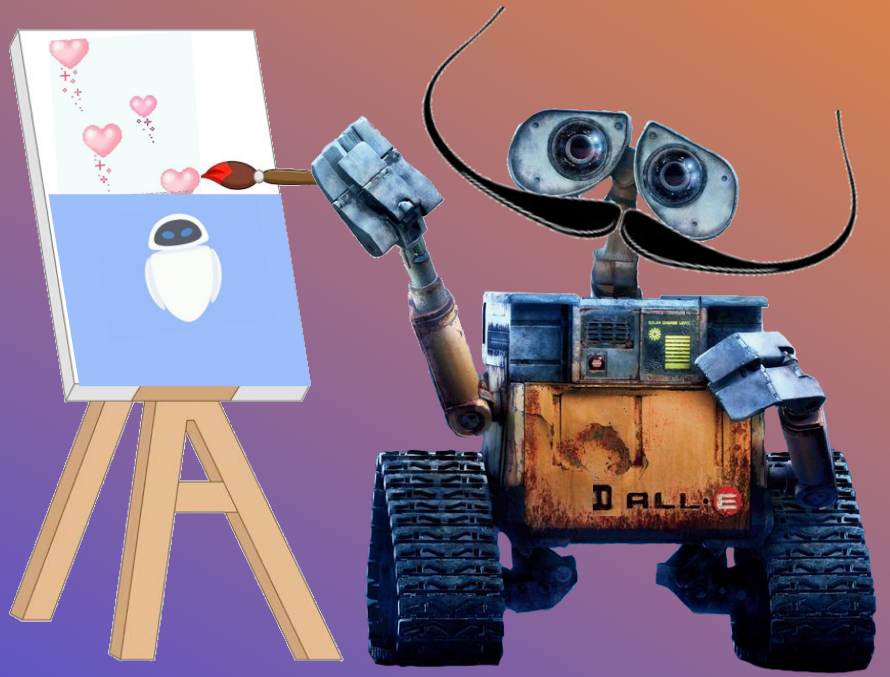
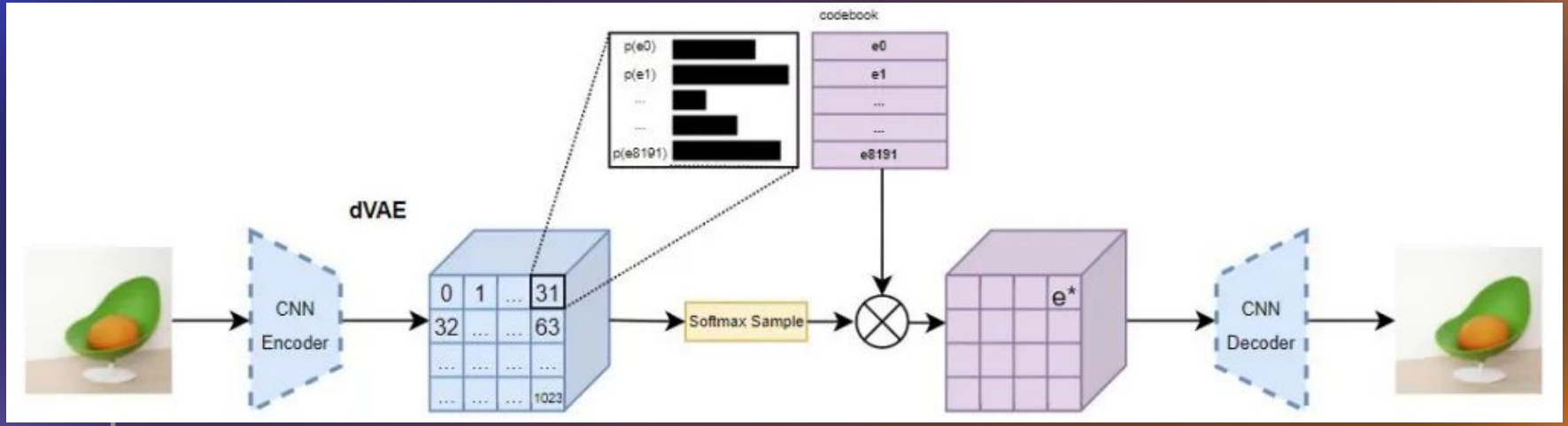
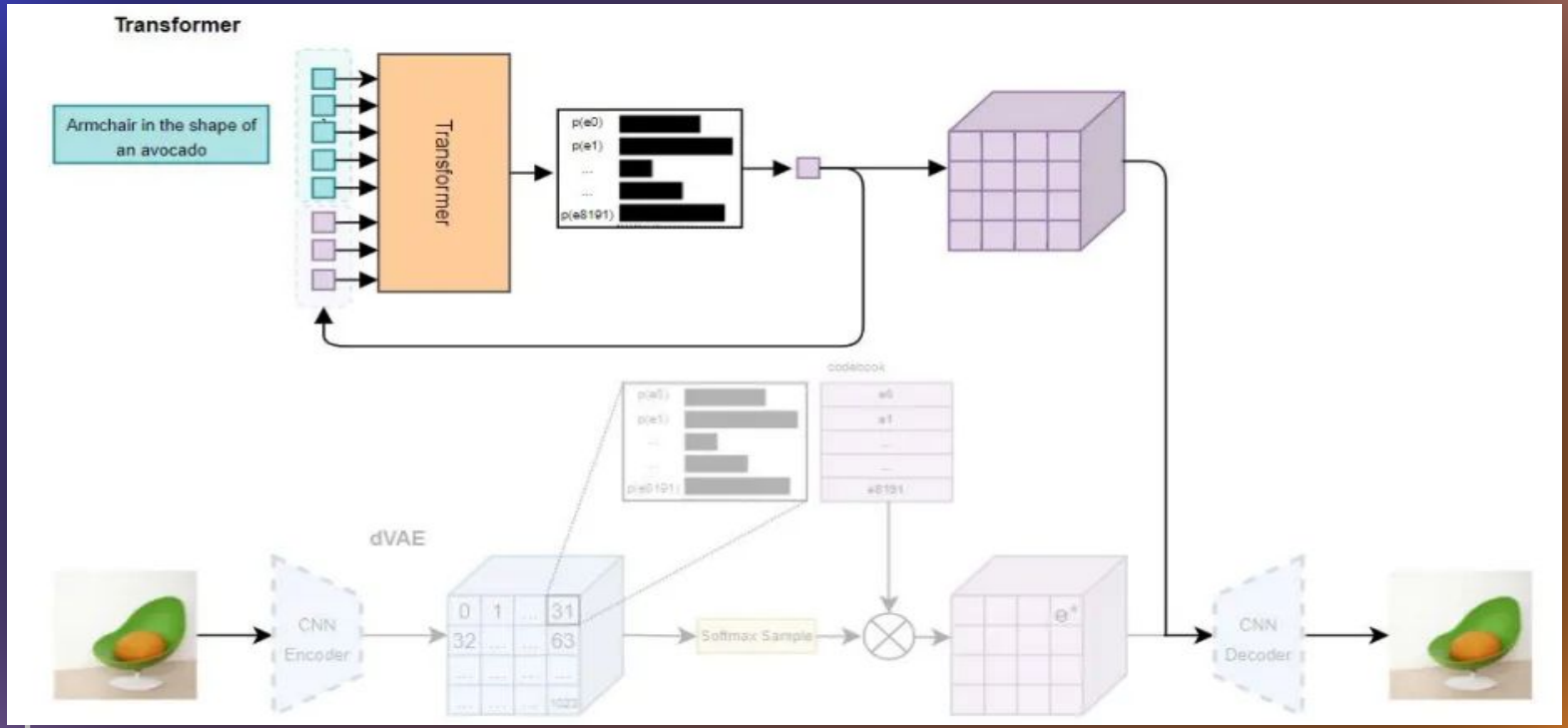
DALL-E 2



