AIGA Executive Leadership Summit Workshop March 12, 2023

Text to Image AI
Latent Diffusion

Hugh Dubberly and Shelley Evenson

AI generated images are becoming more and more prevalent.

A canonical example of the possibilities
is an astronaut riding a horse

This one was generated using Midjourmey

Artist Jason Allen won the Colorado State Fair's digital art competition with this AI generated piece.

Image generating AIs take text as input and output images.

“a picture of a puppy sitting in a feld of poppies”

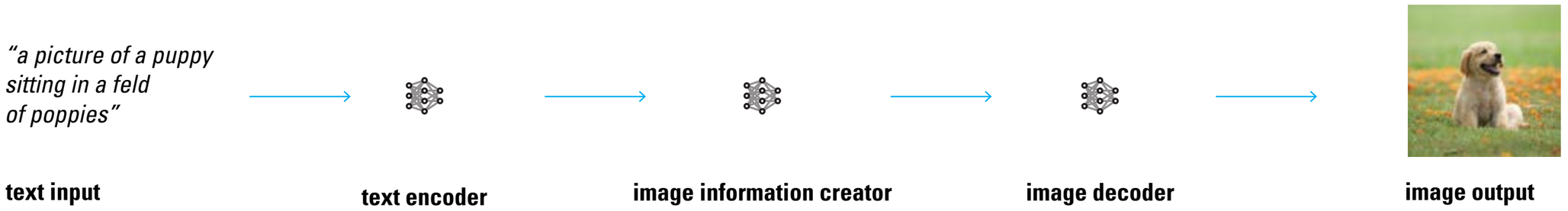
text input

image output

In order to accomplish this, three main components are used.

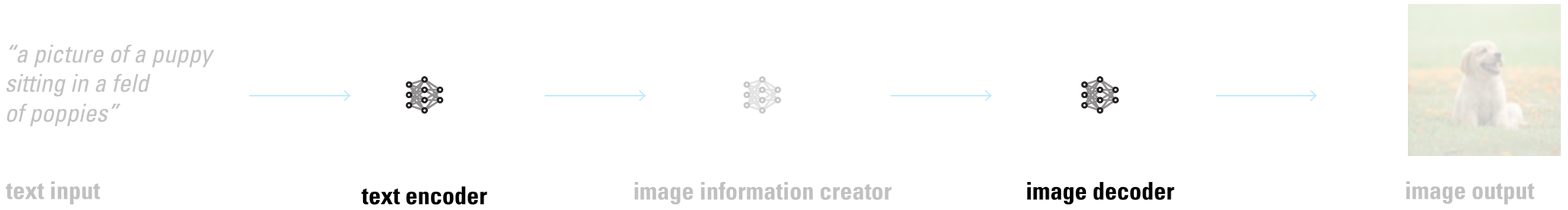
- a text encoder
- an image information creator
- an image decoder

Each with its own neural network.

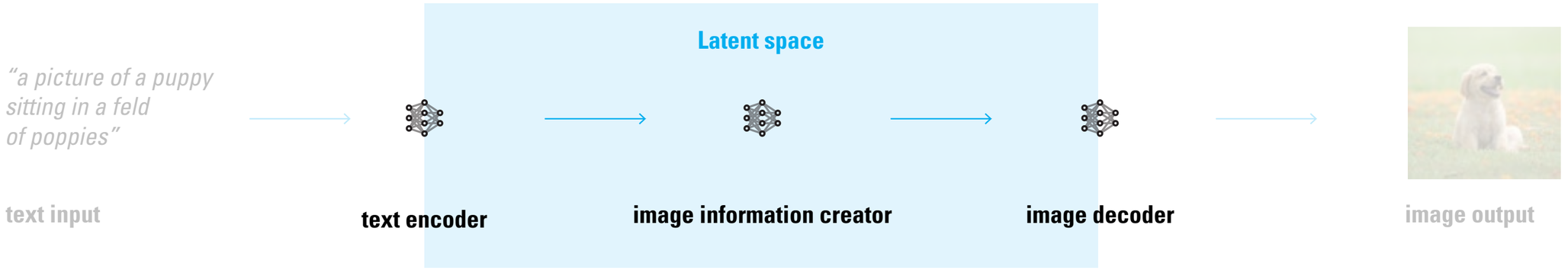


The process is sandwiched between an encoder and a decoder.

