


# DALL-E 2 : Hierarchical Text-Conditional Image Generation with CLIP Latents

*Jiin Kim*

## DALL-E 2 is so hot!




“A CAMERA THAT LOOKS LIKE R2-D2”

5:52


### Creating INSANE Cameras with Artificial intelligence (Dalle.2 demo)

조회수 1.5만회 · 4일 전

 Mathieu Stern ✓

➤ MUSIC 🎵 Here are the best options for great music for monetizing your videos: Epidemic Sound is perfect for cinematic ...

새 동영상




“MOUNT EVEREST MADE OF CAKE”


8:15

### This Is the First A.I. That Really Scared Me (OpenAI DALL-E 2)

조회수 6.4만회 · 1개월 전

 Enrico Tartarotti

DALL-E 2 left me speechless: the text-to-image revolution is here. This is probably one of the most impactful tech advancements in ...




DALL-E CHALLENGE

21:05

### Can AI Replace Our Graphic Designer?

조회수 32만회 · 2주 전

 The Studio ✓

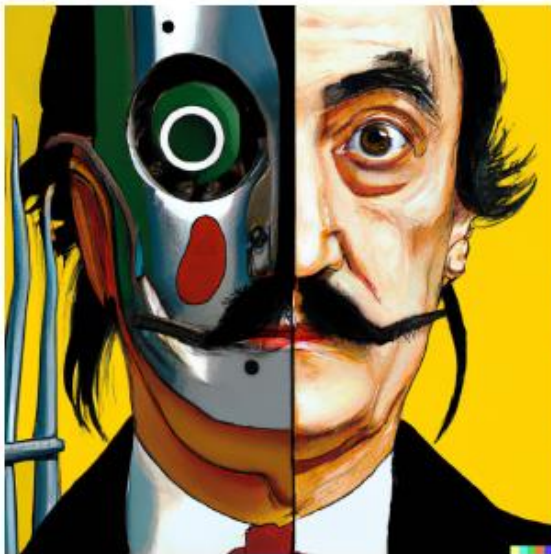
DALL-E 2 from OpenAI can create shockingly impressive images from complex prompts in a matter of seconds — but luckily for ...

4K

Prompt #2

챕터 8 ▾

## Teaser



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula



## Teaser



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor  
napoleon holding a piece of cheese



a teddy bear on a skateboard in times square

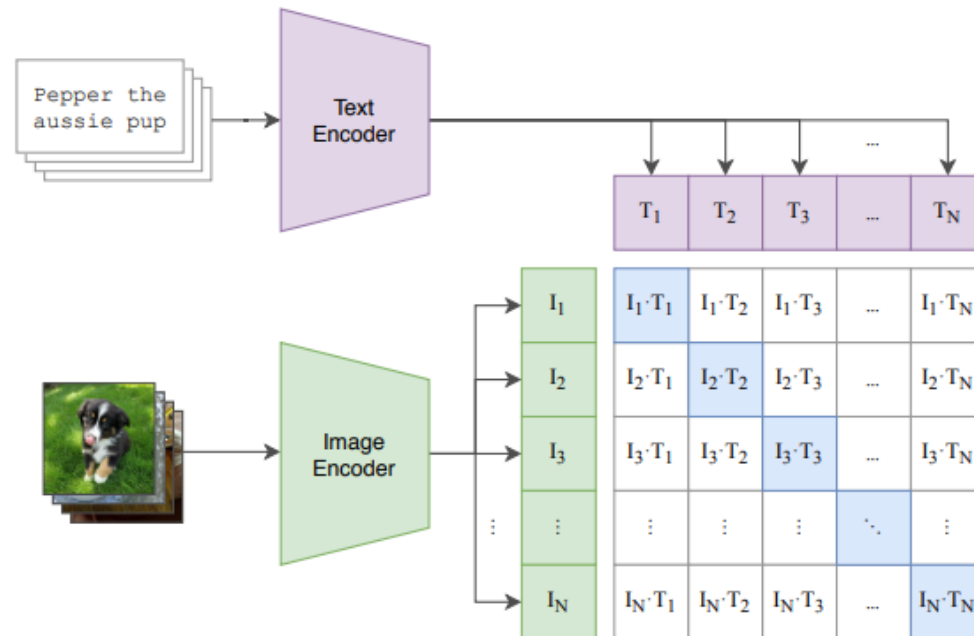
캡션이 주어졌을 때, 캡션과의 유사성을 높게 유지하면서 시각적으로 사실성이 높은 이미지를 생성한다.  
상식과 상상이 필요한 이미지들까지 잘 만들어낸다.

Background knowledge

## CLIP (Radford2021, ICML, OpenAI)

- CLIP 는 text 로 이미지의 class 를 예측하는 모델이고, zero-shot 이 가능하다.
- 4억개의 (이미지, 텍스트) 쌍으로 모델을 학습하였다.
- Image encoder 는 ResNet 또는 Vision Transformer, text encoder 는 CBOW 또는 Text Transformer 를 사용하였다.

