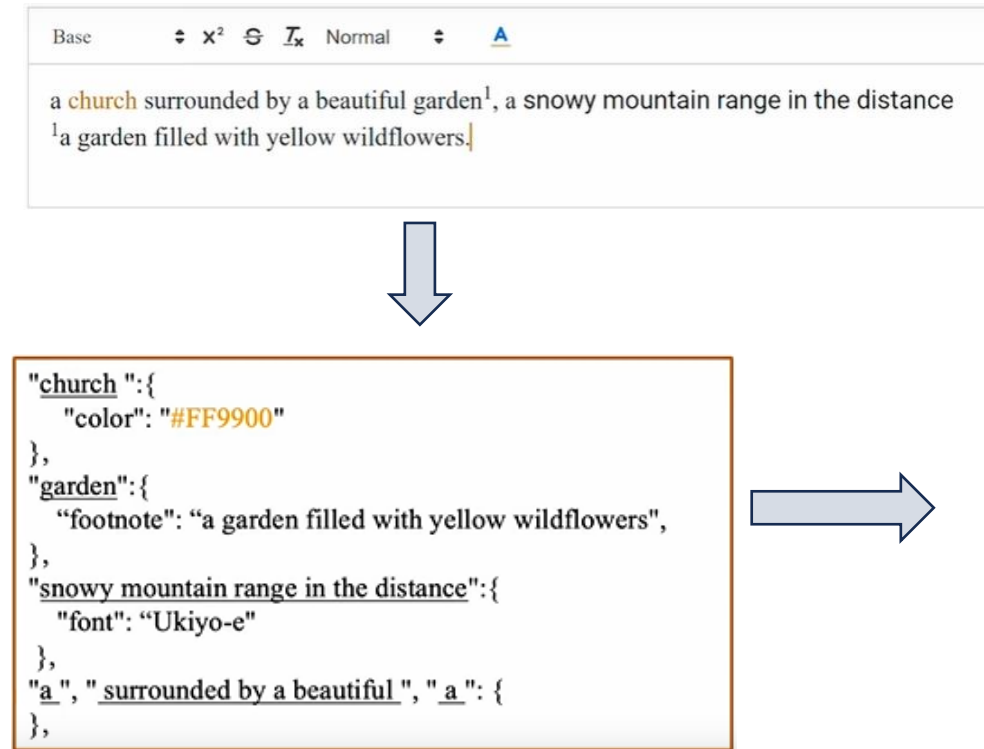
“church”



“garden”

Expressive Text-to-Image Generation with Rich Text, Ge et al., 2023 [5]

Token map



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,
N. Ruiz et al, 2023 [10]