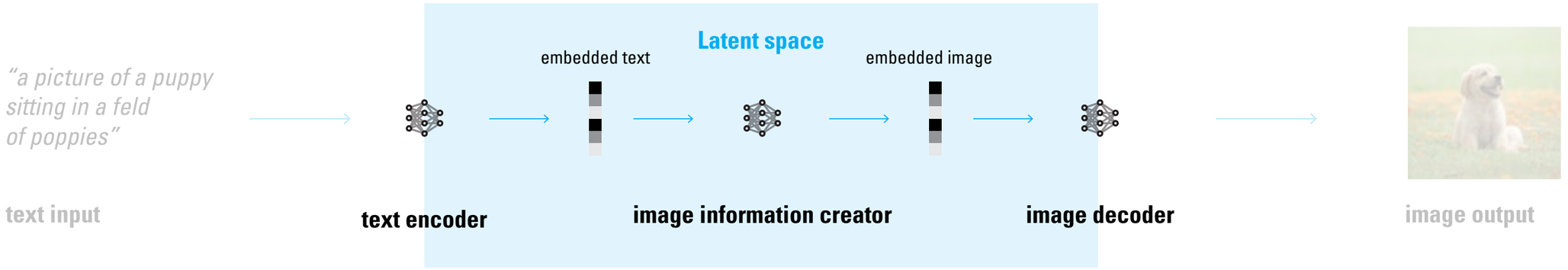
This is because the diffusion, or image generation process, does not handle text or images directly.



Rather it operates in the latent space.



Images and text are embedded as vectors, and can be mapped to the same space.



Humans input text.

The text is embedded.

The embedded text is used to generate an embedded, or latent image.

The embedded image information is converted to pixels.

Text encoder

Text encoders embed text and images as vectors.

Vectors are sets of coordinates, they represent locations within a space.

“a picture of a puppy sitting in a feld of poppies”

(45, 67, ... 98)

text that humans can read

embedded text

Text encoders are trained on images crawled from the internet and their alt tags.

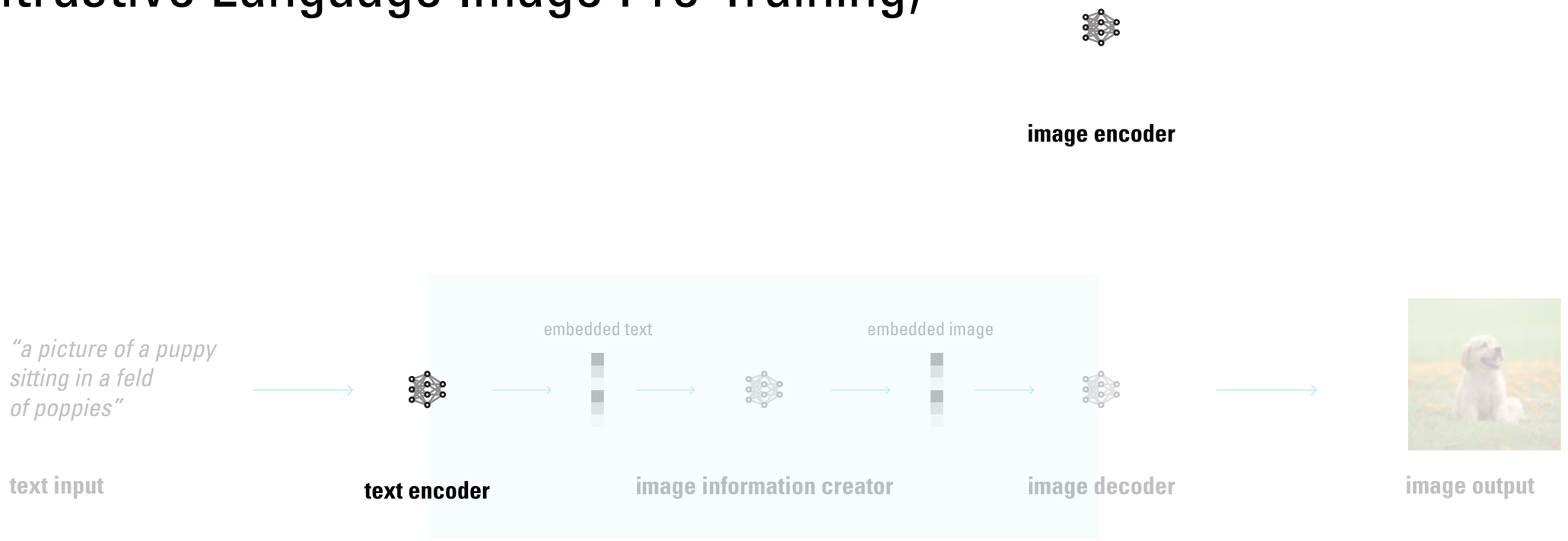
Alt tags are text to describe images on the web,
used by screen readers for accessibility purposes.