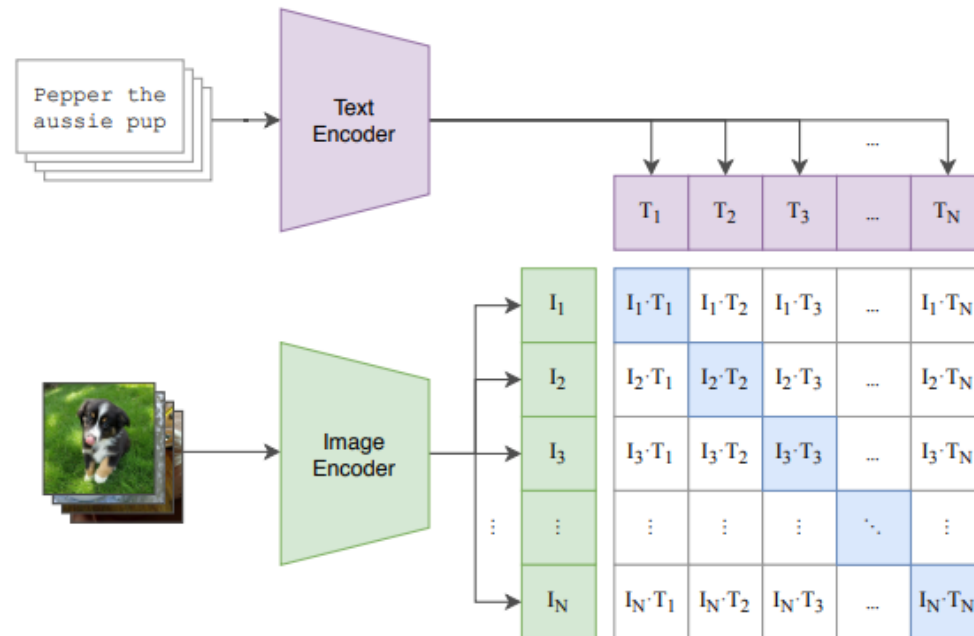
(1) Contrastive pre-training



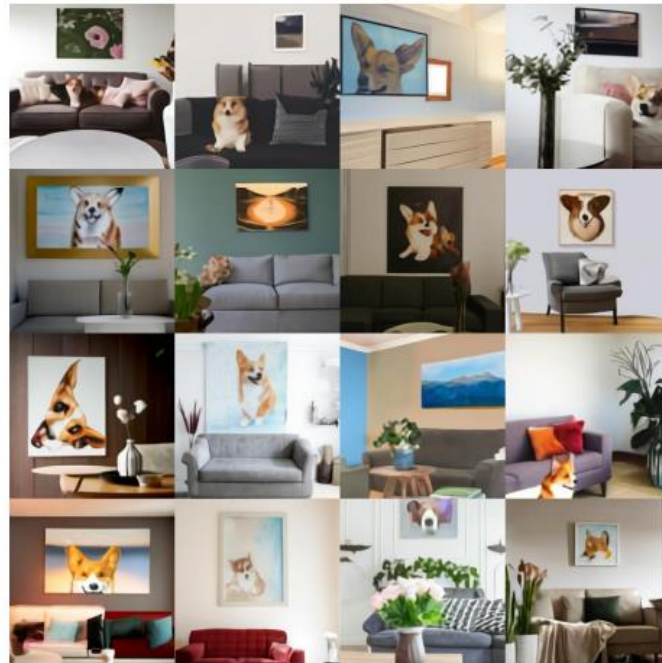
## CLIP (Radford2021, ICML, OpenAI)

- 먼저 image encoder 와 text encoder 를 학습시켰다. 1개의 batch 는  $N$ 개의 (image, text) 쌍으로 구성된다.  $N$ 개의 쌍을 모든  $i, j$ 에 대해 비교하면  $N$  개의 positive pair 와  $N^2 - N$  개의 negative pair 를 얻을 수 있다.
- Image 와 text 를 하나의 공통된 space 로 보낸 다음, **positive pair** 에서의 유사도는 최대화하고 negative pair 에서의 유사도는 최소화하도록 CE loss 를 사용하여 학습하였다.

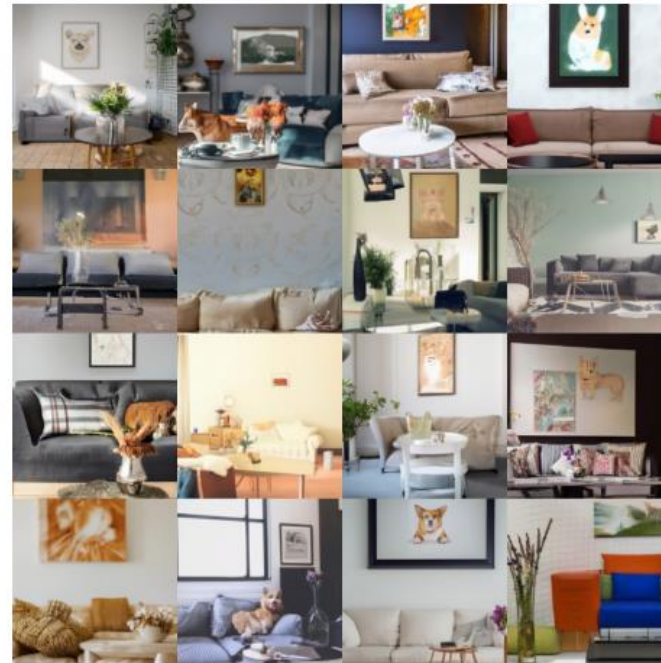
(1) Contrastive pre-training



- 기존 GAN 이 주축이던 text-to-image domain 에 **diffusion model** 을 도입해 사실적인 이미지를 생성했다.
- 실제 사람들에게 DALL-E 1 로 만든 이미지와 GLIDE 가 만든 이미지를 비교하는 설문을 실시했을 때, GLIDE 이미지의 photorealism 선호도가 87% 더 높았고, 캡션과의 유사도는 69% 높았다.



(a) DALL-E (Temp 0.85, CLIP reranked top 16 out of 512)



(b) GLIDE (Unguided)



- GLIDE 결과



“a hedgehog using a calculator”



“a corgi wearing a red bowtie and a purple party hat”



“robots meditating in a vipassana retreat”



“a fall landscape with a small cottage next to a lake”



“a surrealist dream-like oil painting by salvador dalí of a cat playing checkers”



“a professional photo of a sunset behind the grand canyon”