A new approach for personalization of text-to-image diffusions models



Input images



in the Acropolis



swimming



sleeping



in a doghouse



in a bucket



getting a haircut

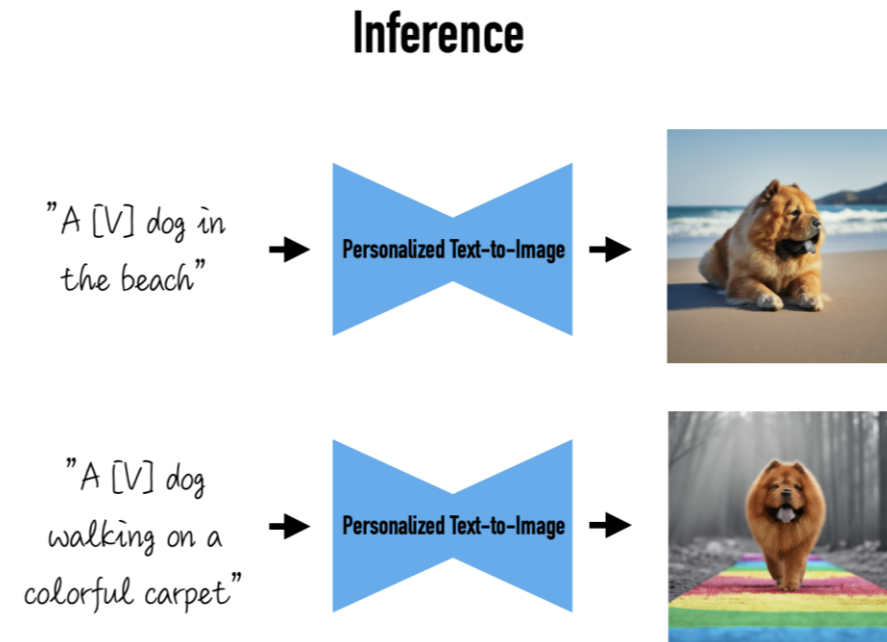
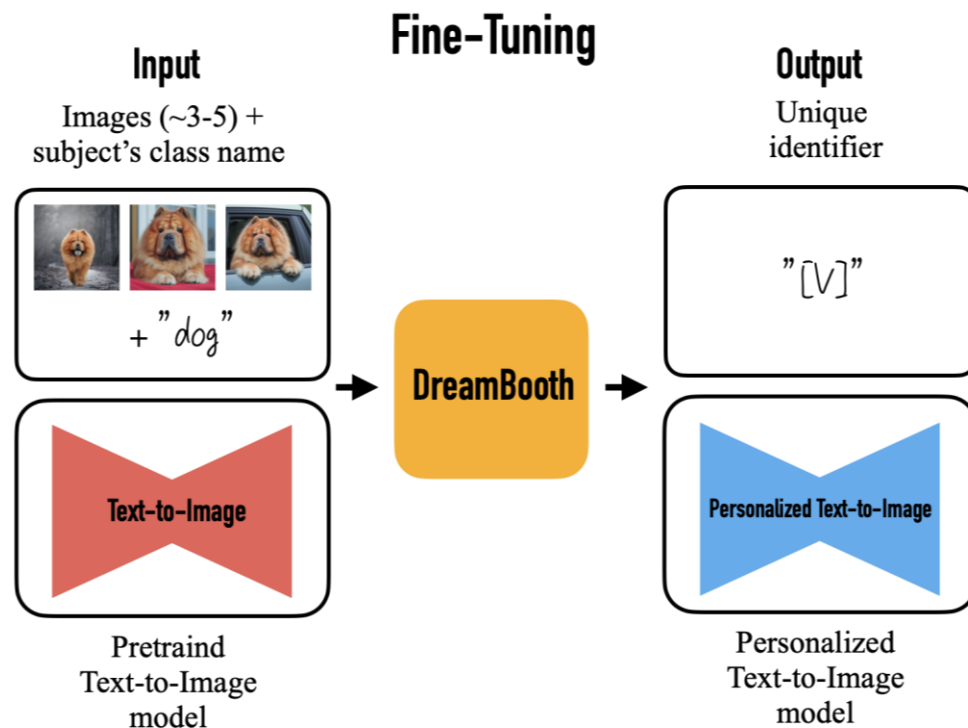
<https://arxiv.org/pdf/2208.12242.pdf>

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, N. Ruiz et al, 2023 [10]

Takes as input a few images of a subject and a class name

Returns a fine-tuned text-to-image model with that encodes a unique identifier referring to the subject