*“From top left to right:
the African bush elephant,
the Asian elephant
and African forest elephant.”*

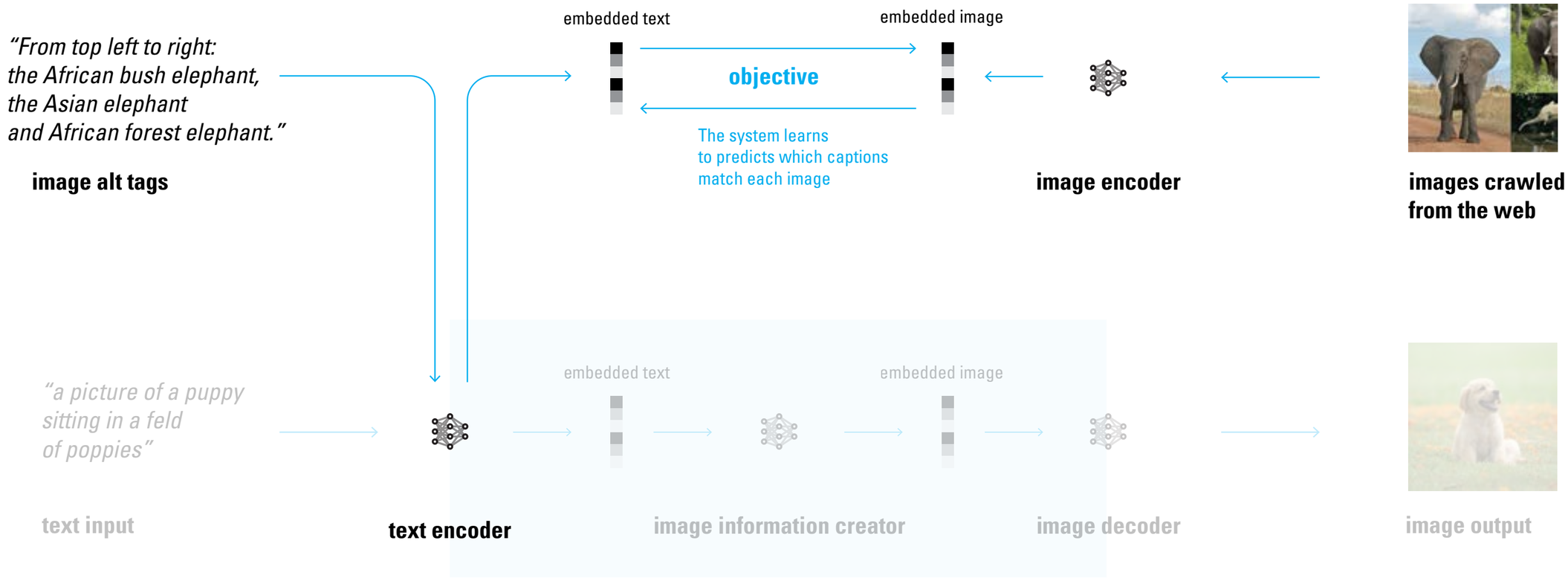
"San Francisco from the Marin Headlands"

To accomplish this, an image encoder is trained at the same time as the text encoder.

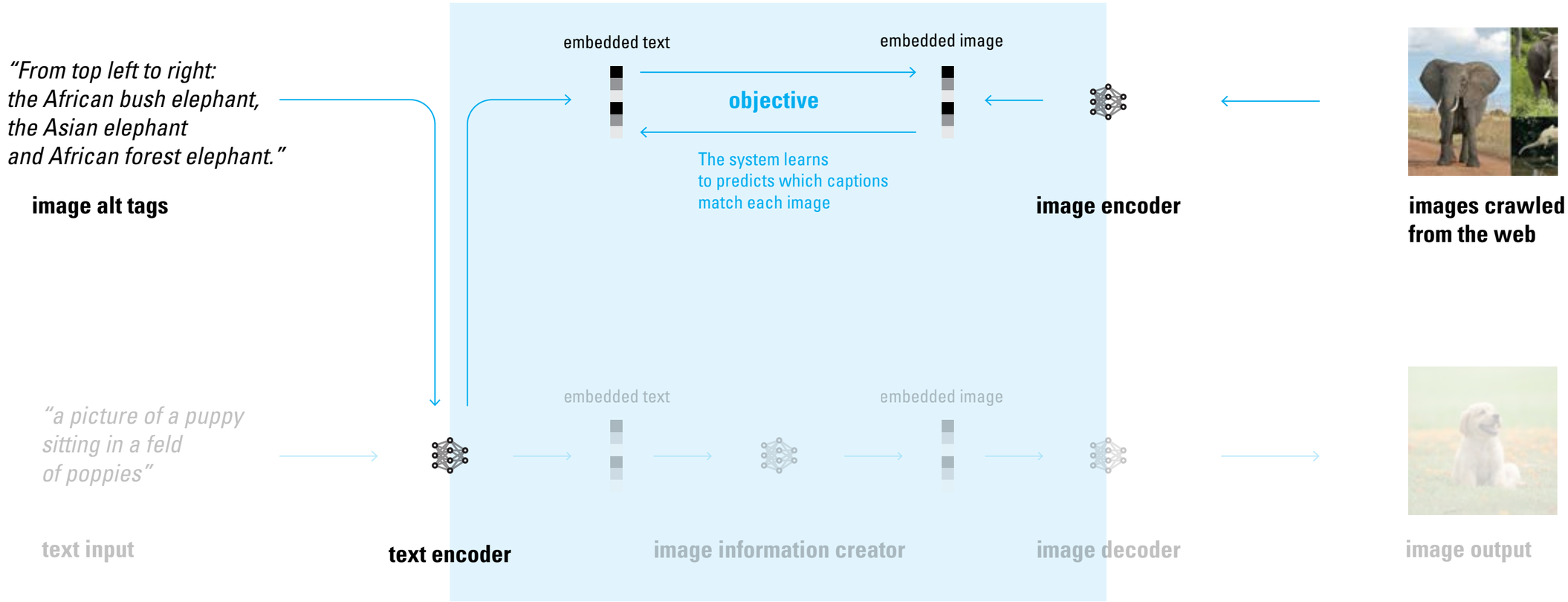
The system used by DALL · E 2 is CLIP
(Contrastive Language Image Pre-Training)



