Games & AI Techniques
COSC2527 / COSC3144

Week 4:
Generative AI

RMIT UNIVERSITY

Generative AI

- Generative AI = AI that generates text, images, sound…
- Concept has been around for a while, but huge breakthroughs in the last few years, mostly led by *transformer models* (outside our scope).
- Text generation: GPT-4, Claude-2, Llama 2, etc.
    - Useful for writing resumes and cover letters, generating code, converting dot points to proper paragraphs, answering questions (although unreliably), "helping" with basic programming challenges :p
- Image generation: DALL-E 3, Stable Diffusion, Midjourney.
    - Can generate photos that fool a lot of people.
    - Potential to replace stock photography. Some companies are already investigating replacing human fashion models (and getting backlash).
    - **Potentially huge for videogames**, movie special effects, etc.
- Video generation: Steadily improving, but often creepy, or ‘faked’ through generating consecutive images. May take a few years to develop.
- Music generation: Google’s MusicLM is extremely impressive. Text-to-speech and voice cloning are also improving rapidly.

A definitely ‘legally’ distinct video game character concept

Image Generation

Currently, there a few well-known text-to-image diffusion models:

- **Stable Diffusion (SD1.5 and SDXL)** – Free to use online, installable at home if you have an ok GPU, and small models if you don't.
- **DALL-E 3** – Available via Bing Chat, Bing Image Creator, and newer ChatGPT interface. Currently has the best prompt understanding (although Stable Diffusion 3 may change that).
- **Midjourney** – Access via the Midjourney Discord server. Historically was the most photorealistic, although they're all pretty close these days.
- **FLUX AI** – from Black Forest Labs free to use (dev model) and run locally, much larger in size than other models available from HuggingFace.
- **Adobe Firefly** – available through RMIT (express) and in Photoshop.
- Many, more... although they may just wrap around other models.

Stable Diffusion is easily the most flexible, because the models are downloadable and can be **finetuned**.

What is our focus today?

Today we'll look at a very high level at **Stable Diffusion**, which is a free-and-open-source image generation system. Our focus will be on the practical use of stable diffusion and benefits to us for generating artwork. We'll discuss the advantages of generative AI and limitations.

We'll start with a short intro in the principles of Machine Learning to understand the high-level of how stable diffusion works. Today we'll treat the internals of stable diffusion as a 'black-box'. We will discuss implementation details of neural networks in later weeks.

Today *is not* about "prompt engineering".

Note: Several of the showcase examples were prepared last year by Dr. Michael Dann using his own likeness.

Dr. Michael Dann

RMIT UNIVERSITY

Introduction to Machine Learning

RMIT UNIVERSITY

What is Machine Learning?

Machine learning is programming computers to optimise a ***performance*** criterion/perform a particular ***task*** by ***generalising*** from examples of ***past experience(s)*** to ***predict*** what is occur in future experience(s).

More technically

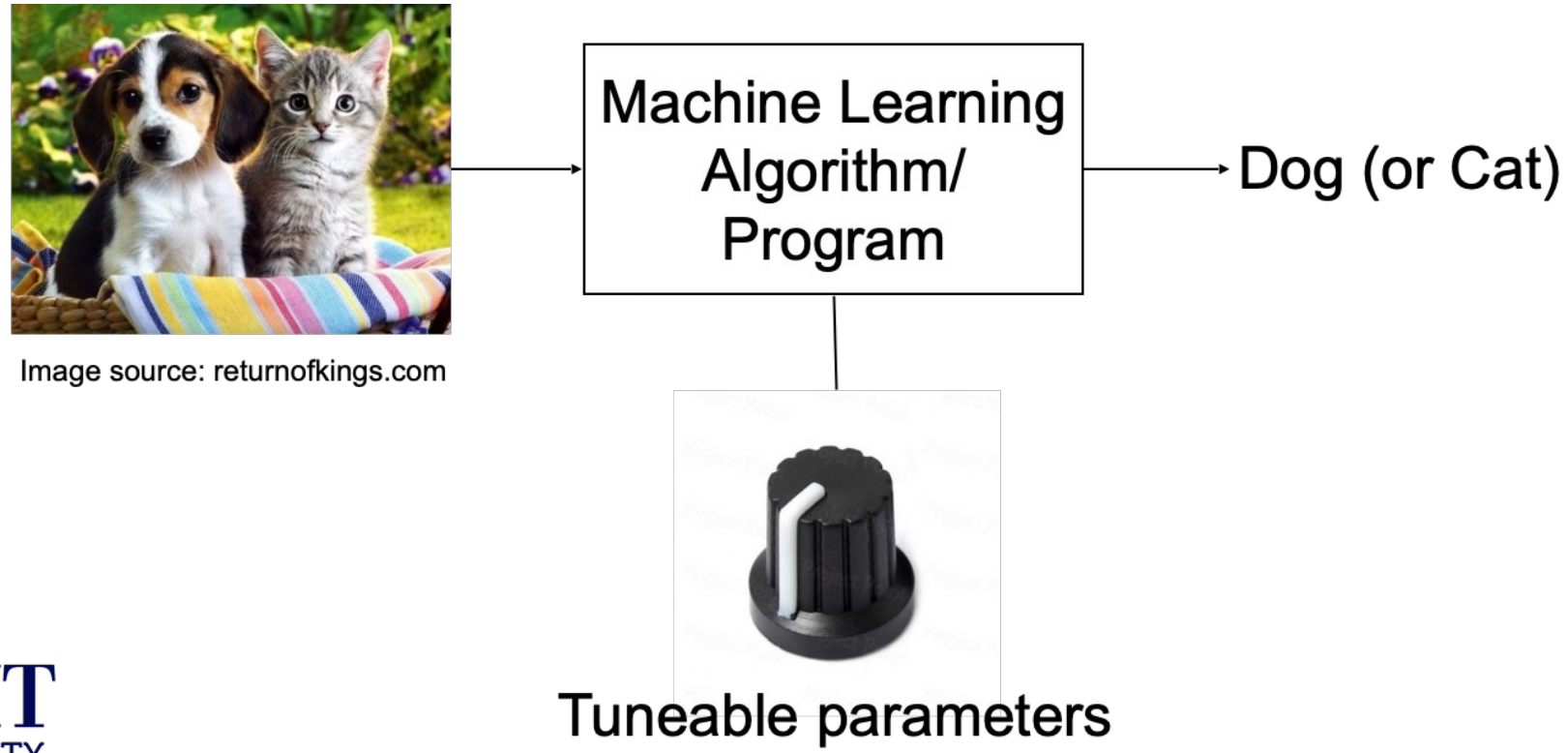
- *"A computer program is said to learn:*
    - *Some class of tasks T*
    - *From experience E, and*
    - *Performance measure P*
- *If its performance at tasks in T, as measured by P, improves with experience E"*