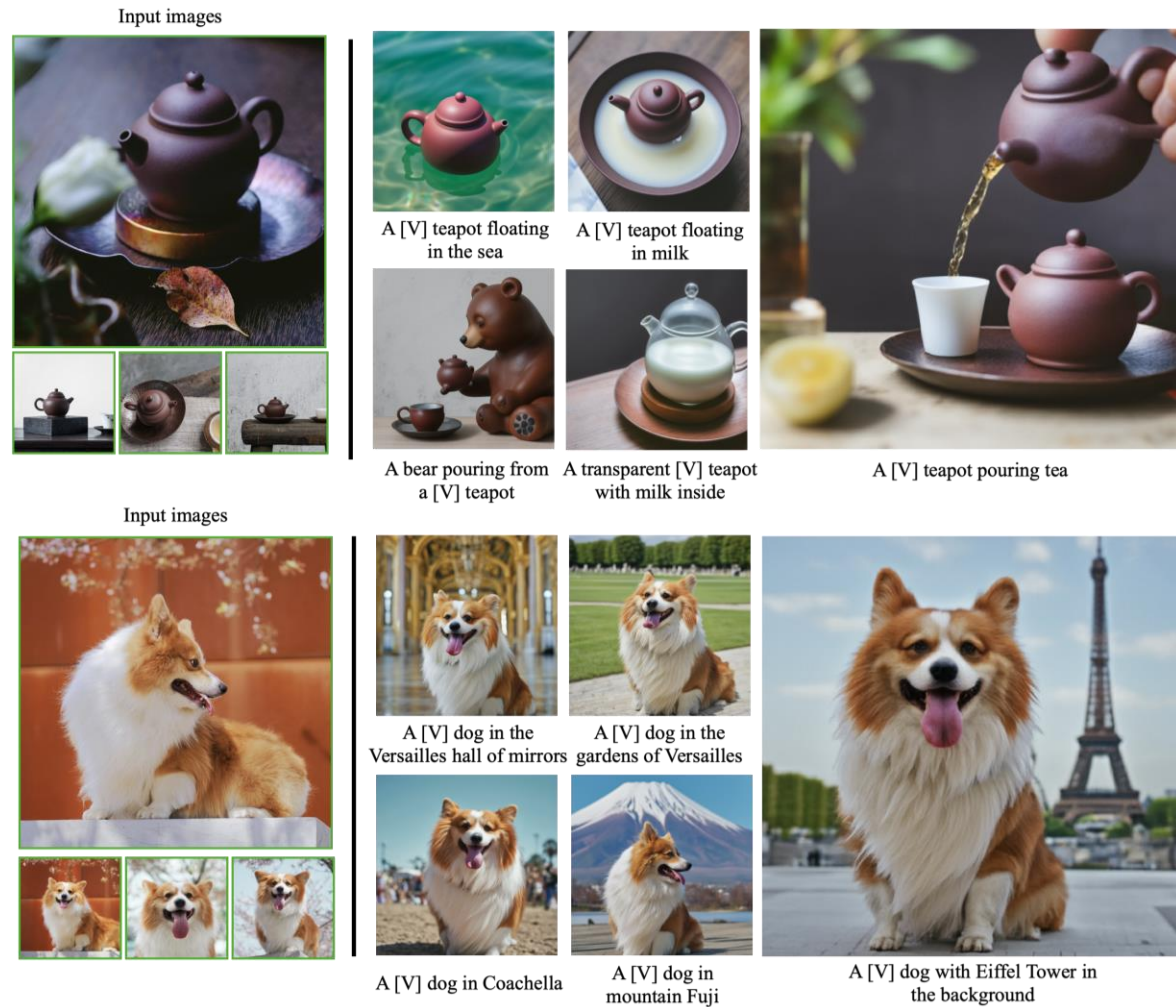
In this work they used Imagen for the pretrained model as the base model



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,
N. Ruiz et al, 2023 [10]

Recontextualization :

a [V] [class noun] [context description]

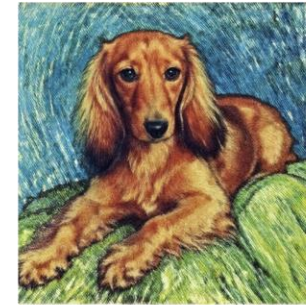


DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, N. Ruiz et al, 2023 [10]

Art Rendition :

a painting/sculpture of a [V]
[class noun] in the style
of [famous painter/sculptor]

Input images



Vincent Van Gogh



Michelangelo



Rembrandt



Johannes Vermeer



Pierre-Auguste Renoir



Leonardo da Vinci

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, N. Ruiz et al, 2023 [10]