The images and captions are embedded, and the system predicts the likelihood of a match.



After training on hundreds of millions of pairs, the system learns a joint representation space for images and text.



This is the latent space, but it can be helpful to think of it as the meaning space or essence space

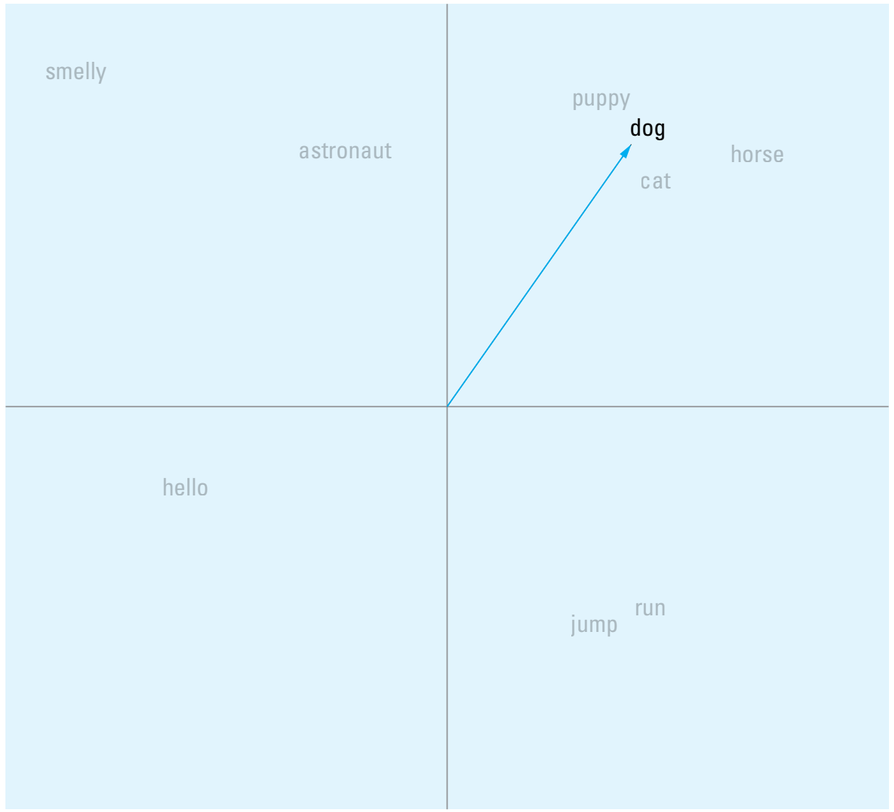
“From top left to right: the African bush elephant, the Asian elephant and African forest elephant.”