Recap: DALL-E 1

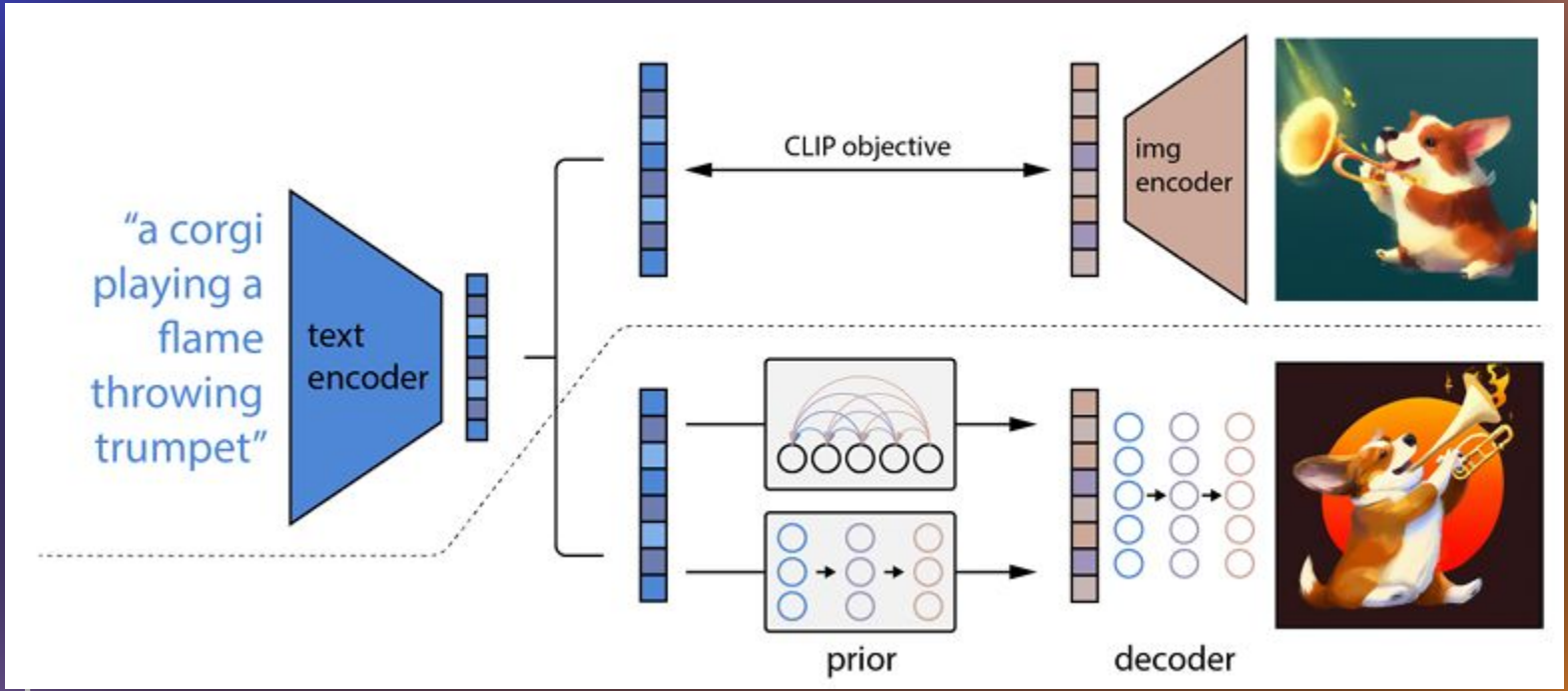


Recap: DALL-E 1



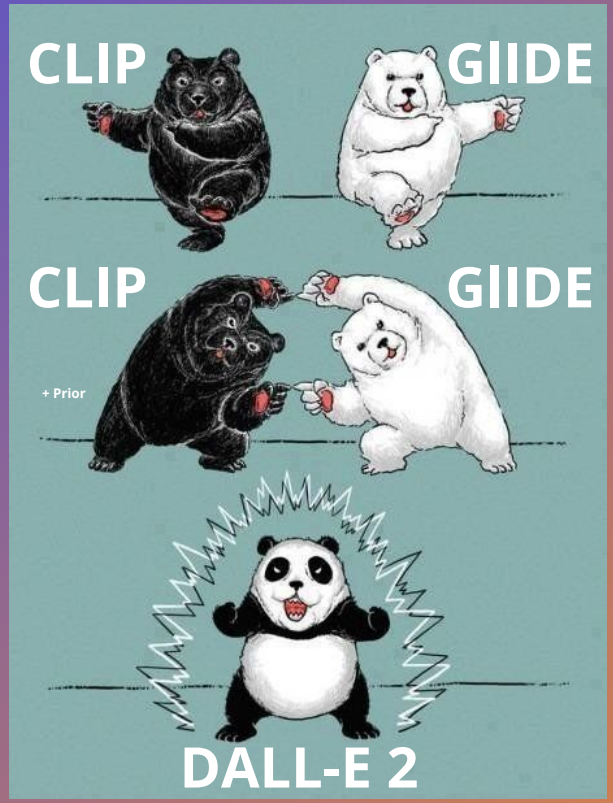
DALL-E 1 Examples

DALL-E 2 Architecture



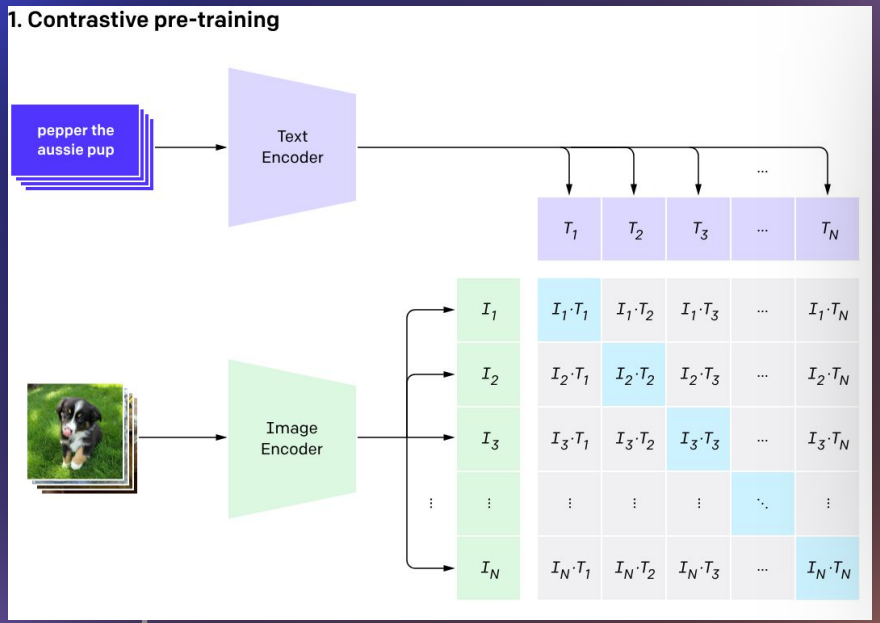
DALL-E 2 Architecture

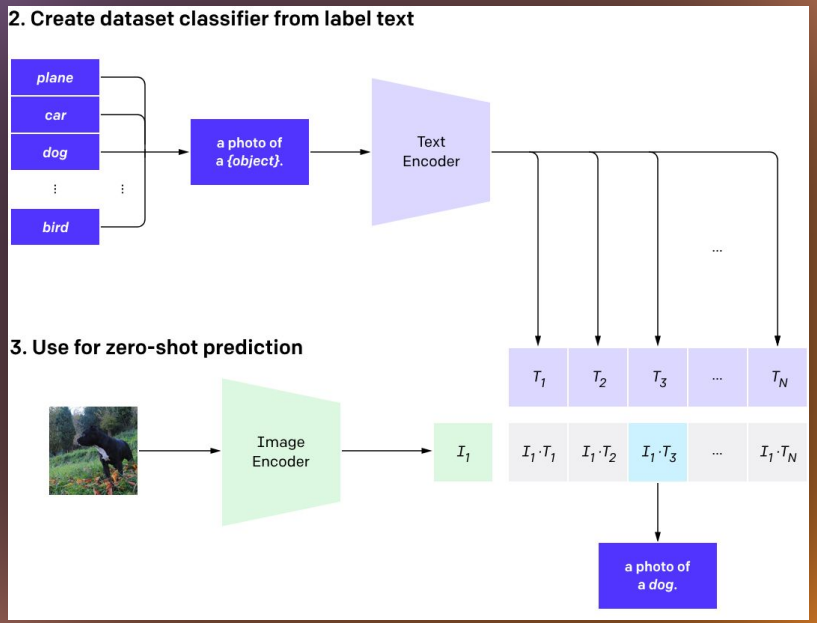
Image + Text Latent Space



Diffusion-based Image generation

CLIP Embeddings





CLIP Example

A human in a banana costume

Text Encoder

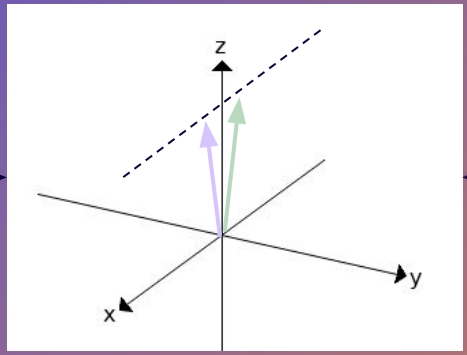


Image Encoder

A human in a banana costume

