ROBOT INTELLIGENCE LAB

Generative Model

ADB, CFG, GLIDE, DALL-E2

Sungjoon Choi, Korea University

kakaobrain

ROBOT INTELLIGENCE LAB

ADM

"Diffusion Models Beat GANs on Image Synthesis," 2021

Ablated Diffusion Model

ROBOT INTELLIGENCE LAB

- This paper presents **Ablated Diffusion Model** (ADM):
  - Several improvements techniques over DDPM
  - Conditioned generation via classifier-guidance

Ablated Diffusion Model

ROBOT INTELLIGENCE LAB

- ADM improves existing DDPM via:
  - Non-constant variance:

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v)\log \tilde{\beta}_t)$$

where $v_\theta(x_t, t)$ is the output of the neural network,

β_t (upper bound) is a scheduled constant ($\beta_1 = 10^{-4}$ to $\beta_T = 0.02$)

,and $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t$ (lower bound), $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$

  - Fewer steps sampling using DDIM

Architecture Improvements

ROBOT INTELLIGENCE LAB

- Following architectural changes are considered:
  - Increasing depth versus width, holding model size relatively constant
  - Increasing the number of attention heads
  - Using attention at 32×32, 16×16, and 8×8 resolutions rather than only 16×16
  - Using the BigGAN residual block for upsampling and downsampling the activations
  - Rescaling residual connections with $\frac{1}{\sqrt{2}}$.

Classifier Guidance

ROBOT INTELLIGENCE LAB

- For conditioned generation,
  - an additional classifier $p_\phi(y \mid x_t, t)$ is trained and use gradients $\nabla_{x_t} \log p_\phi(y \mid x_t, t)$ to guide the diffusion sampling process towards an arbitrary class label y.
  - Note that the classifier $p_\phi(y \mid x_t, t)$ should be trained on noisy images x_t (how?)

Algorithm 1 Classifier guided diffusion sampling, given a diffusion model $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, and gradient scale s.

Input: class label y, gradient scale s
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for all t from T to 1 **do**
  $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
  $x_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$
end for
return x_0

Samples from ImageNet

ROBOT INTELLIGENCE LAB

Samples from LSUN (bedroom)

ROBOT INTELLIGENCE LAB

Samples from LSUN (horse)

ROBOT INTELLIGENCE LAB

Samples from LSUN (cat)

ROBOT INTELLIGENCE LAB

ROBOT INTELLIGENCE LAB

CFG

"Classifier-Free Diffusion Guidance," 2021

Classifier-Free Guidance

ROBOT INTELLIGENCE LAB

- Classifier guidance mixes a score estimate with a classifier gradient during sampling

$$x_{t-1} \leftarrow \mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y \mid x_t), \Sigma)$$ where s is gradient scale

- A downside classifier guidance is that it requires an additional classifier model and thus complicates the training pipeline.
- CFG only modifies the existing diffusion model by making the reverse process function approximator receives c an input (i.e., $e_\theta(z_\lambda, c)$).

$$\tilde{\epsilon}_\theta(z_\lambda, c) = (1 + w)\epsilon_\theta(z_\lambda, c) - w\epsilon_\theta(z_\lambda)$$

where c is the class label,$\epsilon_\theta(z_\lambda) = \epsilon_\theta(z_\lambda, c = 0)$, and w is the implied-classifier weights

Tradeoffs

- **FID**: sample variety
- **IS**: individual sample fidelity

Method	FID (↓)	IS (↑)
ADM [3]	2.07	-
CDM [6]	**1.48**	67.95
Ours, no guidance	1.80	53.71
Ours, with guidance		
$w = 0.1$	1.55	66.11
$w = 0.2$	2.04	78.91
$w = 0.3$	3.03	92.8
$w = 0.4$	4.30	106.2
$w = 0.5$	5.74	119.3
$w = 0.6$	7.19	131.1
$w = 0.7$	8.62	141.8
$w = 0.8$	10.08	151.6
$w = 0.9$	11.41	161
$w = 1.0$	12.6	170.1
$w = 2.0$	21.03	225.5
$w = 3.0$	24.83	250.4
$w = 4.0$	26.22	**260.2**