How its written

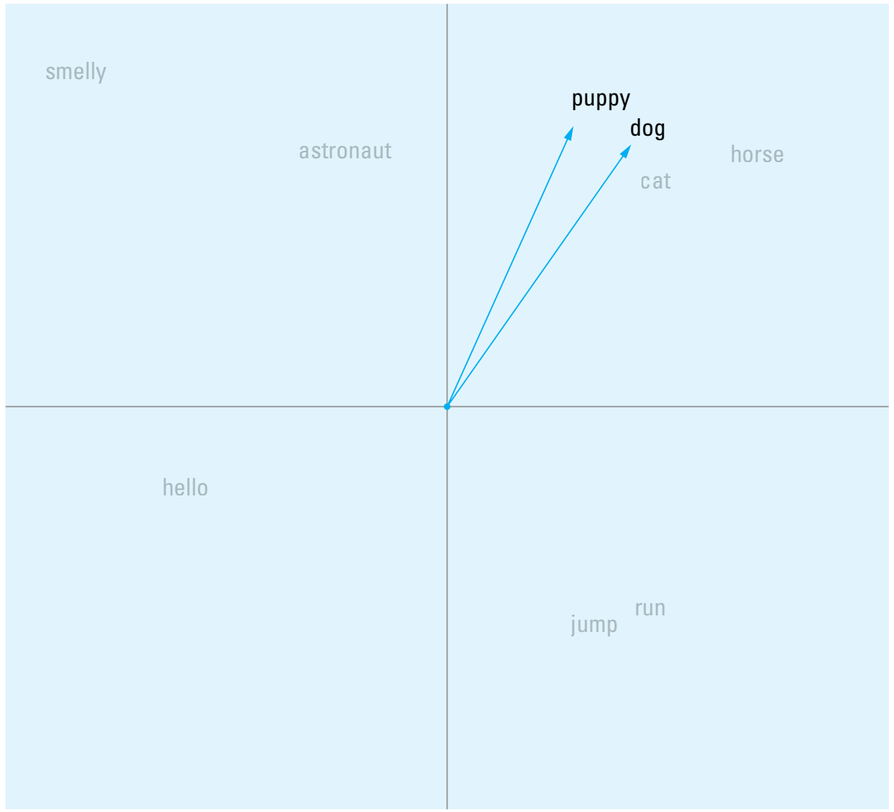
What it is

What is looks like

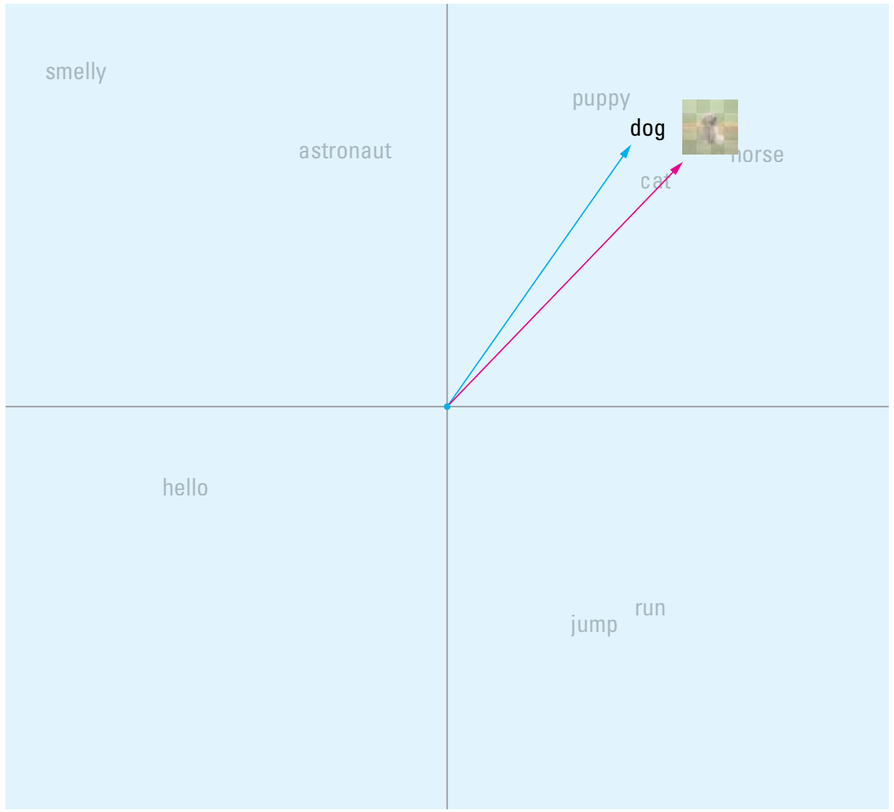
Embedded words are coordinates in this space.



Embedded words with similar meanings are closer together.



Embedded words and images of similar meaning also sit close together.



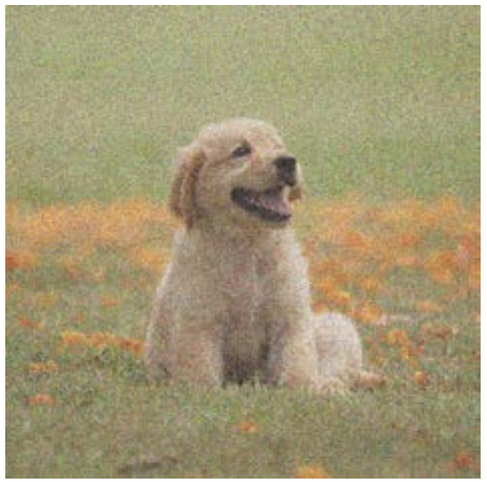
Text and images
can be compared directly
when they are embedded
in the latent space

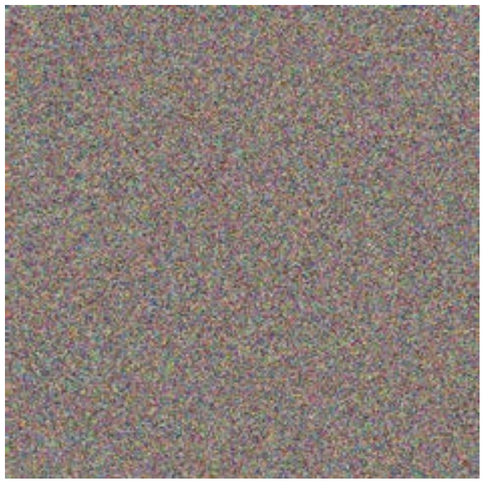
PART TWO

Image information creator and image decoder

Forward diffusion is the process of adding noise to an image until the image losses all structure, or becomes pure noise.