Text Encoder

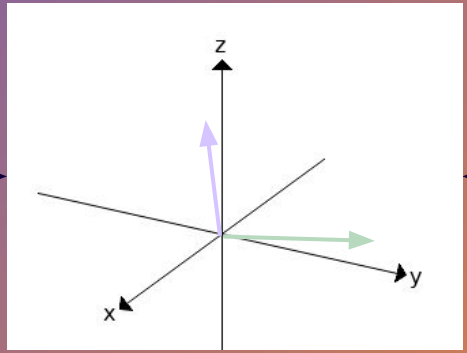


Image Encoder

unCLIP

A human in a banana costume

Text Encoder

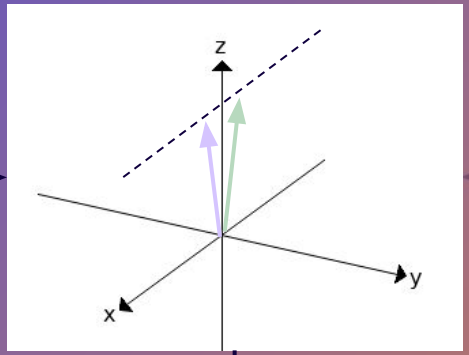


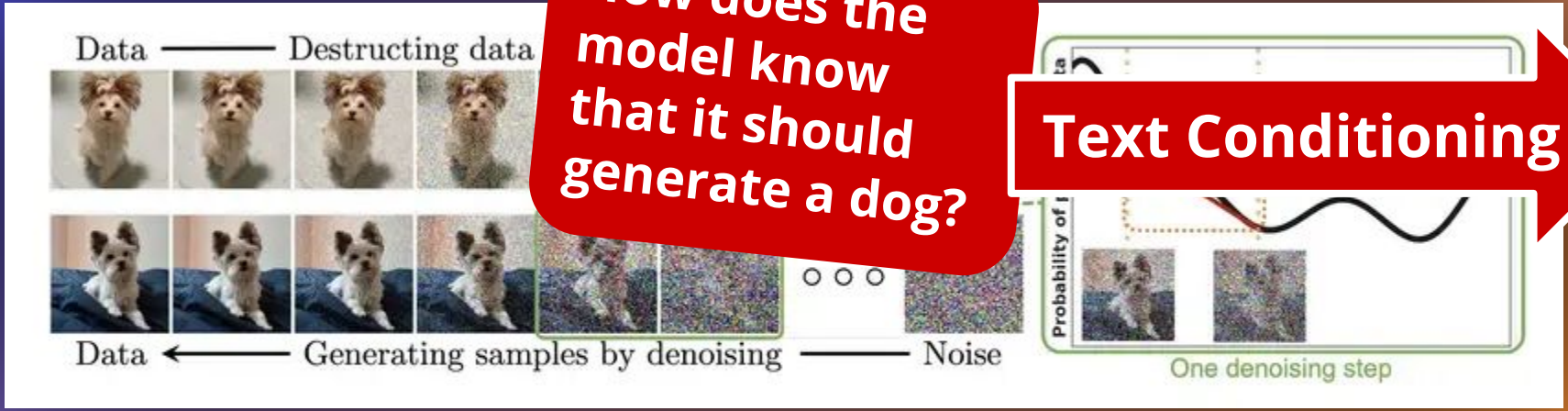
Image Encoder

Image Decoder

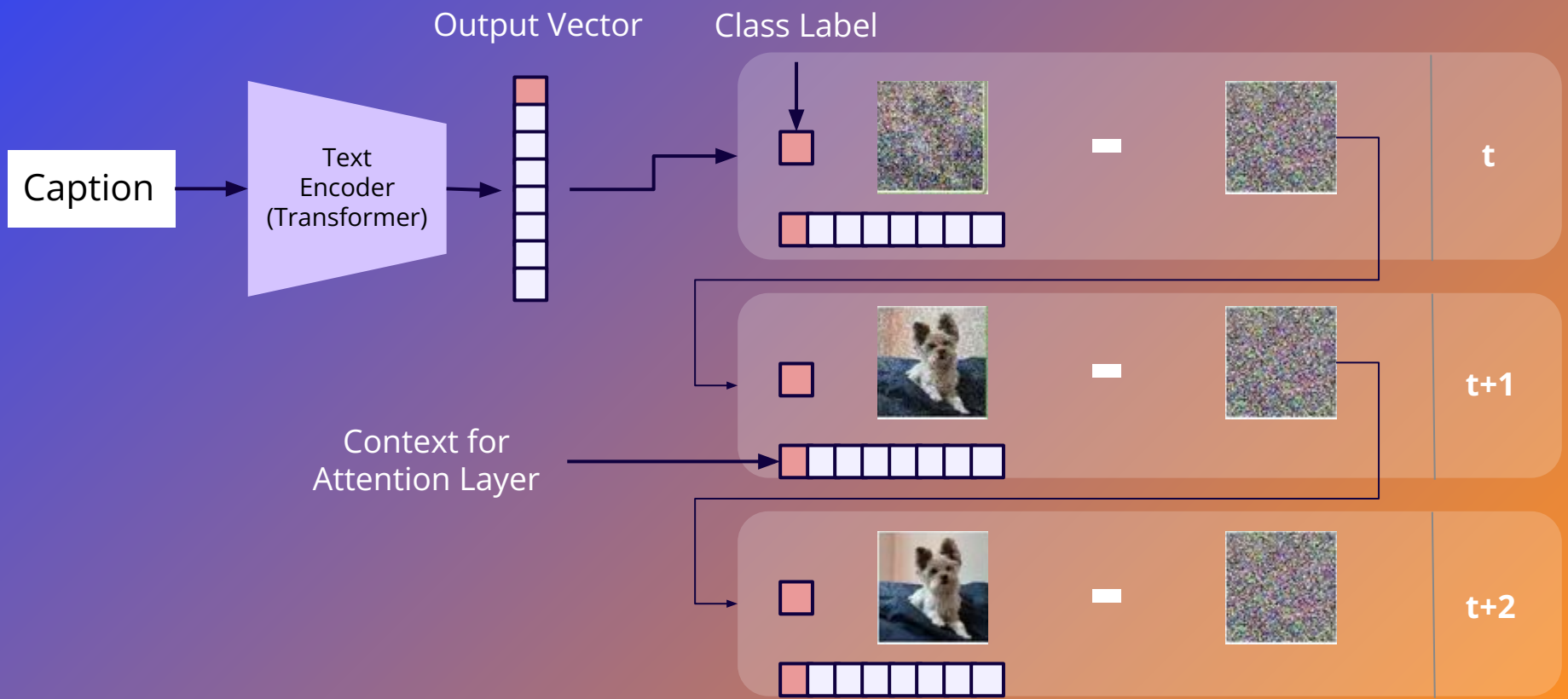
GLIDE

- Image generation from Text input
- Can also work with CLIP guidance
- Diffusion-based

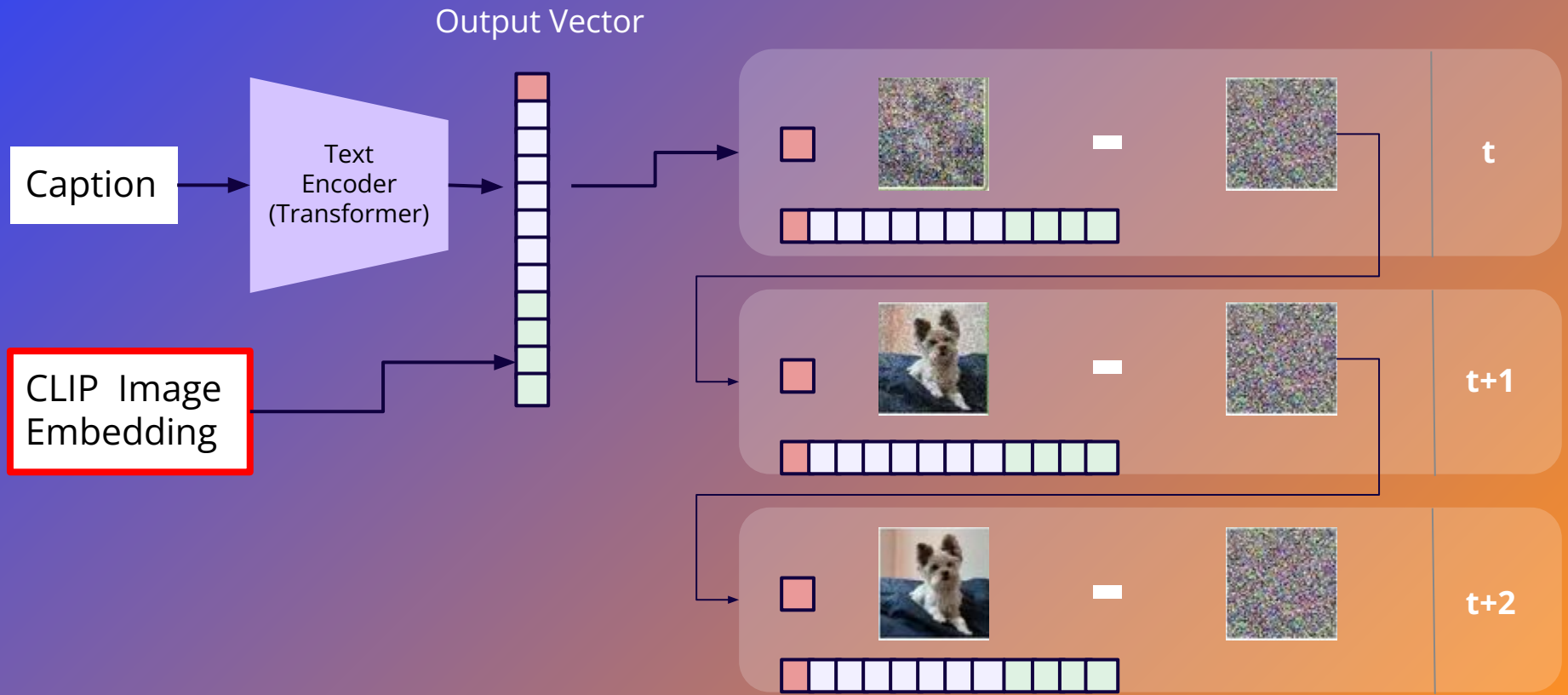
Diffusion-based Image generation



GLIDE


Output Vector
Class Label
Caption
Text