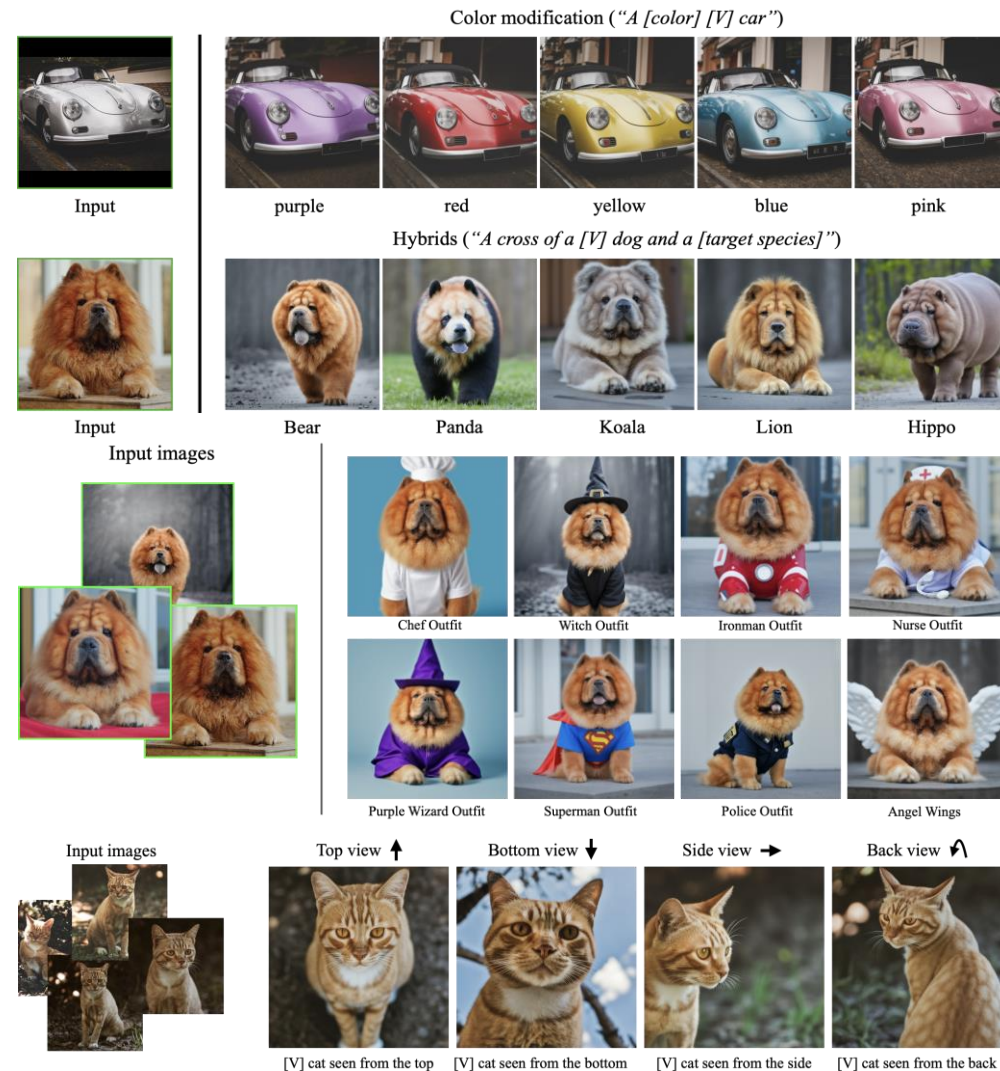
Property modification :