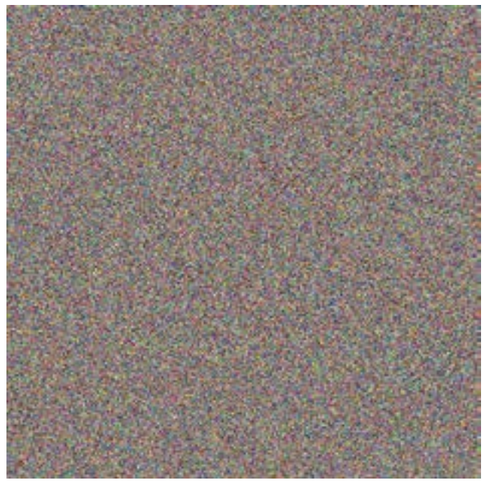
This is how the neural network is trained.



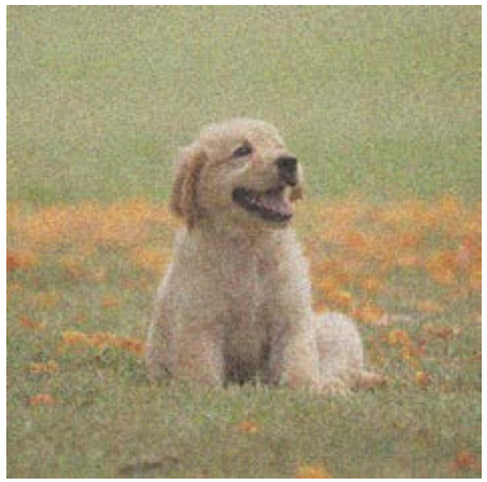
When trained, the network learns to reverse the process.

Reverse diffusion is the process of de-noising an image by iterating through a specified number of steps (t).



t0

t1

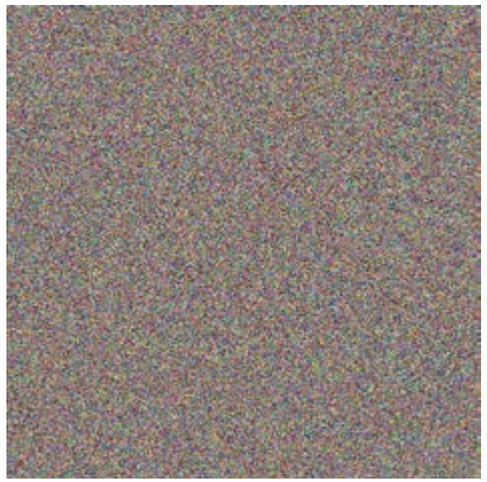


t2

t3

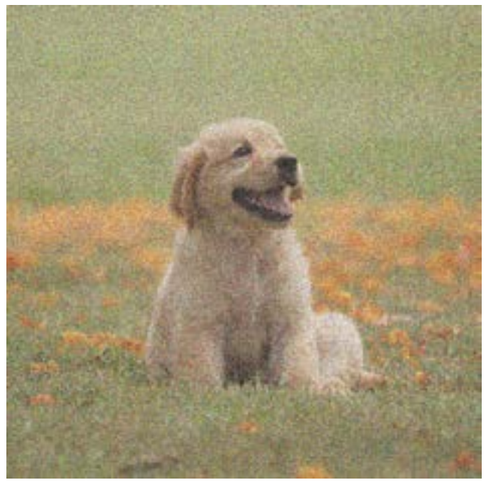
Typically there are between 50 and 100 steps

Both the image information creator (called the ‘prior’ is DALL · E 2) and the image decoder are diffusion models.



