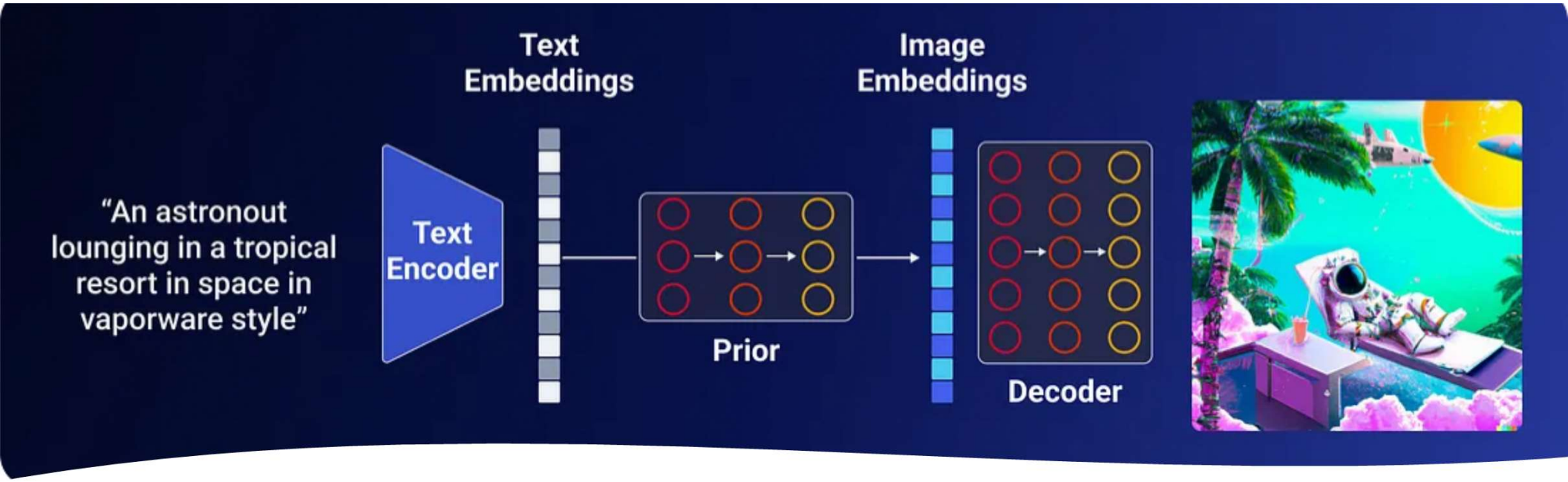

IMAGE – GENERATION
DALL-E 2
CREATED BY AND PRESENTED BY
BLAIESE SWARTVOOD and BRIAN BEASLEY

Image Generation

- Provide the machine with a prompt and it will attempt to generate that image for you.

- Prompt: Make me an image of a class watching a presentation over the new cutting edge DALLE software.



The Encoder Decoder Model

1. First, **CLIP** is the text encoder used to create the text embedding based on the caption and hold a spatial representation of caption and image relationships.
2. A model called the **Prior** takes this text embedding, along with the original caption, and generates the corresponding image embedding.
3. Finally, the **Decoder** generates an actual image conditioned on the image embeddings.

Source: Aditya Singh, How Does DALLE-2 Work?

CLIP

A neural network model that **returns the best caption given an image** by learning how *related* any given caption is to an image.