Color modification

Hybridation

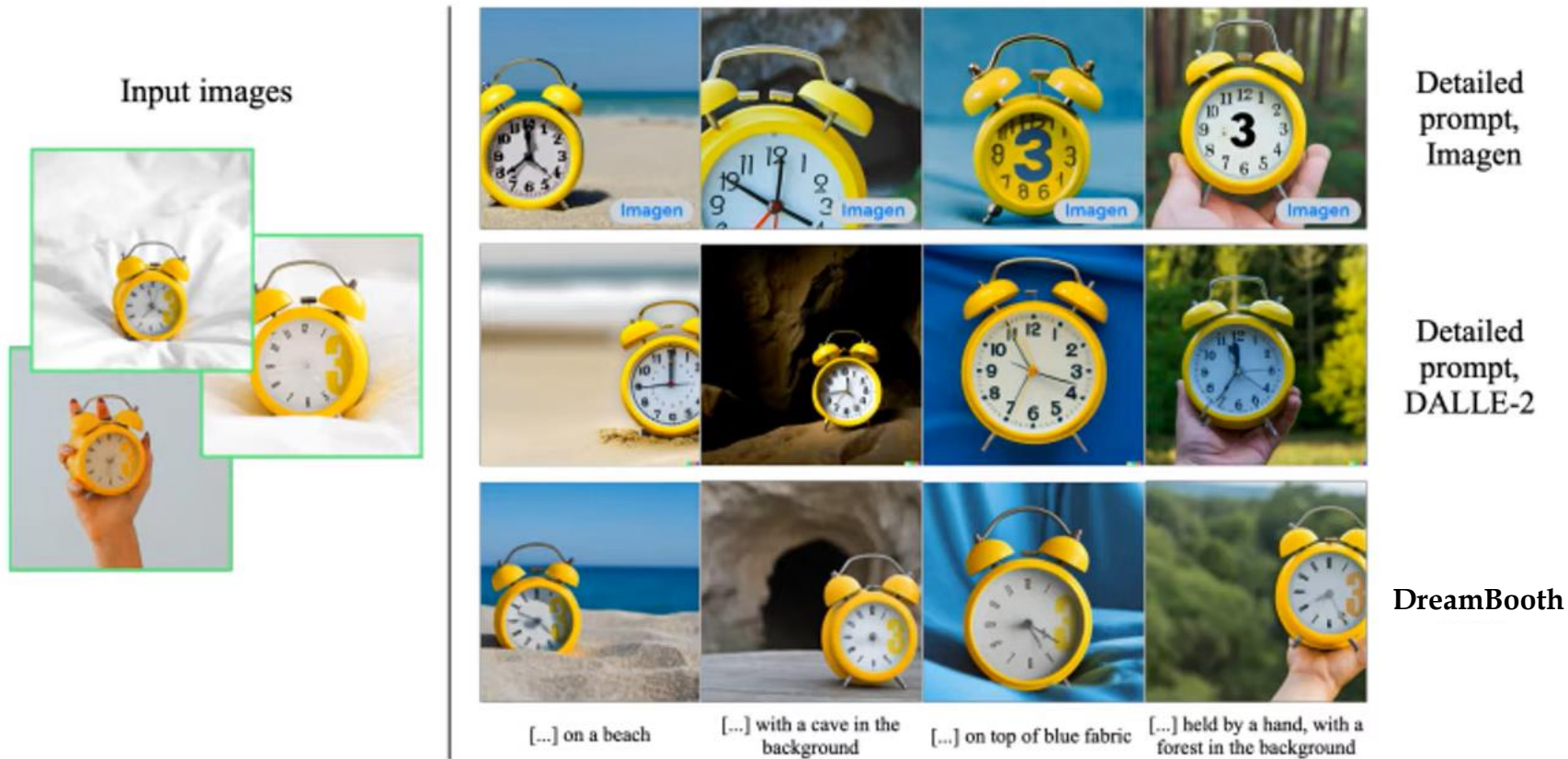
Accessorization

Text-Guided View Synthesis



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, N. Ruiz et al, 2023 [10]

Comparison with other models :



Prompt : “retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face”

DALL-E : Zero-Shot Text-to-Image Generation, A. Ramesh et al., 2021, [11]

Goal : Train a Transformer (GPT-3) to autoregressively model the text and image tokens as a single stream of data.

Stage 1

Compress each 256×256 RGB image into a 32×32 (factor 64) grid of image tokens using a VAE.

Each element can take 8192 possible values to ensure a diversity.

Objective : Reduce the input space of the transformer by a factor of 192 (64×3)



Comparison of original images (top) and reconstructions from the discrete VAE (bottom)

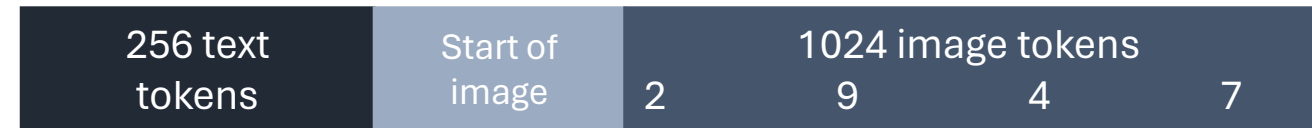
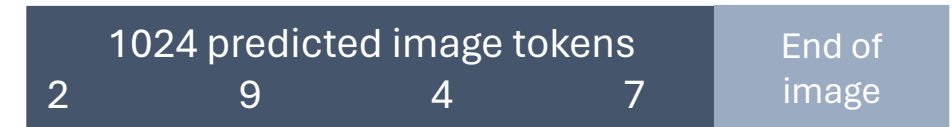
DALL-E : Zero-Shot Text-to-Image Generation, A. Ramesh et al., 2021, [11]

Goal : Train a Transformer (GPT-3) to autoregressively model the text and image tokens as a single stream of data.