Figure 1: ImageNet 64x64 results

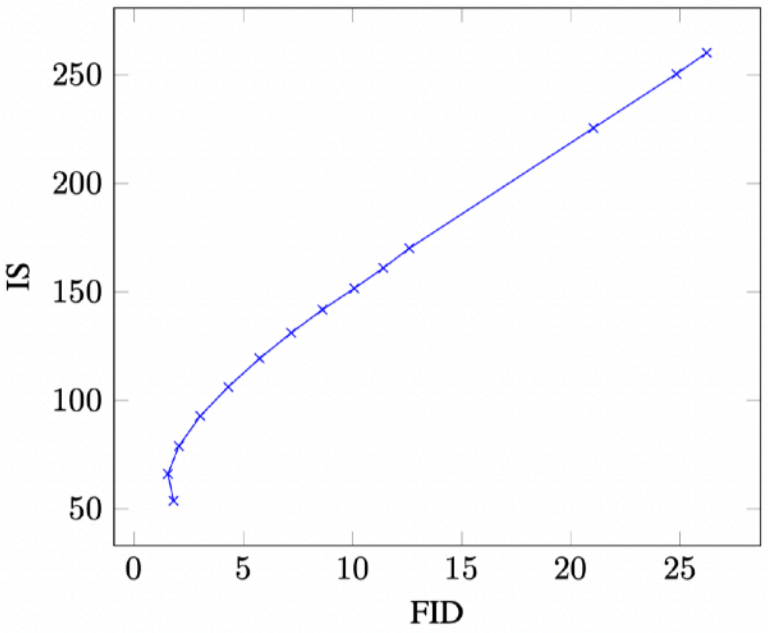
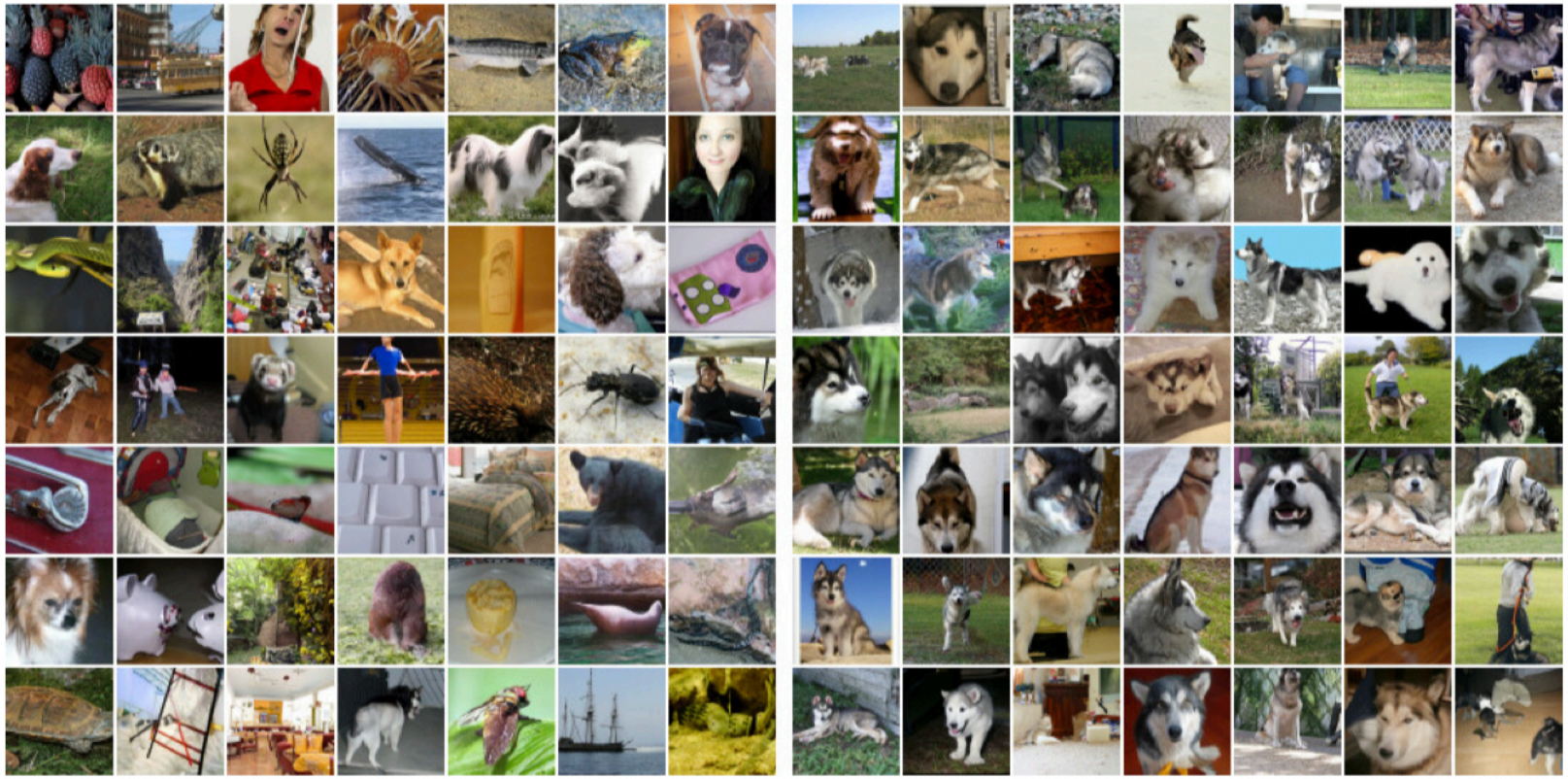