Encoder
(Transformer)
t
t+1
t+2
Context for
Attention Layer

GLIDE for DALL-E 2



Upsampling

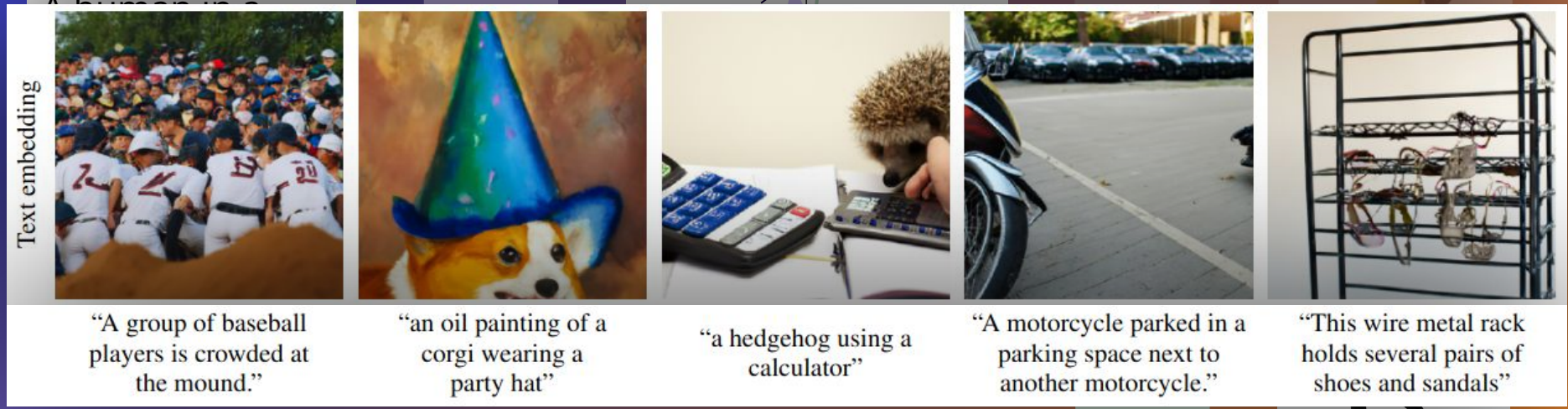
64x64

256x256

1024x1024

Without any caption conditioning or guidance

unCLIP


Text embedding
"A group of baseball players is crowded at the mound."
"an oil painting of a corgi wearing a party hat"
"a hedgehog using a calculator"
"A motorcycle parked in a parking space next to another motorcycle."
"This wire metal rack holds several pairs of shoes and sandals"