Stage 2

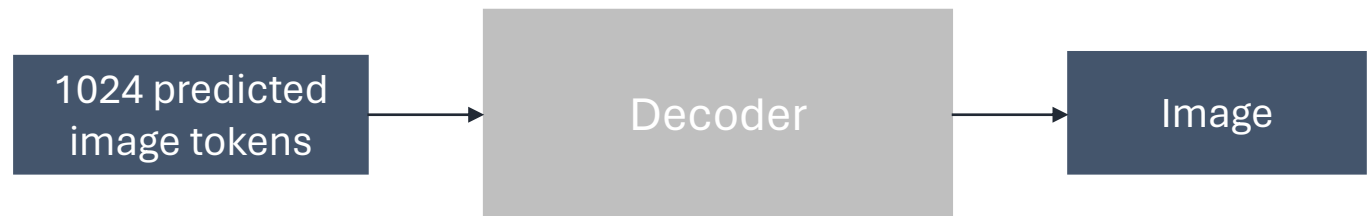
Concatenate in one sequence the 256 BPE-encoded **text** tokens and the $32 \times 32 = 1024$ **images** tokens.

Predict image tokens



« Finally, the text and image tokens are concatenated and modeled autoregressively as a single stream of data »

Create image



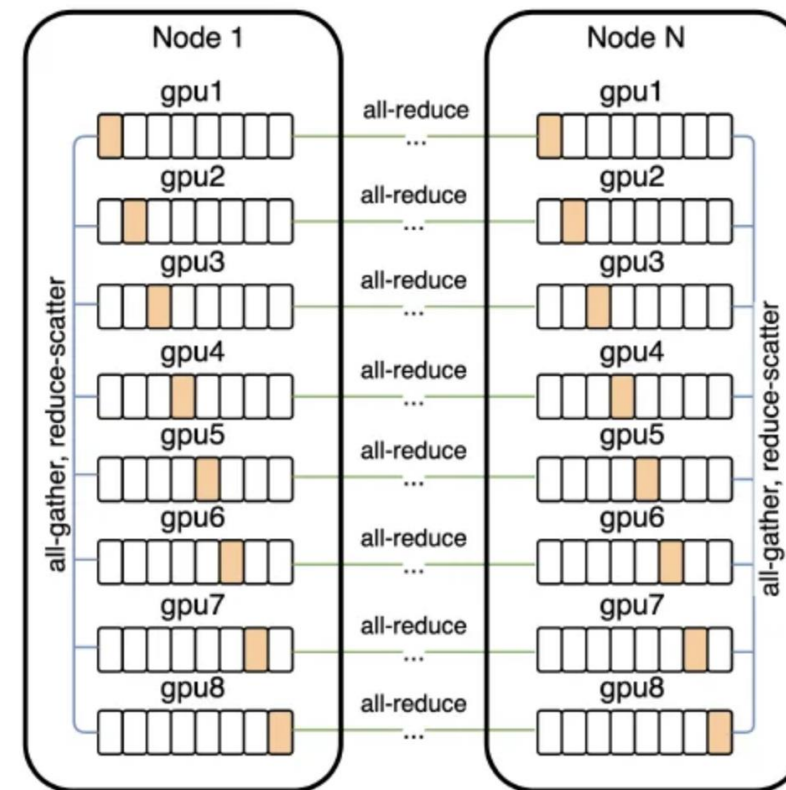
DALL-E : Zero-Shot Text-to-Image Generation, A. Ramesh et al., 2021, [11]

Goal : Train a Transformer (GPT-3) to autoregressively model the text and image tokens as a single stream of data.

Dataset : 250 millions text-images pairs from the internet.