Figure 2: ImageNet 64x64 FID vs. IS

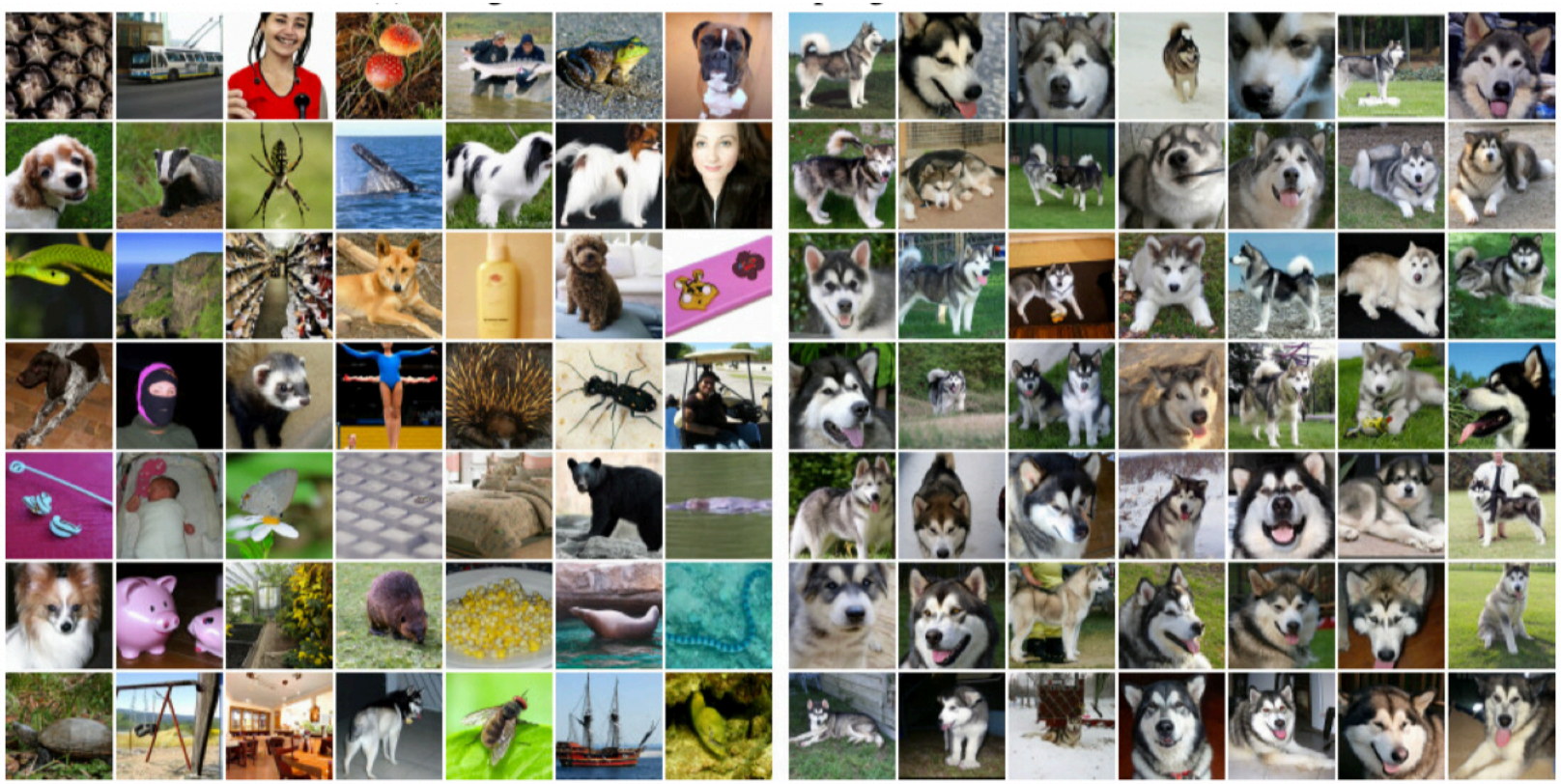
Non-guided



(a) Non-guided conditional sampling: FID=1.80, IS=53.71

- Low FID (the lower the better) -> High Variability
- Low IS (the higher the better) -> Low Image Quality