t0

t1



t2

t3

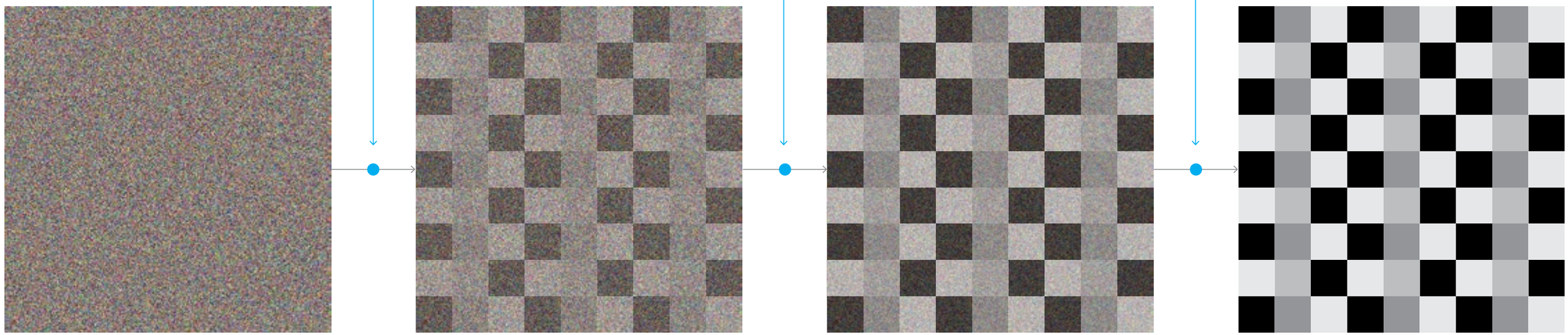
The prior maps the embedded text to a corresponding embedded image.

embedded text + noise

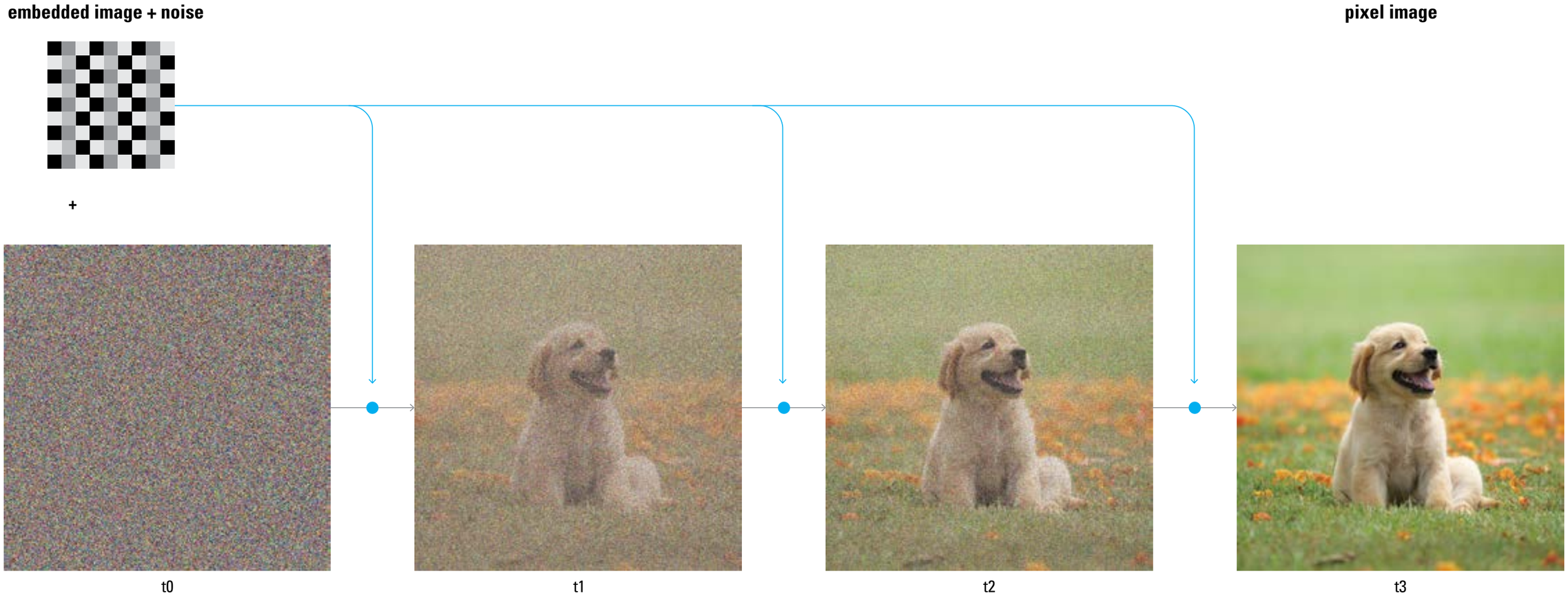
embedded image

“a picture of a puppy sitting in a feld of poppies”