Model : 12-billion parameters, it consumes about 24 GB of memory which exceeds 16 GB NVIDIA V100 GPU

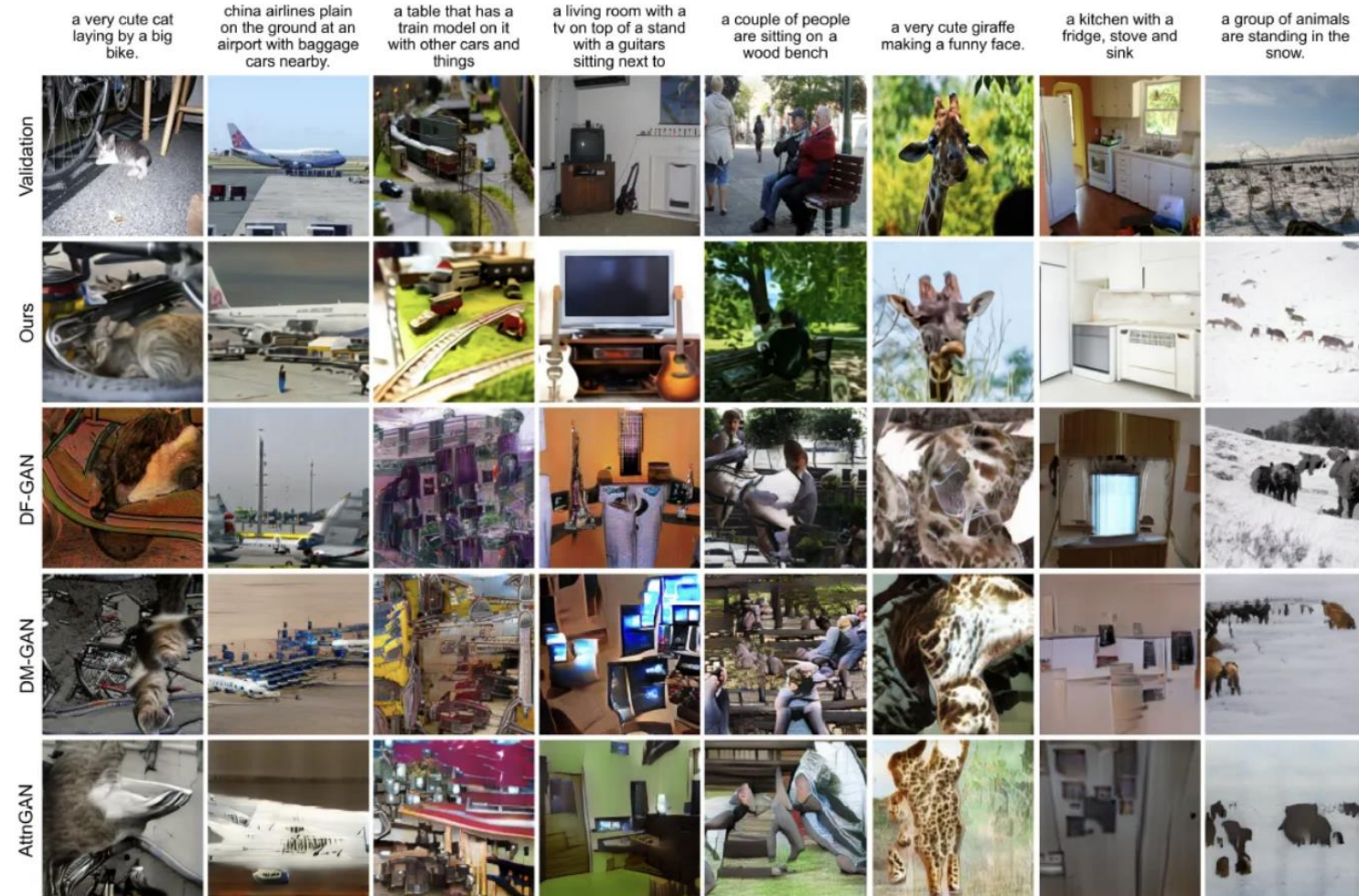
Optimization: Each parameter array is sharded among the eight GPUs on each machine and the gradient is compressed (*Vogels et al., 2019*)



vCommunication patterns used for distributed training

DALL-E : Zero-Shot Text-to-Image Generation, A. Ramesh et al., 2021, [11]

DALL-E's generated images are ranked with a contrastive model (*Radford et al., 2021*). It was the best to create 512 images samples



Comparison of samples from DALL-E model to those from prior approaches on captions from MS-COCO