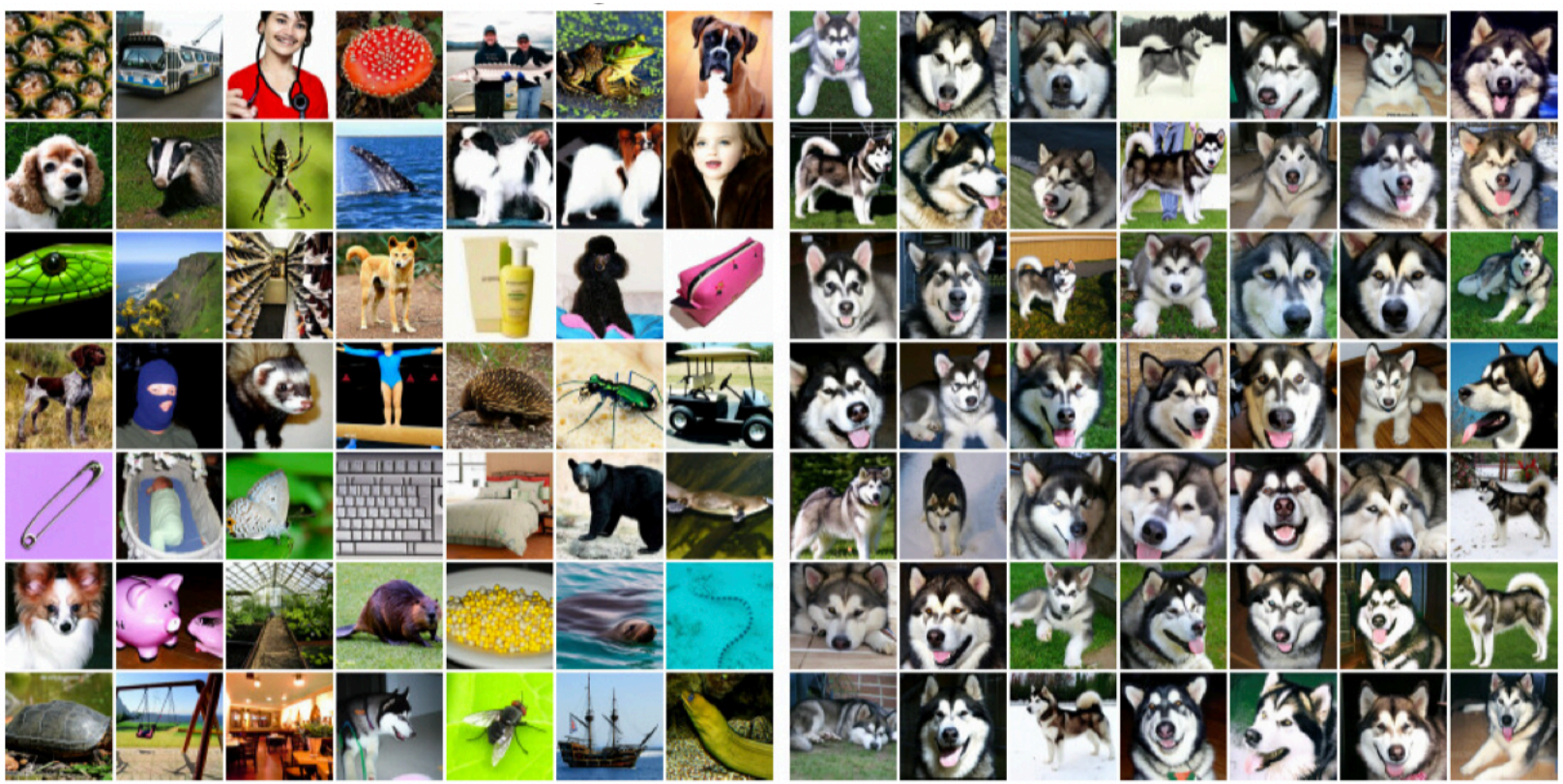
CFG with Small Guidance



(b) Classifier-free guidance with $w = 1.0$: FID=12.6, IS=170.1

- Middle FID (the lower the better) -> Middle Variability
- Middle IS (the higher the better) -> Middle Image Quality

CFG with High Guidance



(c) Classifier-free guidance with $w = 3.0$: FID=24.83, IS=250.4

- High FID (the lower the better) -> Low Variability
- High IS (the higher the better) -> High Image Quality

ROBOT INTELLIGENCE LAB

GLIDE

"GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," 2022

GLIDE

ROBOT INTELLIGENCE LAB

Guided Language to Image Diffusion for Generation and Editing (GLIDE)

Diffusion version of **DALL-E** (or precursor of **DALL-E2**)

GLIDE

ROBOT INTELLIGENCE LAB

Two different approaches are presented: CLIP guidance and classifier-free guidance
+
Editing capabilities (e.g., image inpainting)

Background

ROBOT INTELLIGENCE LAB

- Diffusion Models
  - Forward diffusion: $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)I)$
  - Backward model: $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$
  - Objective: generate samples $x_t \sim q(x_t \mid x_0)$ by applying Gaussian noise ϵ to x_0, then train a model ϵ_θ to predict the added noise using a standard mean-squared error loss

Background

ROBOT INTELLIGENCE LAB

- Guided Diffusion
  - Perturbed mean: $\hat{\mu}_\theta(x_t \mid y) = \mu_\theta(x_t \mid y) + s \cdot \Sigma_\theta(x_t \mid y) \nabla_{x_t} \log p_\phi(y \mid x_t)$
  - Note that $p_\phi(y \mid x_t)$ is a classifier trained with noisy images and s is the guidance scale
- Classifier-free Guidance
  - Class-conditional diffusion model: $\epsilon_\theta(x_t \mid y)$
  - Extrapolated error prediction: $\hat{\epsilon}_\theta(x_t \mid y) = \epsilon_\theta(x_t \mid \varnothing) + s(\epsilon_\theta(x_t \mid y) - \epsilon_\theta(x_t \mid \varnothing))$ where $s \geq 1$ is the guidance scale
- CLIP Guidance
  - CLIP: $f(x) \cdot g(c)$ is high for matching image x and caption c
  - Perturbed mean: $\hat{\mu}_\theta(x_t \mid c) = \mu_\theta(x_t \mid c) + s \cdot \Sigma_\theta(x_t \mid y) \nabla_{x_t}(f(x_t) \cdot g(c))$
  - Note that the CLIP $f(x_t) \cdot g(c)$ should be trained on noisy images

Text-Conditional Diffusion Models

- ADM model with text conditioning information (CLIP guidance and classifier-free guidance)
  - Visual part: ImageNet 64×64 model of ADM with 512 channels -> 2.3B parameters
  - Text part: Transformer encoder with 24 residual blocks of width 2048 -> 1.2B parameters
  - Upsampler: 64×64 resolution to 256×256 resolution -> 1.5B parameters
- Trained on the same dataset as DALL-E.. ~~(not public)~~

Image Inpainting

ROBOT INTELLIGENCE LAB

- Inpainting with a diffusion model can be done by replacing the known region of the image with a sample from $q(x_t|x_0)$ after each sampling step.
  - It is **NOT** replacing the known region of the image directly.
  - But what about the unknown region of x_0?
  - It has the disadvantage that the model cannot see that entire context but only a noised version sampled from $q(x_t|x_0)$.
- Glide explicitly **fine-tune** the model to preform inpainting.
  - During fine-tuning, random regions of training images are erased, and the remaining portions are fed into to model with a mask channel as additional conditioning information.
  - The input of the model has four additional channels: a second set of RGB channels and a mask channel.