unCLIP

A human in a banana costume

Text Encoder

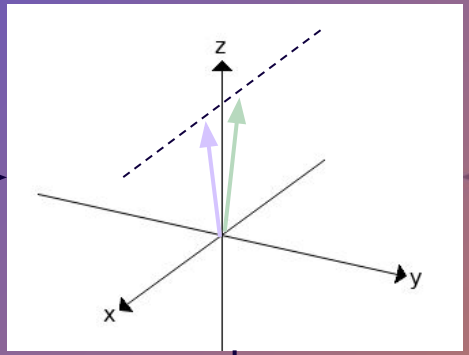


Image Encoder

Text Embedding ≠ Image Embedding

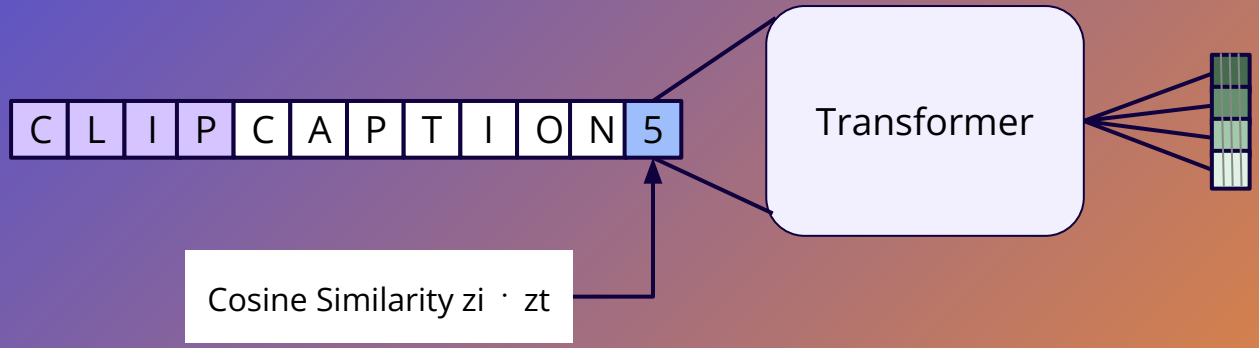
Image Decoder

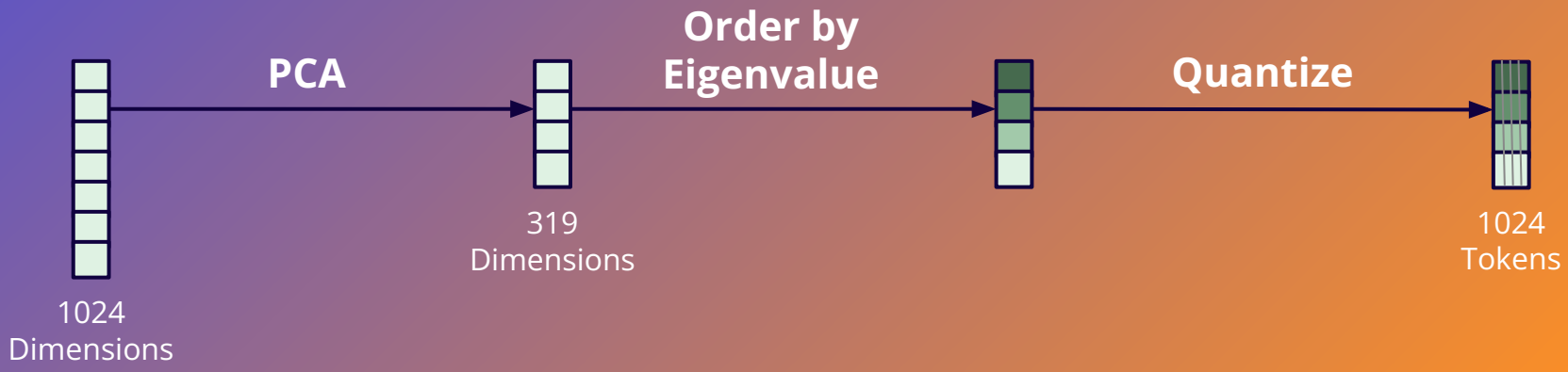
The Prior

- Converts CLIP Text Embeddings into Image Embeddings

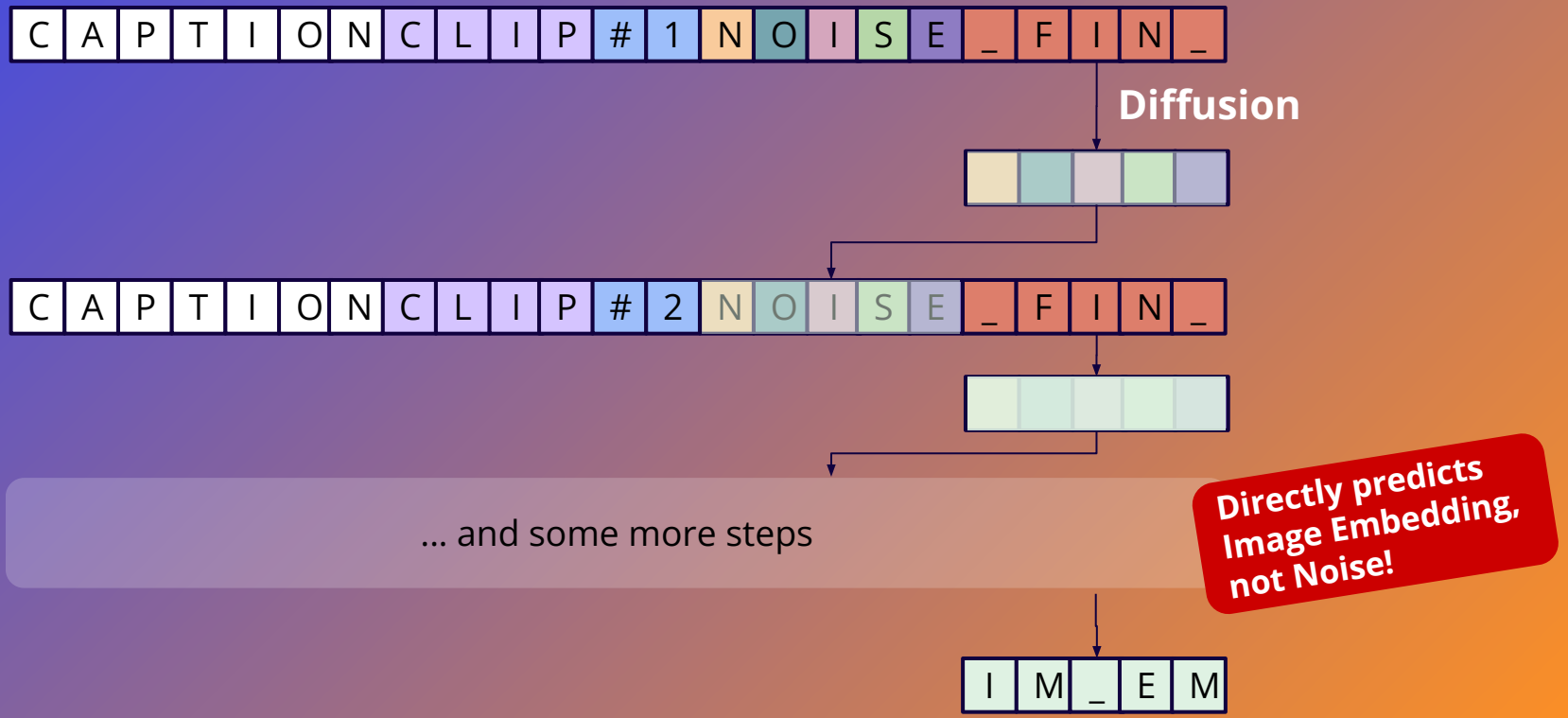
- Two versions:
  - Autoregressive Prior
  - Diffusion Prior