Matches images to their corresponding caption

2 Encoders; image and text embeddings

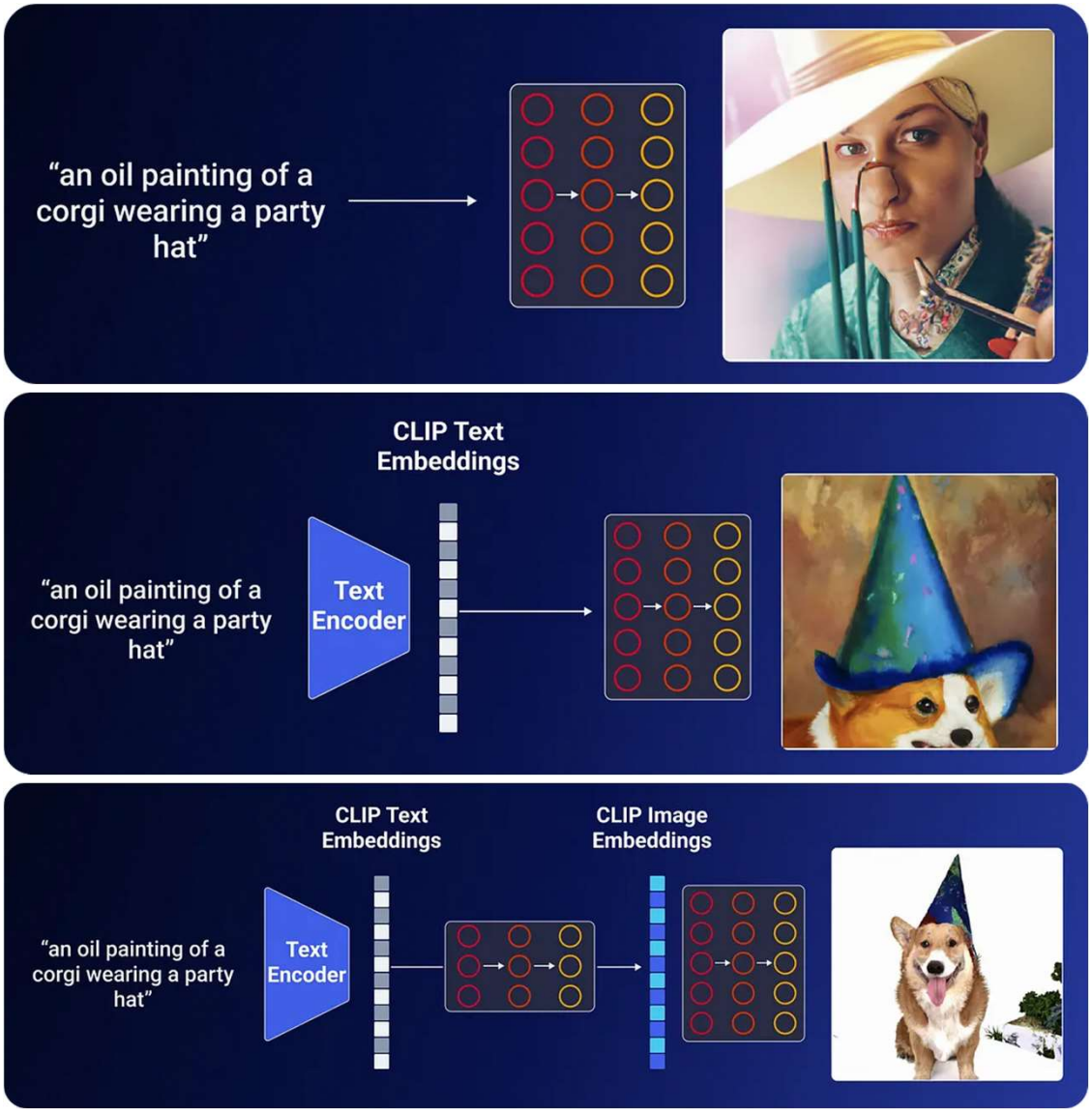
Training:

1. Generate the image and text encoding of each of the image-caption pairs using Transformers.
2. Calculate how "similar" the image and text are
3. Iteratively minimize the distance between incorrect image-caption pairs and maximize the distance between the correct image-caption pairs.

Source: Aditya Singh, How Does DALLE-2 Work?

PRIOR: Transformer

- Generates the CLIP Image embedding given a CLIP text embedding
- Trained using a decoder-only Transformer with a causal attention mask on a sequence consisting of, in order:
  - The tokenized text,
  - The CLIP text embedding,
  - An embedding for the diffusion timestep,
  - The noised image passed through the CLIP image encoder,
  - A final embedding whose output from the Transformer is used to predict the unnoised CLIP image embedding.



Decoder: Diffusion Model

- Takes the image embedding from the prior and produces an image from it.

GLIDE – invert image encoding process to generate the image

- Includes the text embedding as an input to support image creation
- Also includes the CLIP embedding
- Generate 64x64 images and up-samples twice for better resolution

Source: Aditya Singh, How Does DALLE-2 Work?