Results

ROBOT INTELLIGENCE LAB

"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

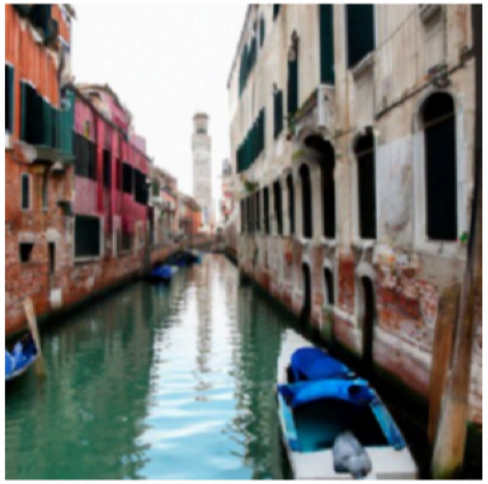
"a professional photo of a sunset behind the grand

"a high-quality oil painting of a psychedelic hamster

"an illustration of albert einstein wearing a superhero

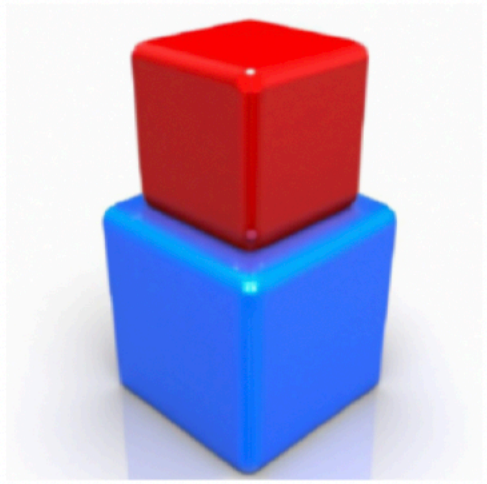
Results

ROBOT INTELLIGENCE LAB

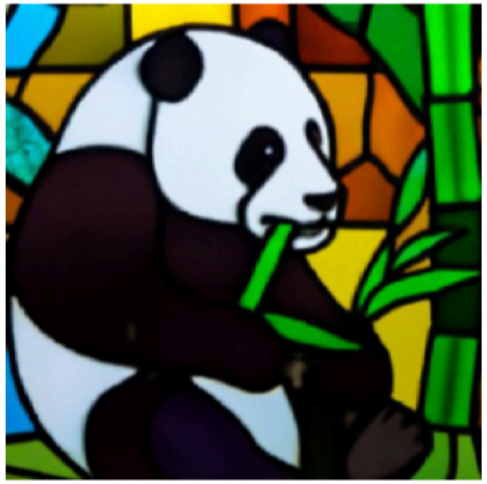


"a boat in the canals of venice"

"a painting of a fox in the style of starry night"



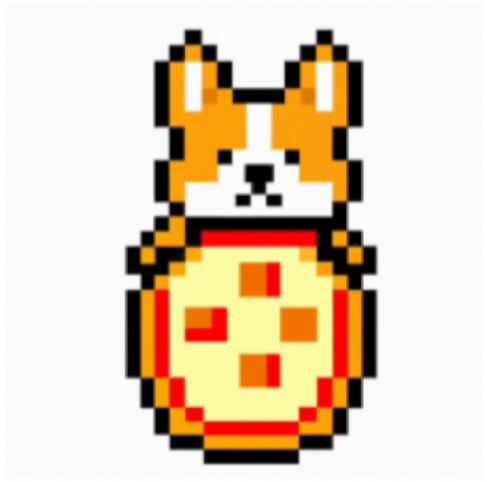
"a red cube on top of a blue cube"



"a stained glass window of a panda eating bamboo"

"a crayon drawing of a space elevator"

"a futuristic city in synthwave style"



"a pixel art corgi pizza"

"a fog rolling into new york"

Results

ROBOT INTELLIGENCE LAB

"a cozy living room"

"a painting of a corgi on the wall above a couch"

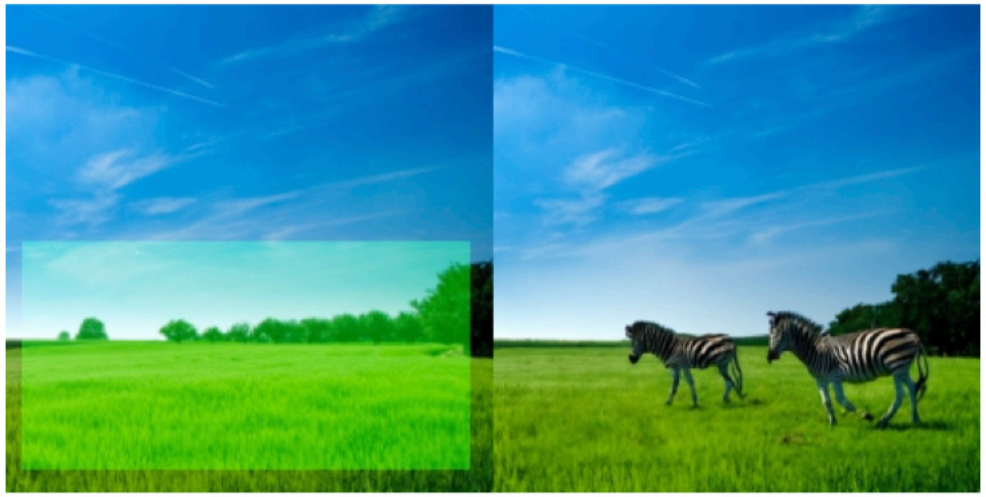
"a round coffee table in front of a couch"

"a vase of flowers on a coffee table"

"a couch in the corner of a room"

Results

ROBOT INTELLIGENCE LAB



"zebras roaming in the field"

"a girl hugging a corgi on a pedestal"

"a man with red hair"

"a vase of flowers"