Autoregressive Prior





Diffusion Prior



unCLIP

A human in a banana costume

Text Encoder

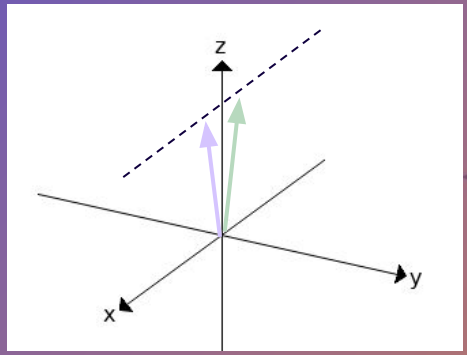


Image Encoder

Prior

Image Decoder

Model Details

CLIP:

- Image Encoder: ViT-H/16 @ 256x256px,
  1280 width, 32 Transformer blocks

- Text Encoder:
  1024 width, 24 Transformer blocks

- CLIP & DALLE Datasets (650M images total)

Model Details

GLIDE:

- 3.5b parameter model from the authors

Upsamplers:

- ADAMNet
- 1. stage: 256x256px, 27 steps, gaussian blur
- 2. stage: 1024x1024px, 14 steps, BSR degradation

Model Details

AR Prior:

- Transformer Text Encoder: 2048 width, 24 blocks
- Decoder: casual attention mask

Diffusion Prior:

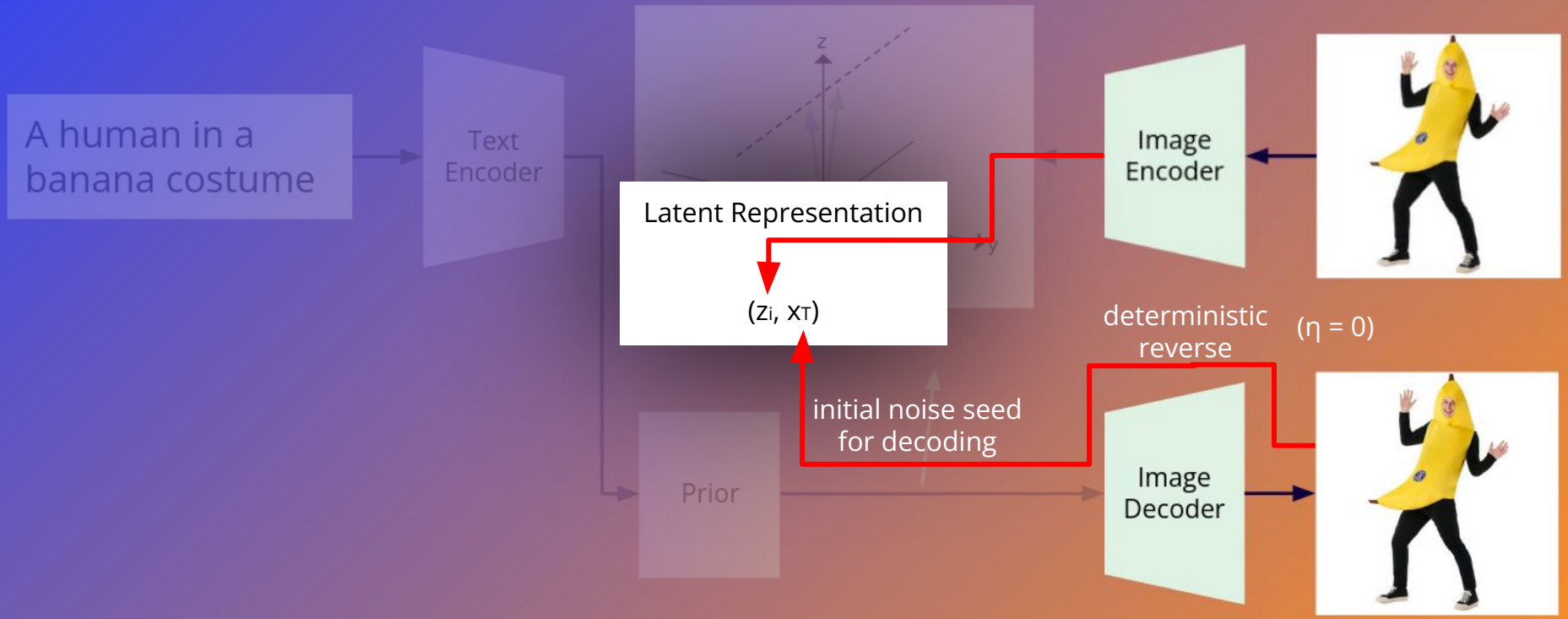
- Transformer: 2048 width, 24 blocks
- sample with Analytic DPM

Training Details

	AR prior	Diffusion prior	64	$64 \rightarrow 256$	$256 \rightarrow 1024$
Diffusion steps	-	1000	1000	1000	1000
Noise schedule	-	cosine	cosine	cosine	linear
Sampling steps	-	64	250	27	15
Sampling variance method	-	analytic [2]	learned [34]	DDIM [47]	DDIM [47]
Crop fraction	-	-	-	0.25	0.25
Model size	1B	1B	3.5B	700M	300M
Channels	-	-	512	320	192
Depth	-	-	3	3	2
Channels multiple	-	-	1,2,3,4	1,2,3,4	1,1,2,2,4,4
Heads channels	-	-	64	-	-
Attention resolution	-	-	32,16,8	-	-
Text encoder context	256	256	256	-	-
Text encoder width	2048	2048	2048	-	-
Text encoder depth	24	24	24	-	-
Text encoder heads	32	32	32	-	-
Latent decoder context	384	-	-	-	-
Latent decoder width	1664	-	-	-	-
Latent decoder depth	24	-	-	-	-
Latent decoder heads	26	-	-	-	-
Dropout	-	-	0.1	0.1	-
Weight decay	4.0e-2	6.0e-2	-	-	-
Batch size	4096	4096	2048	1024	512
Iterations	1M	600K	800K	1M	1M
Learning rate	1.6e-4	1.1e-4	1.2e-4	1.2e-4	1.0e-4
Adam β_2	0.91	0.96	0.999	0.999	0.999
Adam ϵ	1.0e-10	1.0e-6	1.0e-8	1.0e-8	1.0e-8
EMA decay	0.999	0.9999	0.9999	0.9999	0.9999

The funny things we can do with the CLIP latent space

Latent Representation



1. Image Variations

Latent Representation

(z_i, x_T)

$\eta > 0$

"make it more non-deterministic"