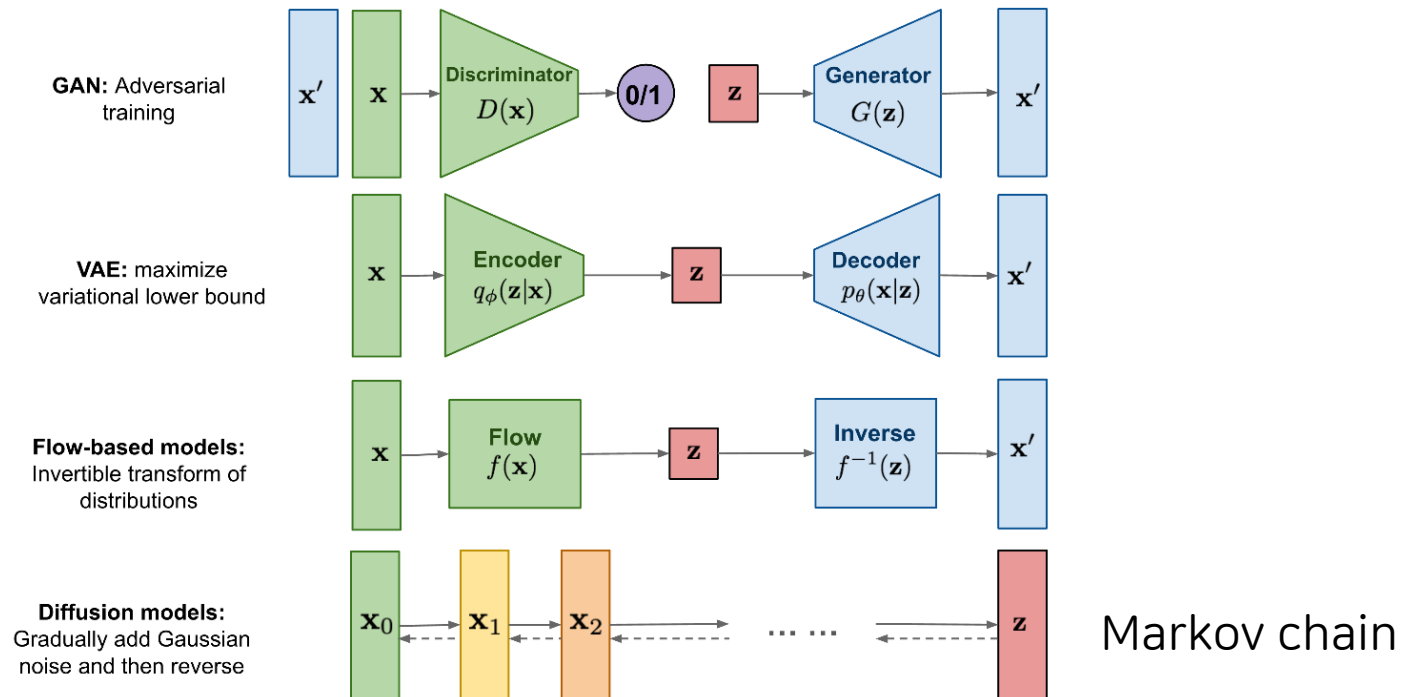
“a high-quality oil painting of a psychedelic hamster dragon”



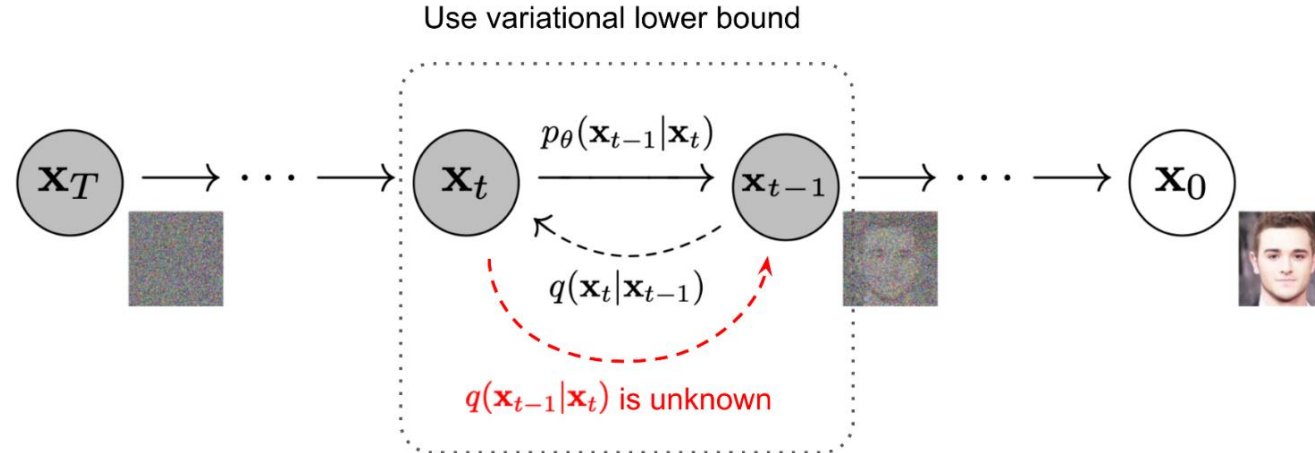
“an illustration of albert einstein wearing a superhero costume”

- Diffusion models

- Diffusion model 은 유명한 생성모델인 GAN 이나 VAE 와는 다르게 원본 이미지와 같은 차원을 가지는 고차원 latent variable 을 이용해 학습을 진행한다.



- Diffusion model : NSCN  $\rightarrow$  DDPM  $\rightarrow$  ADM  $\rightarrow$  GLIDE
  - Noise-Conditioned Score Network (NCSN)
  - Denoising Diffusion Probabilistic Models (DDPM)