ROBOT INTELLIGENCE LAB

DALL-E2

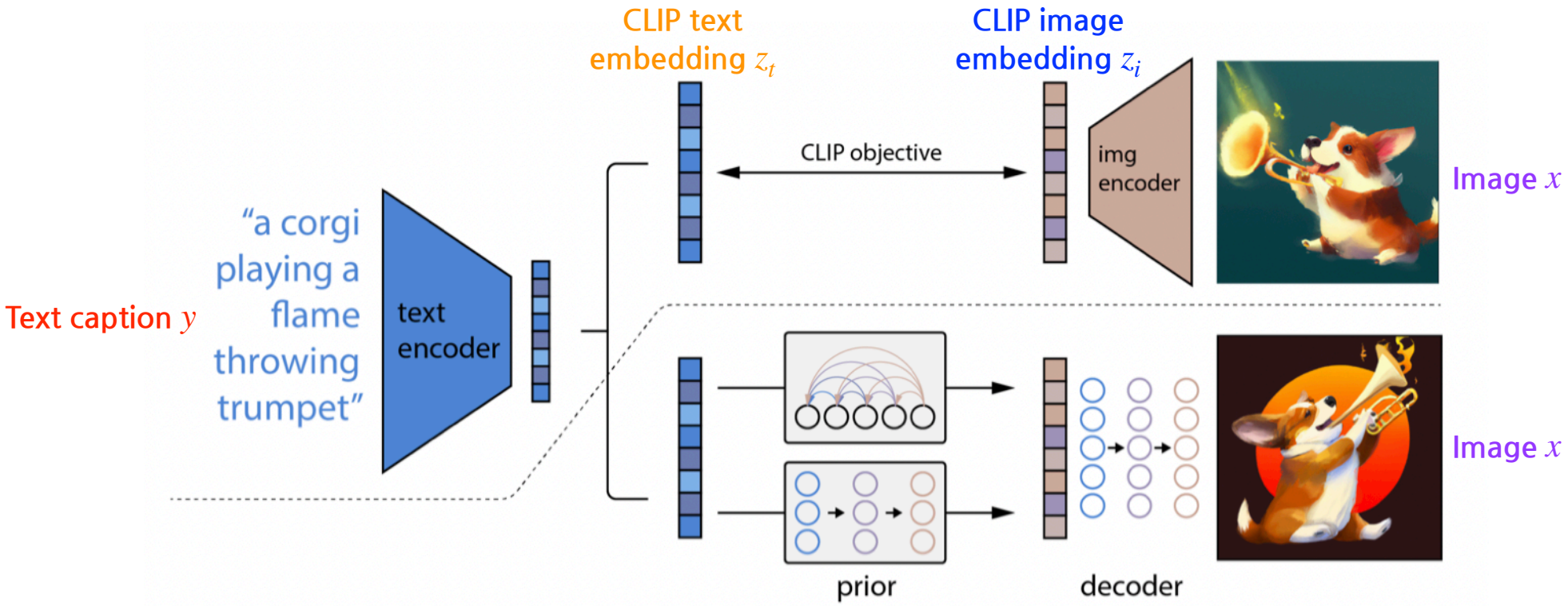
"Hierarchical Text-Conditional Image Generation with CLIP Latents," 2022

unCLIP

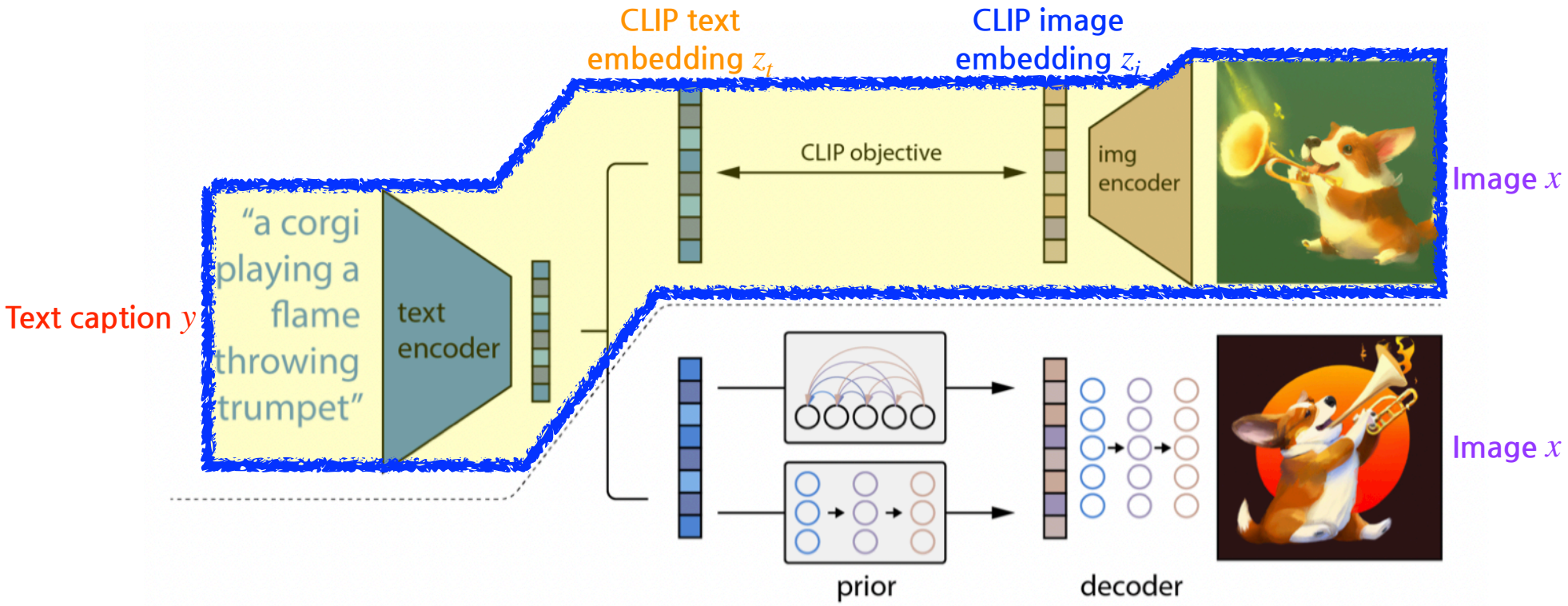
ROBOT INTELLIGENCE LAB

This is the famous **DALL-E 2** paper
(the name of the proposed text-to-image stack is **unCLIP**)