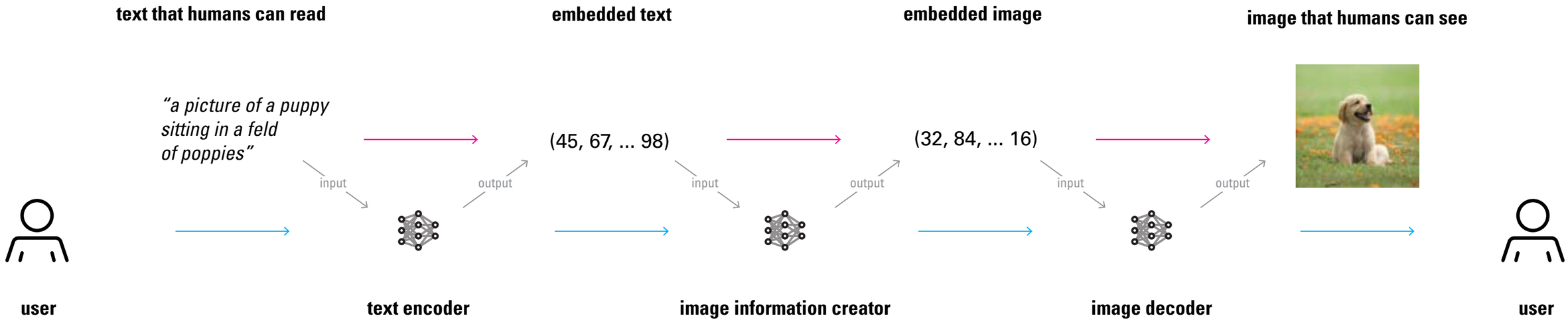
+



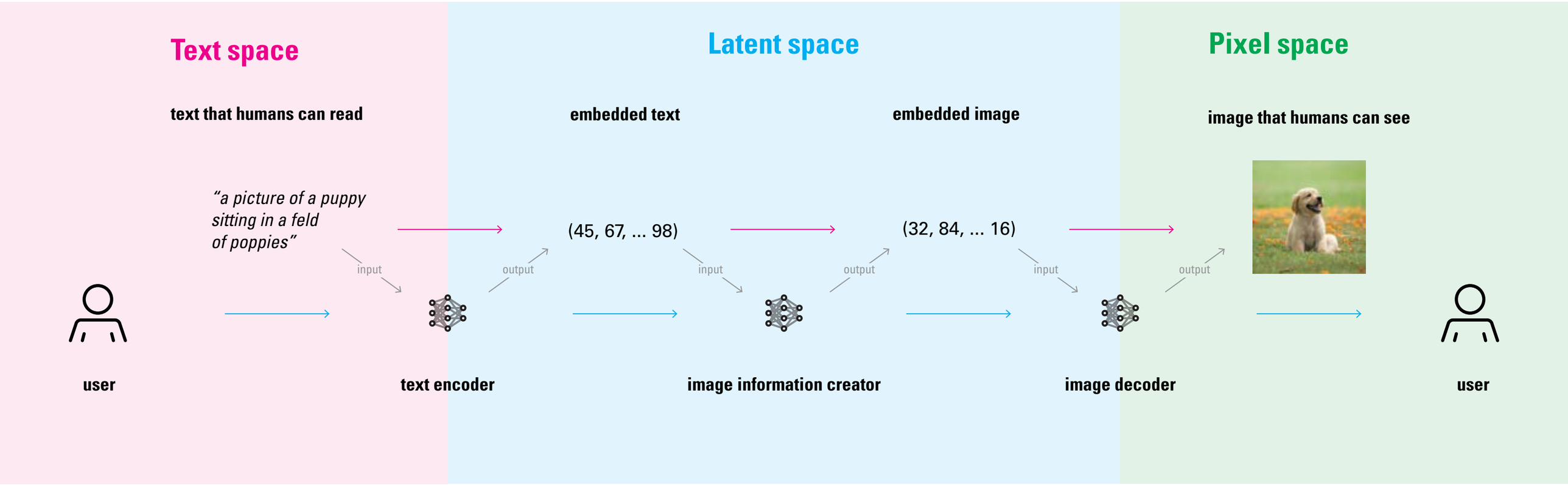
The image decoder converts the embedded image into a pixel image, and outputs it for viewing.



The process converts text into vectors, operates on those vectors, and outputs an image.



The process starts in the text space, the computation is done in the latents space, and the output is in the pixel space.



Appendix

DALL · E 2 can also produce images without using the prior.

Each row is produced using the same prompts.

In the first, the decoder is passed just the caption.

In the second, the decoder is passed the embedded text as if it were an embedded image.

And the third is using the full stack.

Bibliography

https://jalammar.github.io/illustrated-stable-diffusion/

https://eugeneyan.com/writing/text-to-image/#classifier-guidance-increasing-the-strength-of-promptsName

https://www.youtube.com/watch?v=1CIpzeNxIhU

https://arxiv.org/abs/2204.06125

https://www.assemblyai.com/blog/how-dall-e-2-actually-works/

Special thanks to
Gavin Miller
Ryan Reposar
Ian Shadforth
John Cain
Jake Sheiner

Presentation posted at
presentations.dubberly.com/ai_image_generators.pdf