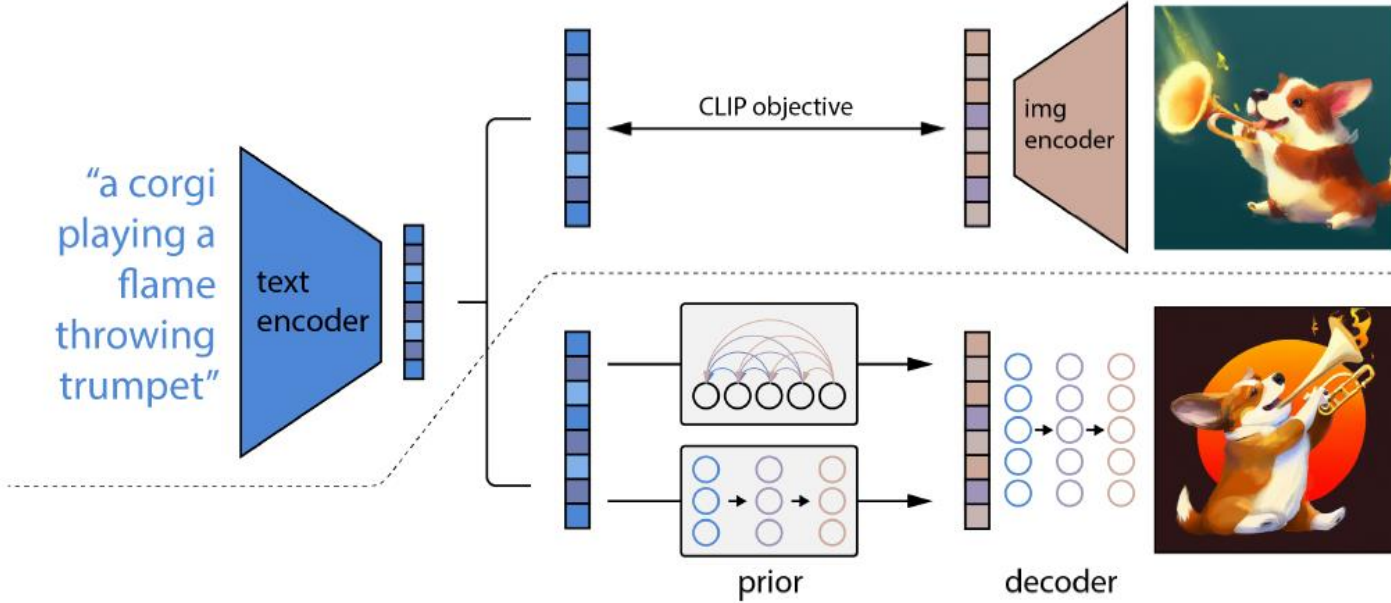


- Ablated Diffusion Model (ADM)
- Guided Language to Image Diffusion for Generation and Editing (GLIDE)

DALL-E 2 → Text to image generation



## Method

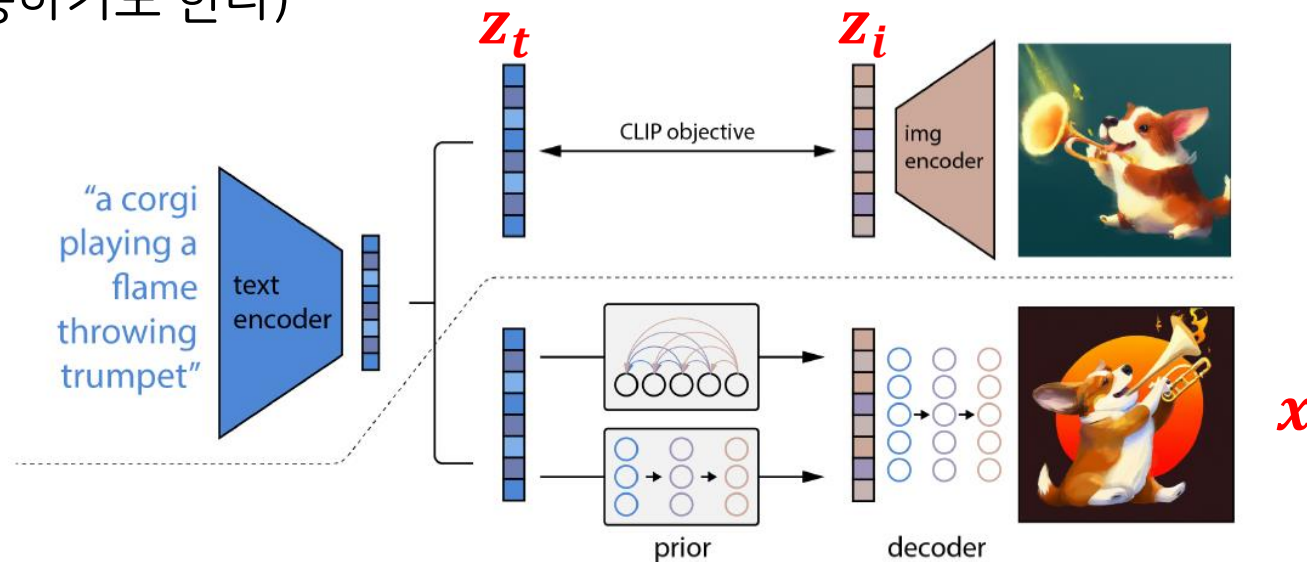


- [CLIP] CLIP 훈련 과정
  - 텍스트와 이미지 사이의 연관성을 학습한다.

- [unCLIP] Text-to-image generation 과정
  - CLIP text embedding 이 prior 에 들어가서 image embedding 을 생성한다.
  - 생성된 image embedding 이 decoder 에 들어가서 최종 결과 이미지를 생성한다.
  - CLIP model 은 prior 와 decoder 의 training 동안에 frozen 된다.