unCLIP

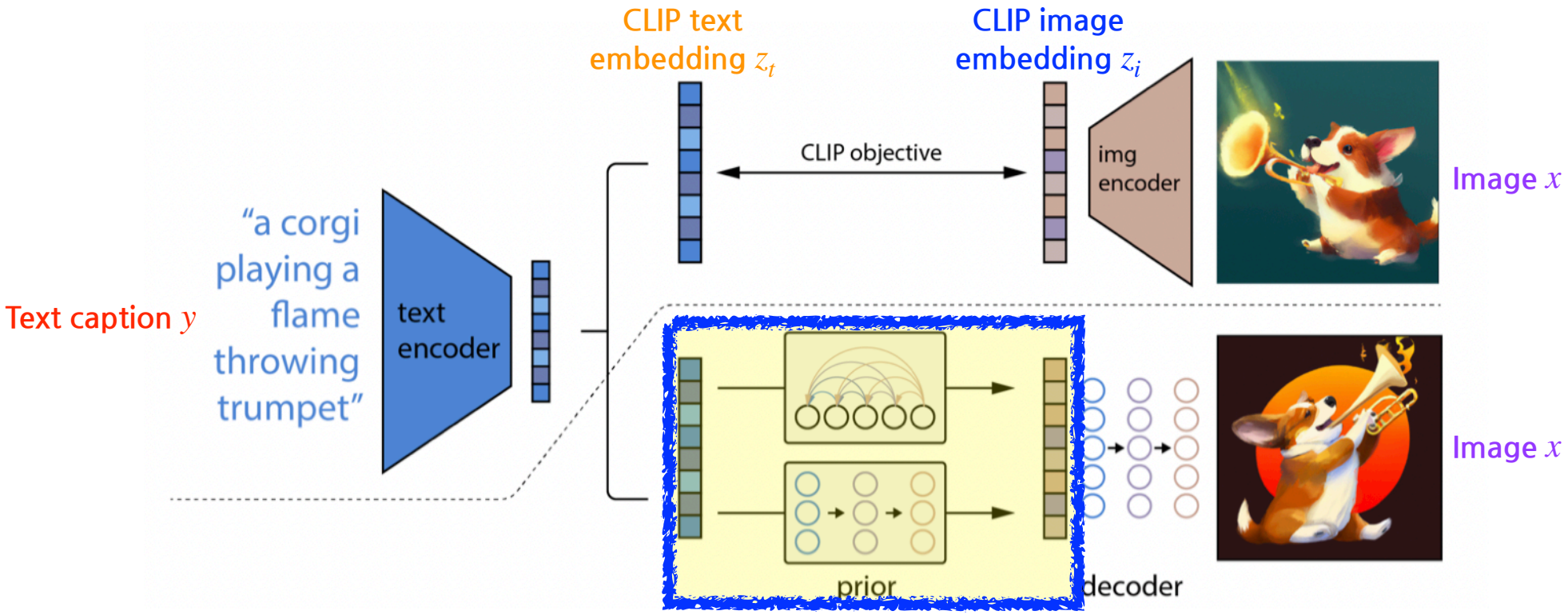

CLIP text embedding z_t
CLIP image embedding z_i
CLIP objective
img encoder
Image x
Text caption y
"a corgi playing a flame throwing trumpet"
text encoder
prior
decoder
Image x