CONTEXT	CURRENT TECHNIQUES	LIMITATIONS	REFERENCES
<p><i>Electricity consumption of GPT</i></p> <ul style="list-style-type: none"> - 1.287 MWh for its training - 564 MWh per day for 3.500 servers - 30,000 GPUs - Future Scenario: Implementing an AI in Google search - 500,000 servers - 4 million GPUs <p>=> <i>Annual consumption of 29.2 TWh > Ireland's annual consumption of 29.3 TWh.</i></p>		<p><i>GPT's Carbon Footprint</i></p> <ul style="list-style-type: none"> - 260-522T of CO2e : 270 flights between Paris and New York - electrical operation (50%) - server manufacturing - refrigerant gas leaks (per year). - 8,4 tCO2e/an : Daily execution <p>=> Paris Agreement's goal is 2T CO2eq per person and actual consumptions is 8T CO2eq in 2021.</p>	
<p><i>Cost of GPT</i></p> <p>=> 700,000 dollars per day for OpenAI.</p>		<p><i>Water consumption of GPT</i></p> <ul style="list-style-type: none"> - 700 cubic meters of water already used - 25 to 50 interactions require half a liter of water - GPT-4 is even more demanding in terms of water consumption - LaMDa and Bard required over 8.7 million cubic meters of water in 2019 in just three U.S. states. 	



Deep Fake & GANS :

- Legal void
- Plagiarism and intellectual property
- Fraud
- Use of private data
- Infringement of someone else's privacy,
- Violation of the right to one's image.
- Lack of transparency (source...)

=> IA act: regulating the use of artificial intelligence throughout the European Union

CONTEXT	CURRENT TECHNIQUES	LIMITATIONS	REFERENCES
CONTEXTE		image editing instructions. <i>CVPR</i> , 2023	
[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. In <i>Advances in Neural Information Processing Systems</i> , Vol. 27. 139–144		[8] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models.arXiv preprint <i>arXiv:2301.13826</i> , 2023.	
[2] R. Scott, A. Zeynep, Y. Xincheng, L. Lajanugen, S. Bernt, L. Honglak, Generative adversarial text to image synthesis, in <i>ICML</i> 2016		[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. High-resolution image synthesis with latent diffusion models. <i>CVPR</i> , 2022	
[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen Hierarchical Text-Conditional Image Generation with CLIP Latents, <i>arXiv:2204.06125v1</i> , 2022	CURRENT TECHNIQUES : DREAMBOOTH		
[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, in <i>CVPR</i> , 2022		[10] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, arXiv:2208.12242, in <i>CVPR</i> 2023	
CURRENT TECHNIQUES : RICH TEXT		CURRENT TECHNIQUES : ZERO SHOT	
[5] S. Ge, T. Park, JY. Zhu, JB. Huang, Expressive Text-to-Image Generation with Rich Text, <i>Ge et al.</i> , 2023		[11] R. Aditya, P. Mikhail, G. Gabriel, G. Scott, V. Chelsea, R. Alec, C. Mark, S. Ilya. 2021. Zero-shot text-to-image generation. In <i>International Conference on Machine Learning</i> . 8821–8831.	
[6] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. <i>arXiv:2208.01626</i> , 2022.			
[7] T. Brooks, A. Holynski, and A. Efros. Instructpix2pix: Learning to follow			

Carbon Footprint and energy consumption

- <https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d>
- <https://www.hellowatt.fr/blog/chat-gpt-empreinte-carbone/>

Water consumption

- https://www.bfmtv.com/tech/intelligence-artificielle/une-bouteille-par-conversation-chat-gpt-est-un-gouffre-de-consommation-d-eau-fraiche_AV-202304120278.html

ChatGPT cost

- <https://www.20minutes.fr/high-tech/4034128-20230424-chatgpt-fonctionnement-chatbot-coute-pres-700-000-dollars-jour>

Deep Fake and IA act

- https://www.challenges.fr/high-tech/ai-act-pour-les-entreprises-la-fin-du-systeme-d-pour-utiliser-l-ia-risque-de-couter-cher_876982
- <https://linc.cnil.fr/dossier-ia-generative-quelles-regulations-pour-la-conception-des-ia-generatives>