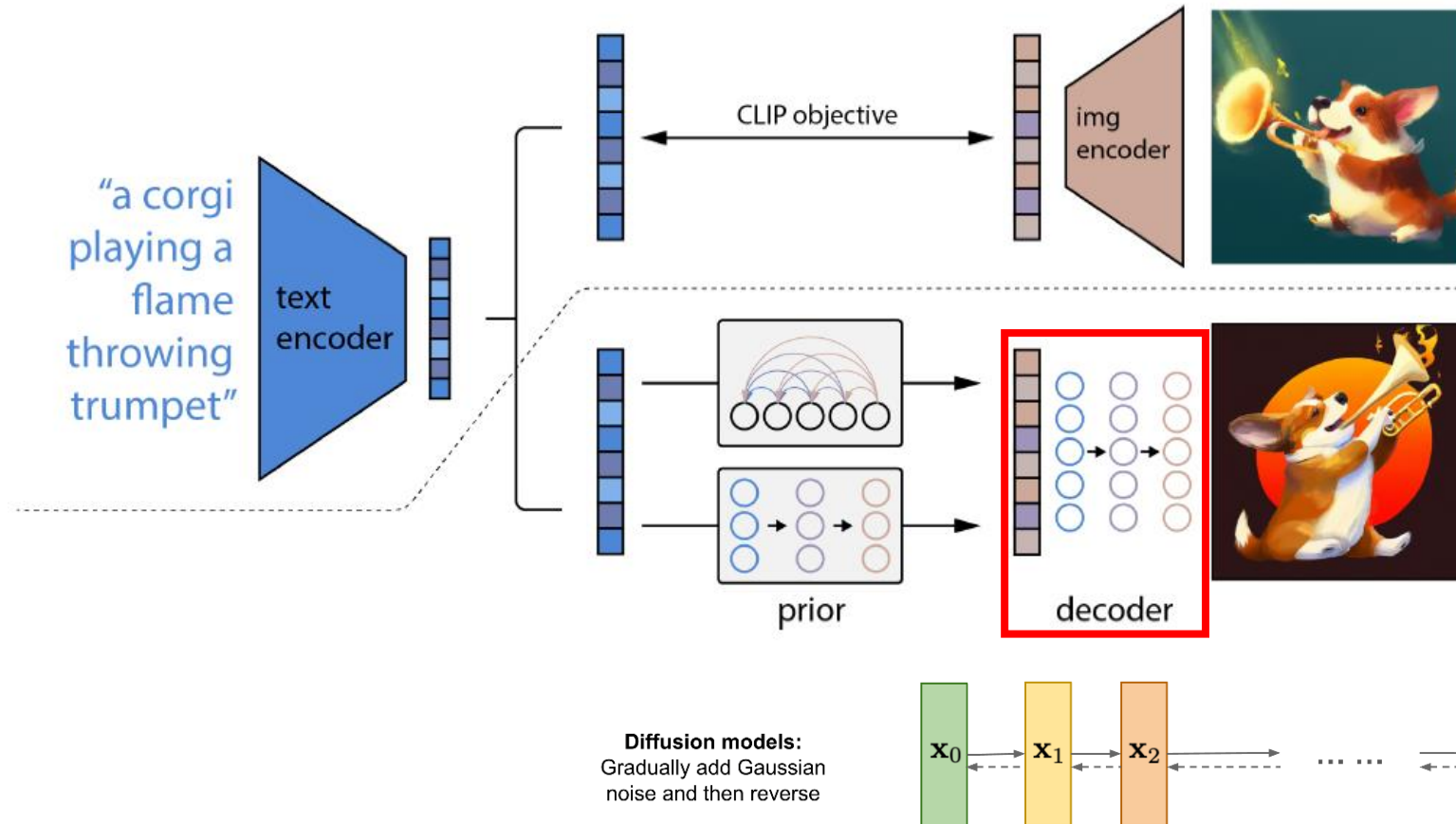
## Method

- Training dataset :  $(x, y)$  쌍으로 구성된다.  $x$  = images,  $y$  = captions
- $z_i$  : CLIP image embeddings
- $z_t$  : CLIP text embeddings
- *prior*  $P(z_i|y)$  : 캡션  $y$  가 주어졌을 때 CLIP image embeddings  $z_i$  를 생성한다.
- *decoder*  $P(x|z_i, y)$  : CLIP image embeddings  $z_i$  가 주어졌을 때 이미지  $x$  를 생성한다. (선택적으로 캡션  $y$  를 사용하기도 한다)

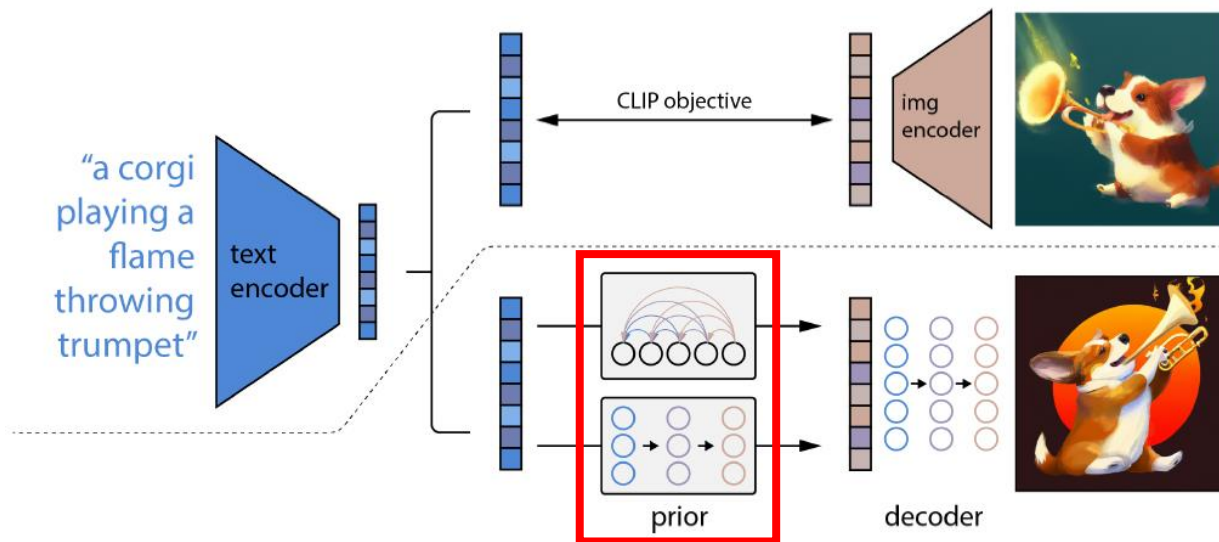


## Decoder

- DALL-E 2는 diffusion models 를 사용해서 CLIP image embeddings 로 이미지를 생성한다.  
Decoder 는 GLIDE 의 구조를 수정하였다.