Image Decoder

1. Image Variations

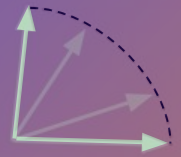
2. Image Interpolation

Image Encoder

z_{i1}

Image Encoder

z_{i2}



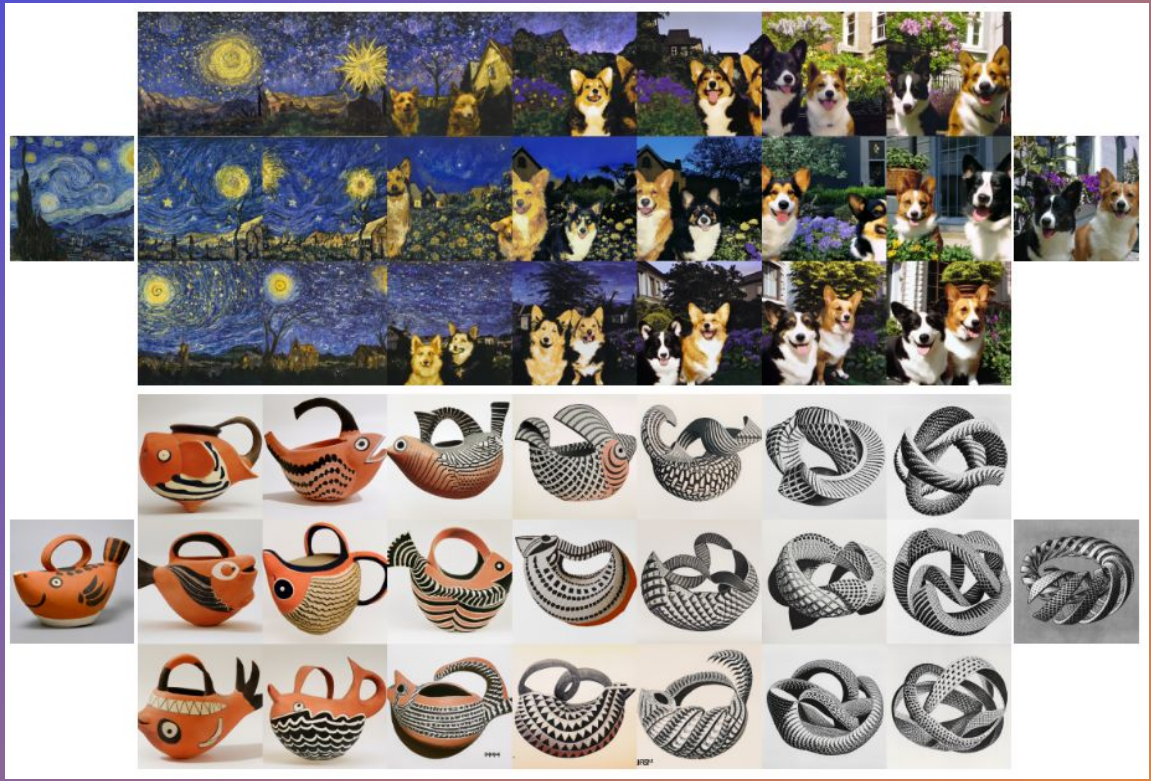
slerp(z_{i1}, z_{i2}, θ)