unCLIP



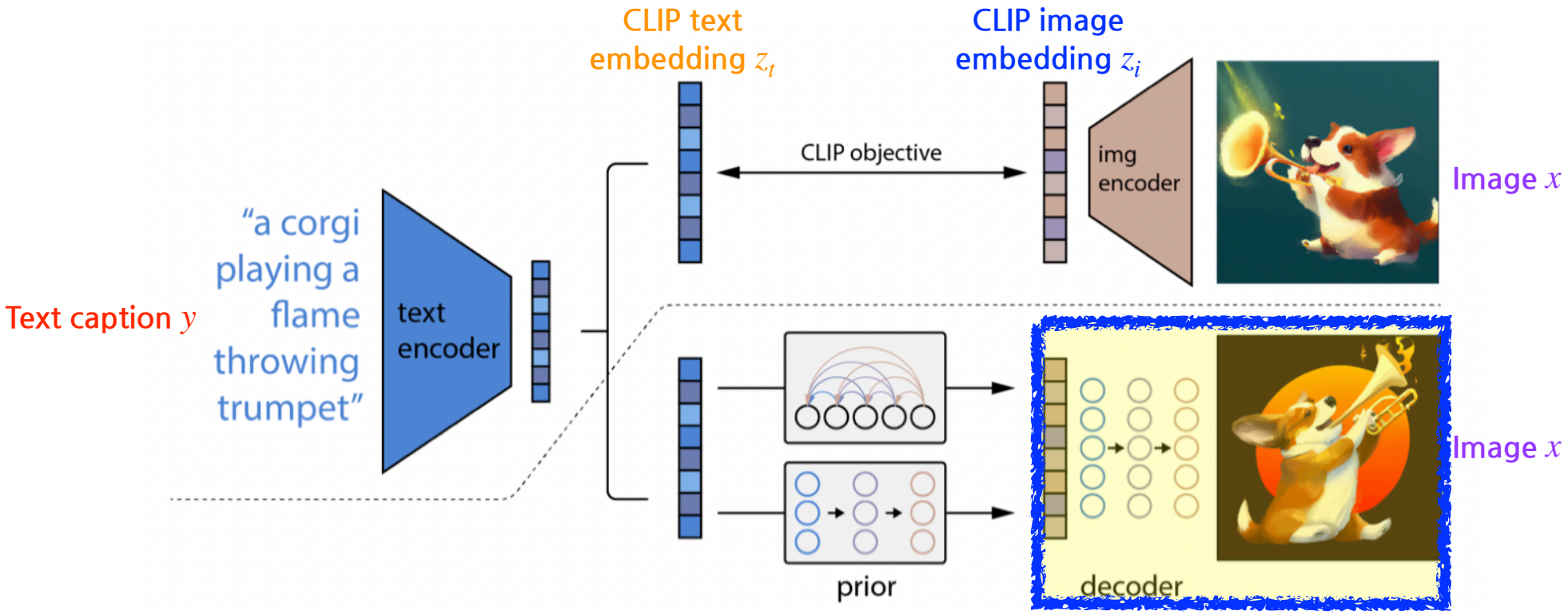
This part is the same as the original CLIP

unCLIP



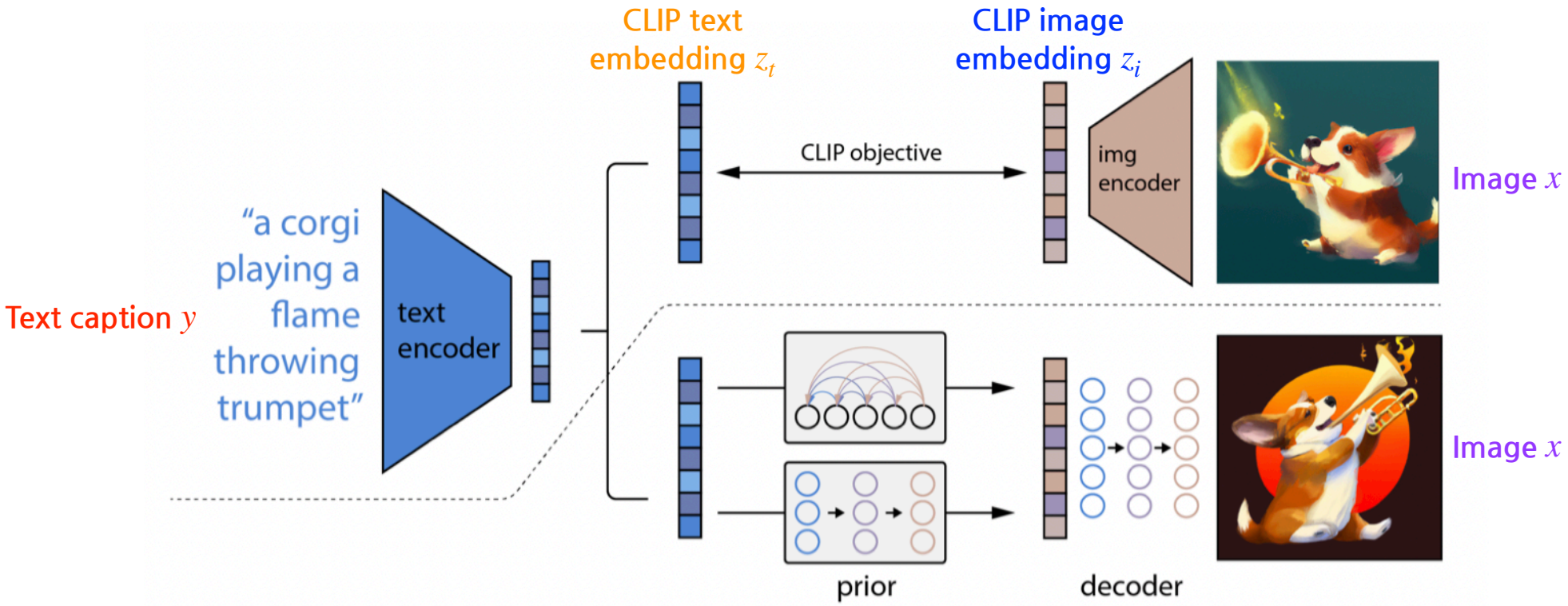
Prior: CLIP text embedding z_t -> CLIP image embedding z_i

unCLIP



Decoder: CLIP image embedding z_i-> Image x

unCLIP



- **unCLIP**: Text encoder: text caption => text embedding; Prior: => image embedding; Decoder: => image

Results

ROBOT INTELLIGENCE LAB

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

Interpolation

ROBOT INTELLIGENCE LAB

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

Failures

Figure 16: Samples from unCLIP for the prompt, “A sign that says deep learning.”

Failures

ROBOT INTELLIGENCE LAB

(a) A high quality photo of a dog playing in a green field next to a lake.

(b) A high quality photo of Times Square.

Figure 17: unCLIP samples show low levels of detail for some complex scenes.

Thank You