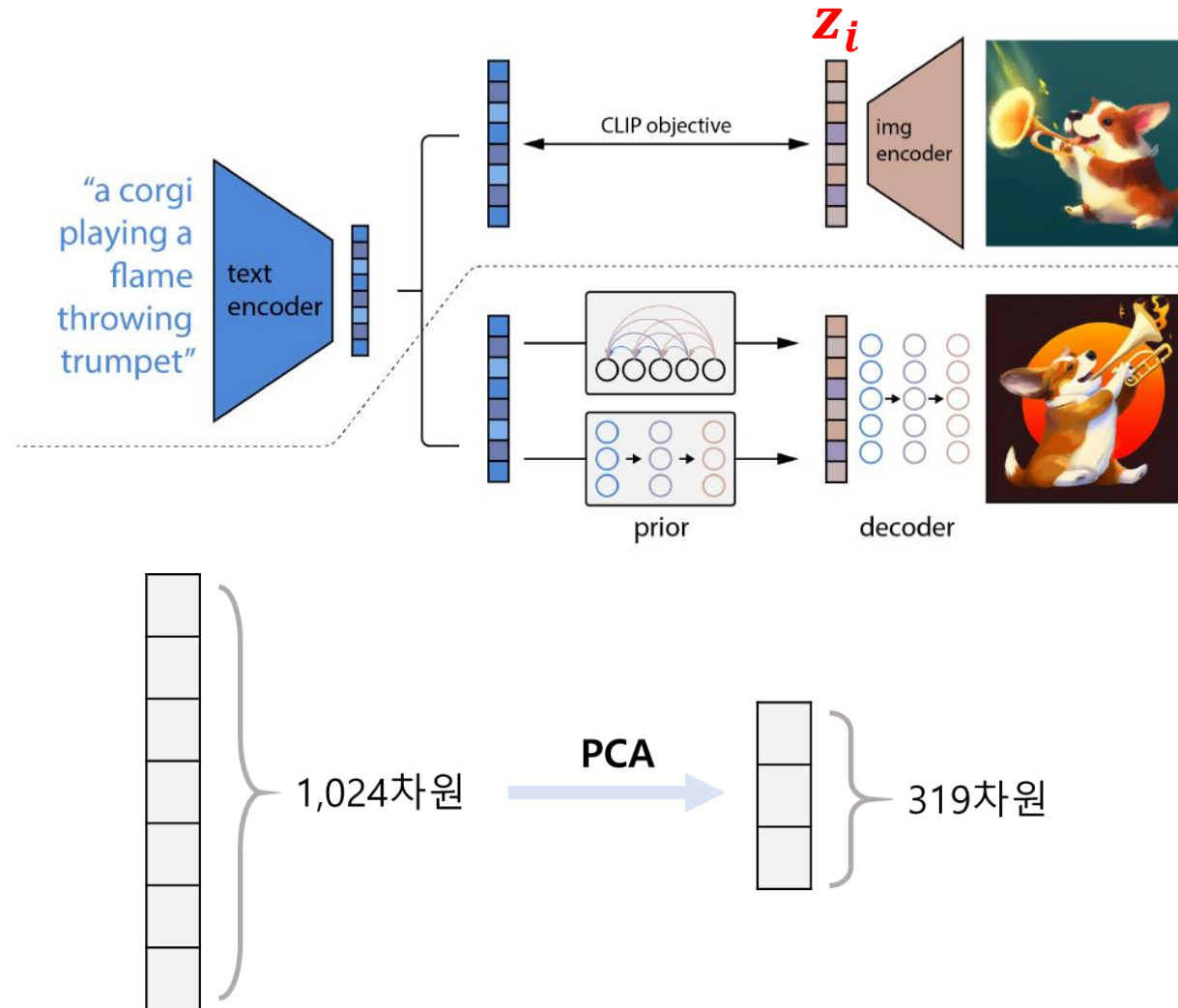
- 텍스트 캡션으로부터 이미지 생성을 하기 위해 캡션  $y$ 로부터  $z_i$ 를 생성하는 prior model 이 필요하다. 본 논문에서는 두 가지 prior 에 대해서 실험을 진행하였다.
- **Autoregressive (AR) prior:** CLIP image embedding  $z_i$ 는 이산적인 코드의 연속으로 변환되고 캡션  $y$ 에 대해서 autoregressively 하게 예측된다.
- **Diffusion prior:** 연속적인 벡터  $z_i$ 는 캡션  $y$ 에 대해서 Gaussian diffusion model 을 사용해서 바로 모델링 된다.





## Autoregressive(AR) Prior

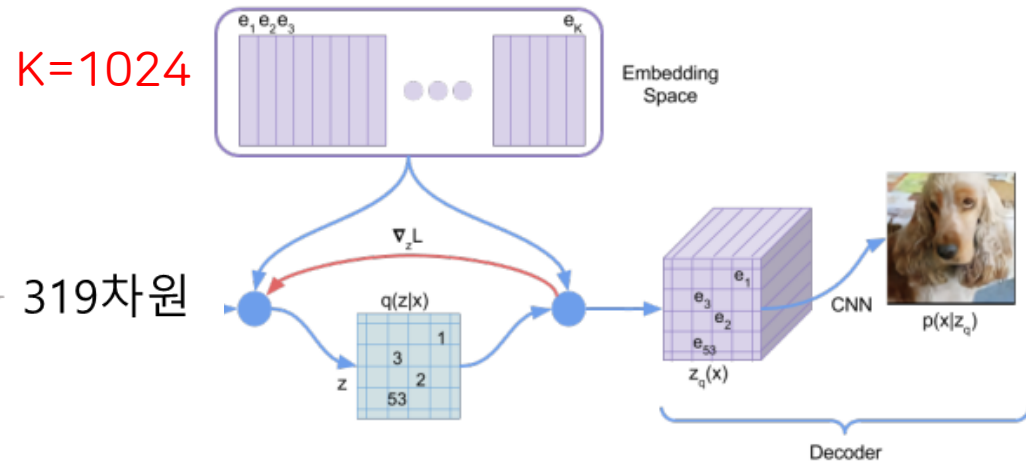
- 먼저 PCA 를 사용해서 CLIP image embeddings  $z_i$  의 차원을 감소시켰다. 1024 개 중 주요한 319 개의 요소만 유지하여 거의 모든 정보를 보존할 수 있었다.



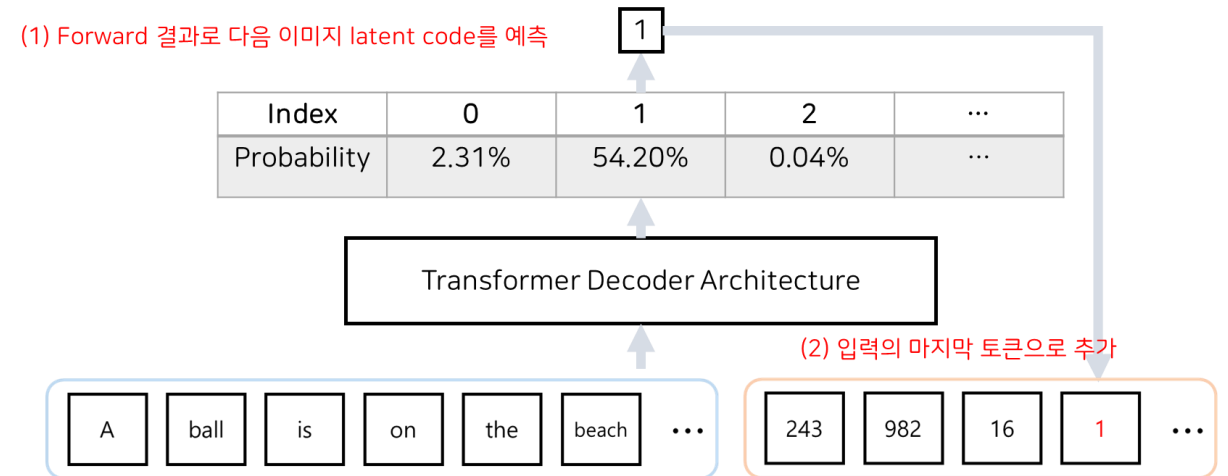
## Autoregressive(AR) Prior

- PCA 를 적용한 후에, DALL-E 1의 과정을 거쳤다.
- 이러한 과정 덕분에 inference 동안에 예측된 토큰들의 수를 3배 감소시키고, 훈련 안정성을 향상시킨다.