z_T

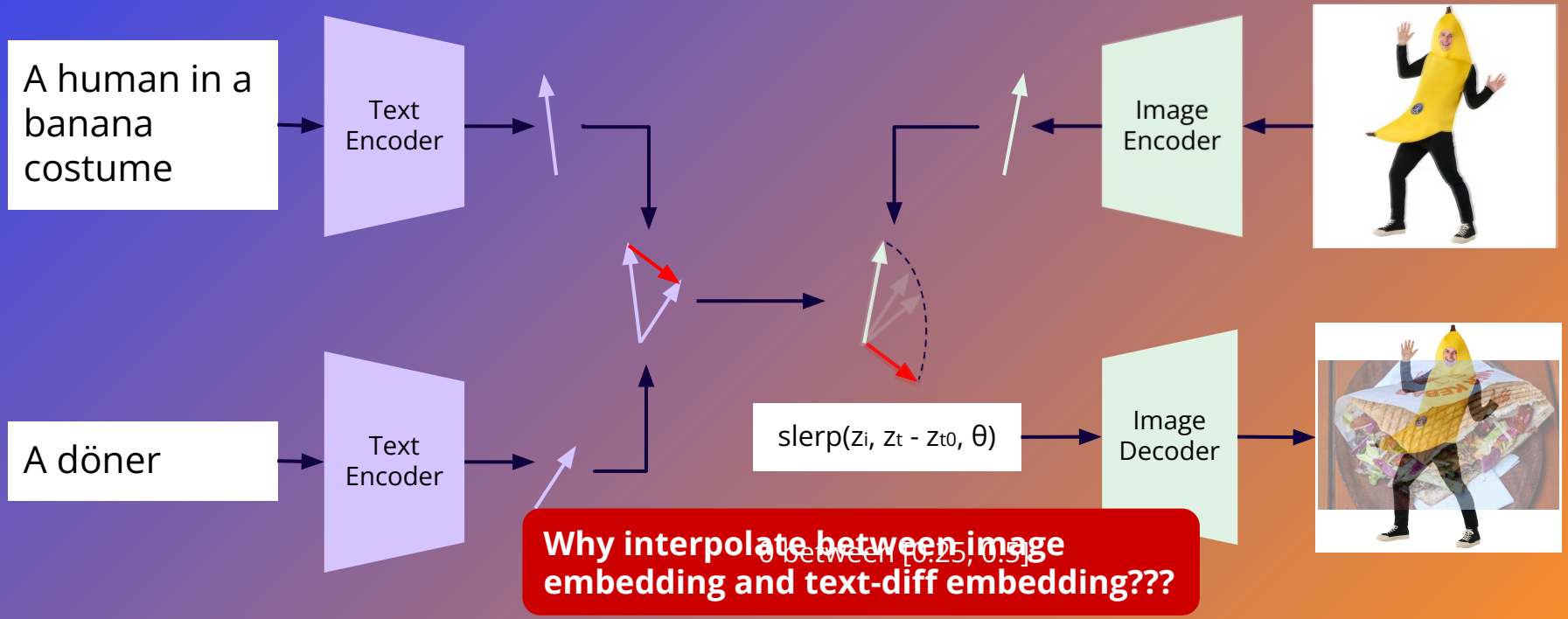
slerp seeds
or
pick random seed

Image Decoder

2. Image Interpolation



3. Text Diffs

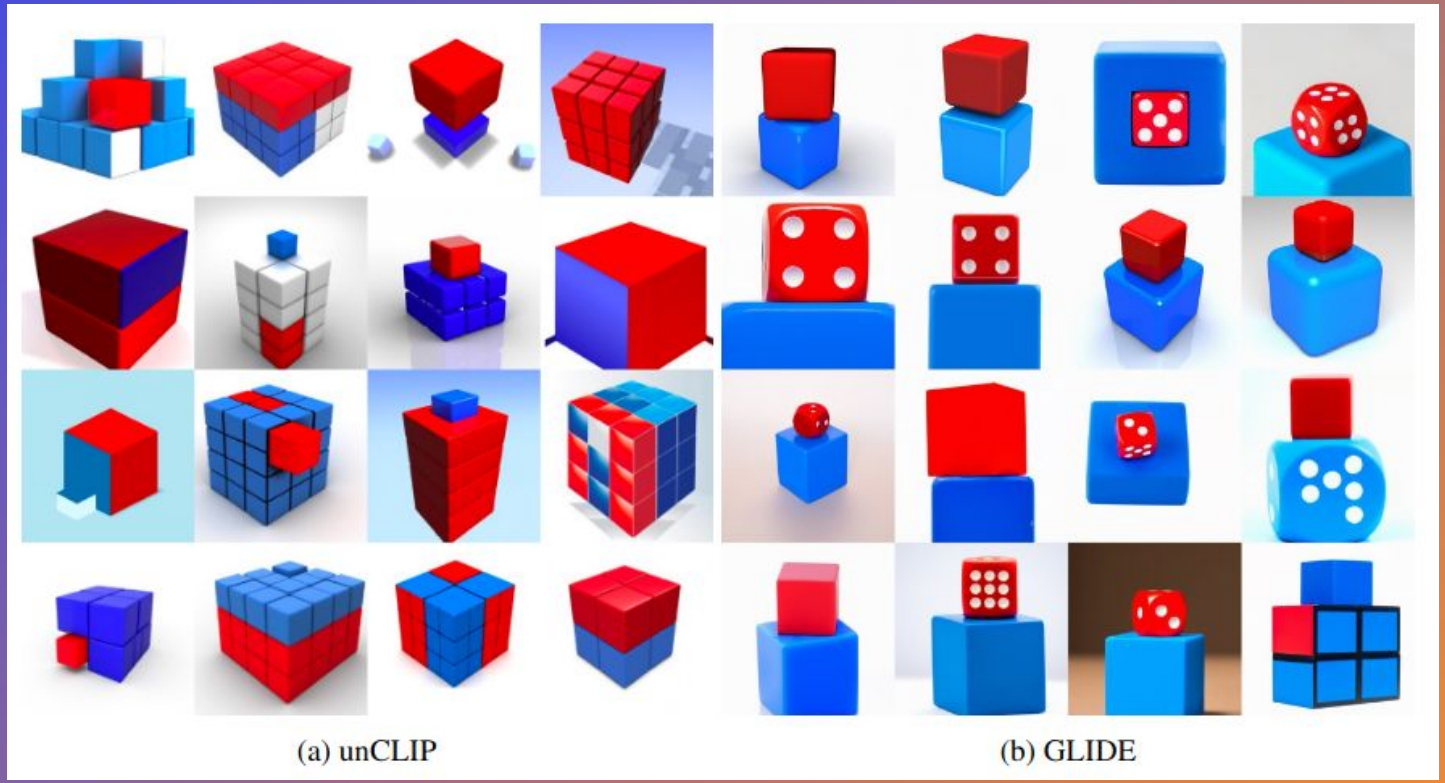


Why interpolate between image embedding and text-diff embedding???

3. Text Diffs


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

Problems & Limitations



“A sign that says deep learning.”

Problems & Limitations

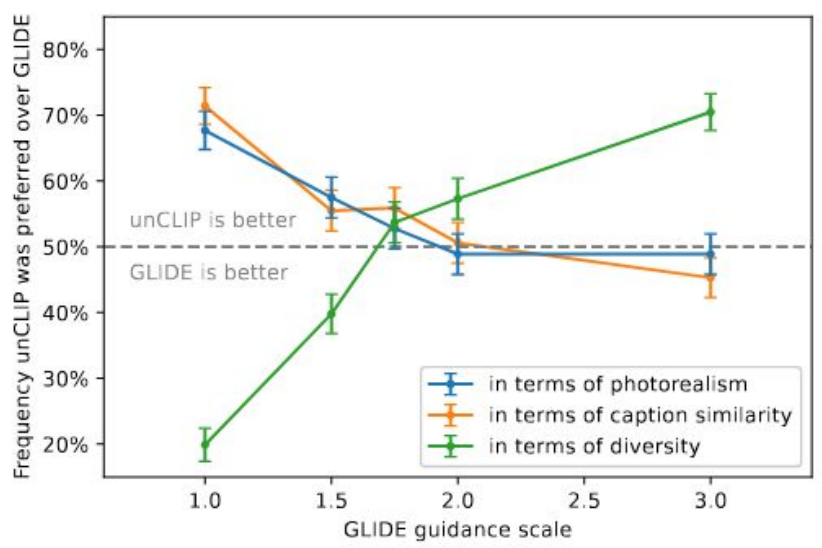
Evaluation

Model	FID	Zero-shot FID	Zero-shot FID (filt)
AttnGAN (Xu et al., 2017)	35.49		
DM-GAN (Zhu et al., 2019)	32.64		
DF-GAN (Tao et al., 2020)	21.42		
DM-GAN + CL (Ye et al., 2021)	20.79		
XMC-GAN (Zhang et al., 2021)	9.33		
LAFITE (Zhou et al., 2021)	8.12		
Make-A-Scene (Gafni et al., 2022)	**7.55**		
DALL-E (Ramesh et al., 2021)		~ 28	
LAFITE (Zhou et al., 2021)		26.94	
GLIDE (Nichol et al., 2021)		12.24	12.89
Make-A-Scene (Gafni et al., 2022)			11.84
unCLIP (AR prior)		10.63	11.08
unCLIP (Diffusion prior)		**10.39**	**10.87**

Table 2: Comparison of FID on MS-COCO 256×256. We use guidance scale 1.25 for the decoder for both the AR and diffusion prior, and achieve the best results using the diffusion prior.

Evaluation

unCLIP Prior	Photorealism	Caption Similarity	Diversity
AR	47.1% ± 3.1%	41.1% ± 3.0%	62.6% ± 3.0%
Diffusion	48.9% ± 3.1%	45.3% ± 3.0%	70.5% ± 2.8%



Evaluation

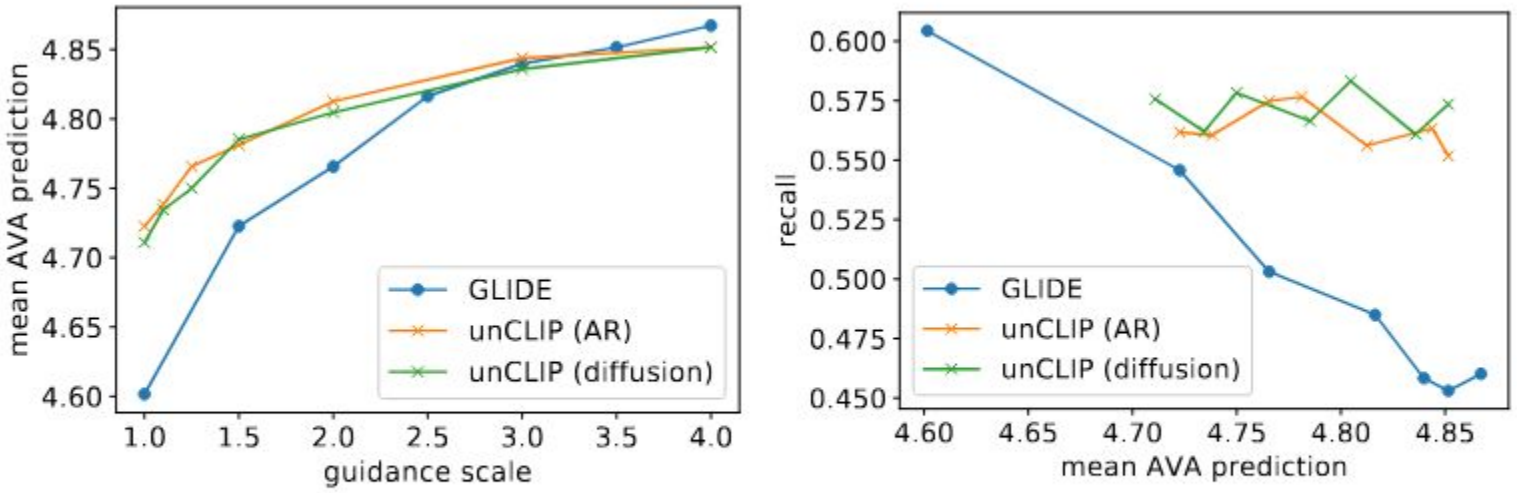


Figure 13: Aesthetic quality evaluations comparing GLIDE and unCLIP using 512 auto-generated artistic prompts. We find that both models benefit from guidance, but unCLIP does not sacrifice recall for aesthetic quality.

Samples