Significance of DALL-E 2

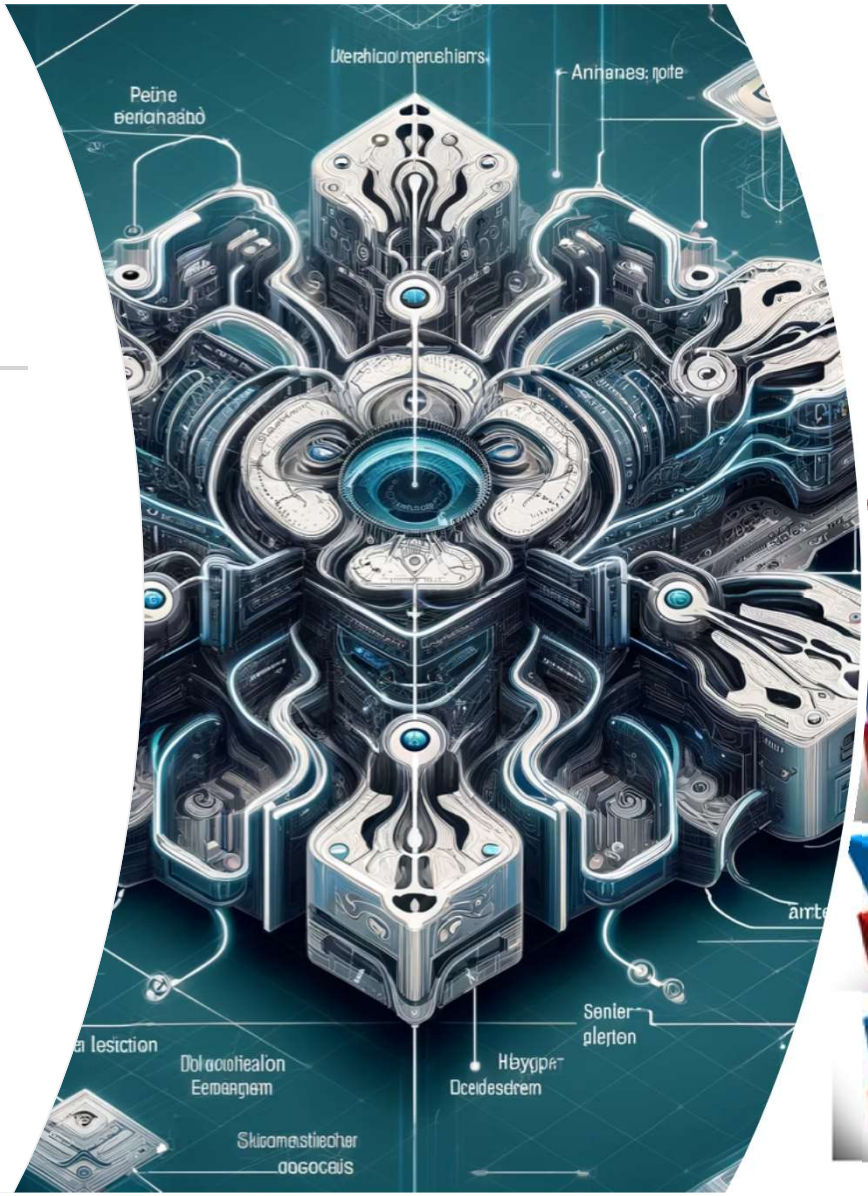
- Firstly, DALL-E 2 also has additional features beyond just image recognition such as variations and interpolations of images.
- Secondly, DALL-E 2 has outperformed previous models at the time based on standard tests.
- Furthermore, humans also evaluated images on three categories:
  - Photorealism
  - Caption Similarity
  - Diversity

Prompt: Give me an image that captures the significance of DALL-E 2

Shortcomings

Technical images

- Prompt: Give me a technical image of how DALLE's internal systems work, the decoder, encoder, etc.



Bind attributes to objects

- Prompt: Put a red block on top of a blue block

Source: Aditya Singh, How Does DALLE-2 Work?

Biases

Biased are being improved but still very much prevalent in DALLE image generation

“Based on our internal evaluation, users were 12x more likely to say that DALL·E images included people of diverse backgrounds after the technique was applied.” –OpenAI

Librarian

Architect

Social Worker

Lawyer

Potential Risks

- Guidelines and policies according to DALLE itself:
- **Intellectual Property**: I can't create images that include copyrighted characters, logos, or recognizable elements from copyrighted works
- **Privacy and Likeness**: I don’t generate images of specific, private individuals unless you provide a description of what they look like. For public figures, I can create images that might resemble them in gender and physique but not exact likenesses.
- **Sensitive Content**: I avoid creating images that could be considered offensive, promote misinformation, or depict violence, explicit content, or anything that could be harmful or sensitive in nature.
- **Artistic Styles Post-1912**: I can't generate images in the style of specific artists or artistic movements whose defining works were created after 1912. However, I can use general descriptions to create images in a similar aesthetic without directly copying their style.

Prompt: give me an image of AI overtaking humanity

The Future of DALLE

- Improving of images to look more realistic and less cartoony
- The ability to generate technical images
- OpenAI has already released DALL-E 3, an improved image generation model, and is now producing videos as well using SORA.
- Or even possibly world domination

Questions?

Thank you for listening!