### [Codebook 학습]

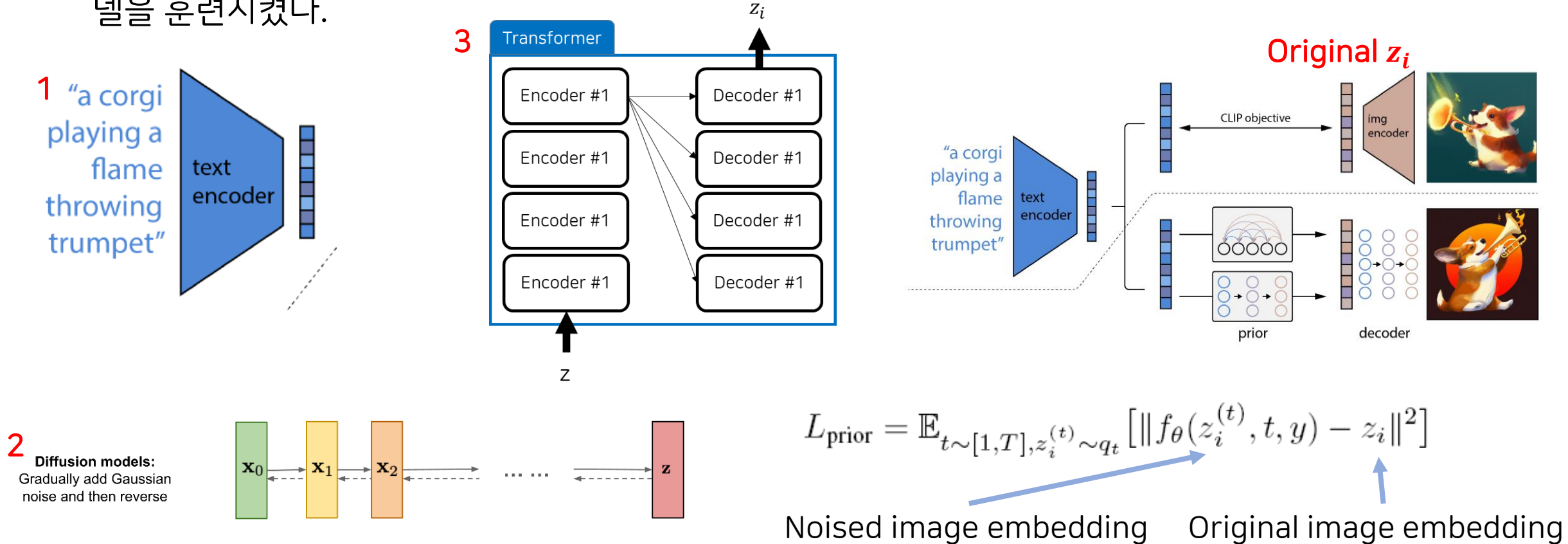


### [Transformer를 사용하여 image token 예측]



# Diffusion Prior

- Diffusion prior 과정: text  $\rightarrow$  **Text Encoder**  $\rightarrow$  CLIP text embedding  $\rightarrow$  **Diffusion model**  $\rightarrow$  embedding ( $z$ )  $\rightarrow$  **Transformer**  $\rightarrow$  unCLIP image embedding 예측
- unnoised  $z_i$  를 바로 예측하기 위해 DDPM 에서 사용한 mean-squared err loss 를 수정하여 모델 을 훈련시켰다.

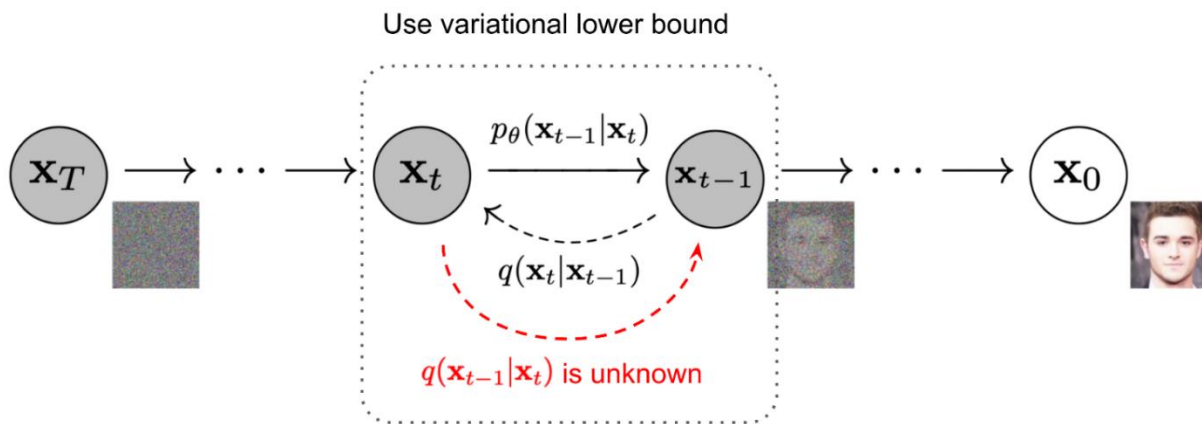
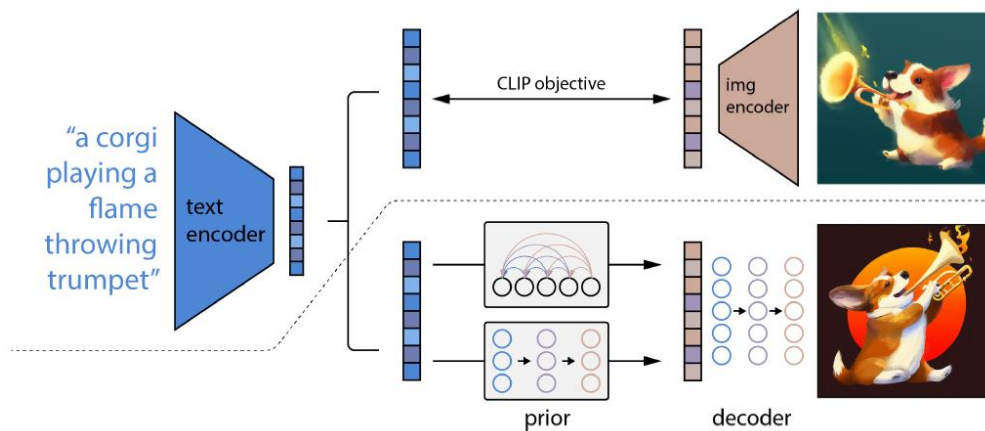


DALL-E 2 → Image manipulation



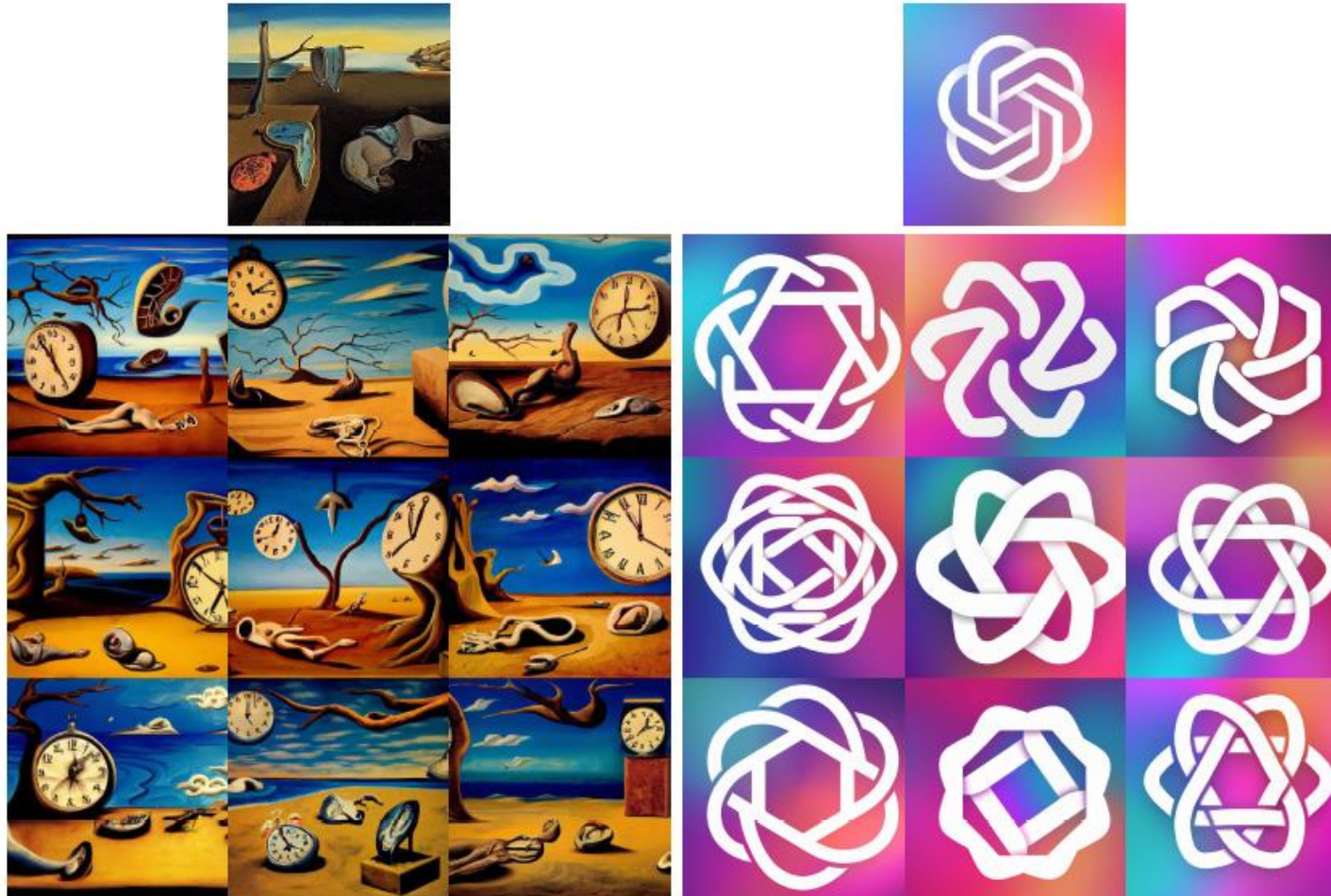
## Image Manipulations

- 앞에서 보았던 unCLIP 구조를 통해 이미지  $x$  를 bipartite latent representation  $(z_i, x_T)$  으로 encode 할 수 있다.
- $z_i$  : CLIP 에 의해 인식하는 이미지에 대한 정보이다. CLIP image encoder 로 간단히 얻을 수 있다.
- $x_T$  : Decoder 가  $x$  를 reconstruct 하기 위해 필요한 모든 residual information 을 포함한다. DDIM inversion 으로 얻을 수 있다.
- $(z_i, x_T)$  로 세 가지 image manipulations 을 할 수 있다.



## Variations

- 이미지  $x$  가 주어지면, 동일한 essential content 를 공유하지만 모양, 객체, 색감 등의 변화를 준 이미지를 생성할 수 있다.





# Interpolations

- 두 개의 이미지를 blend 할 수 있다.



## Text Diffs

- Language-guided image manipulation( = Text Diffs)를 할 수 있다.



a photo of a cat → an anime drawing of a super saiyan cat, artstation



a photo of a victorian house → a photo of a modern house



a photo of an adult lion → a photo of lion cub



a photo of a landscape in winter → a photo of a landscape in fall