RMIT UNIVERSITY

What is Machine Learning?

A Machine Learning algorithm thus transforms an input through its algorithm into some form of output.

A classic image classification task is:



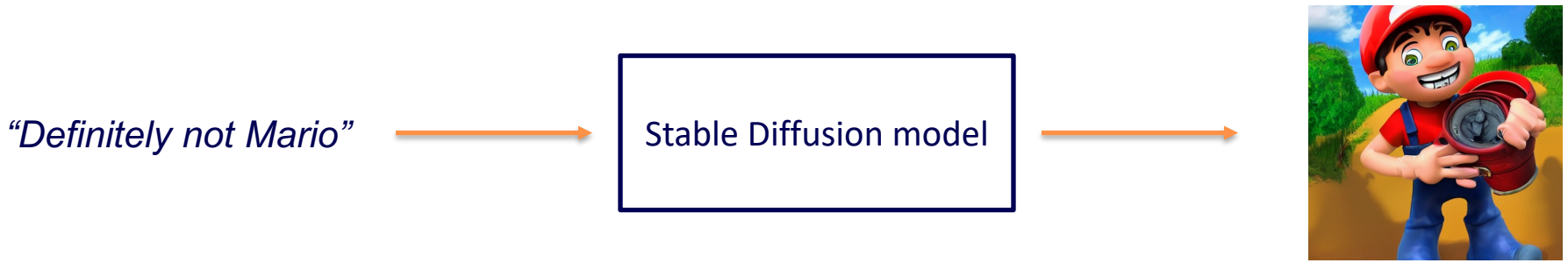
RMIT UNIVERSITY

Stable Diffusion as ML

For Generative AI (as used in Stable Diffusion), the process of Machine Learning is:

- Task – transform an input (primarily a text prompt) into an image.
- Experience –many examples of possible input/output pairs, of text prompts that describe given images.
- Performance – measures how "good" the transform is at producing images. We won't look into this today.

That is, the Stable Diffusion algorithm is trained by giving many examples of an image that *could* be produced from a text prompt input, and then executed by just giving the text prompt and seeing what output is *generated*.



RMIT UNIVERSITY

A note on Terminology

There is a tendency (especially in online sources) to confusingly use some terms interchangeably when describing Generative AI system (and neural network machine learning systems in general). The most over-used word is *'model'* which some may use to describe almost any aspect.

We'll try to relatively consistency use:

- **Architecture** – to refer to a neural network structure of which we'll see more in later weeks.
- **Model** – an architecture combined with a trained set of weights (tuneable parameters) which is the actual 'entity' that takes inputs and produces outputs. Multiple models are possible with the same architecture!
- **Model identifiers** – such as SD1.0, SDXL, MJv6, Flux.A1 which are unique identifiers for a version of a model.
- **Model Collections** – such as Stable Diffusion, Midjourney, etc. that are (now) a library of models trained and produced by a single provider often with a similar (or the same) architecture, named after their base model.
- **Tools/clients/utilities**– such as AUTOMATIC1111, to refer to software that is used to execute some form of a Generative AI system.

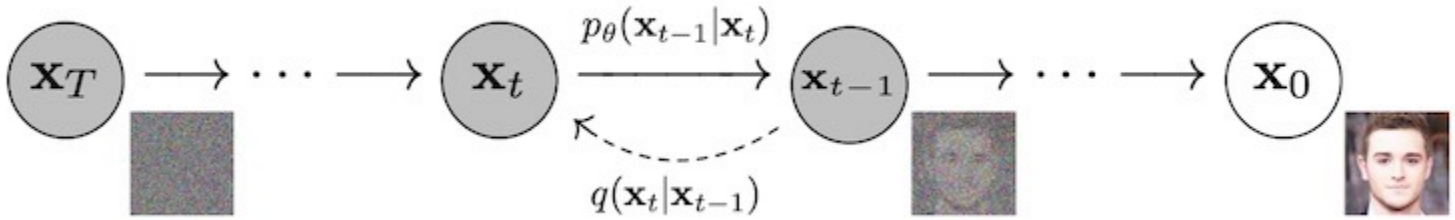
RMIT UNIVERSITY

What is Diffusion?

RMIT UNIVERSITY

Diffusion

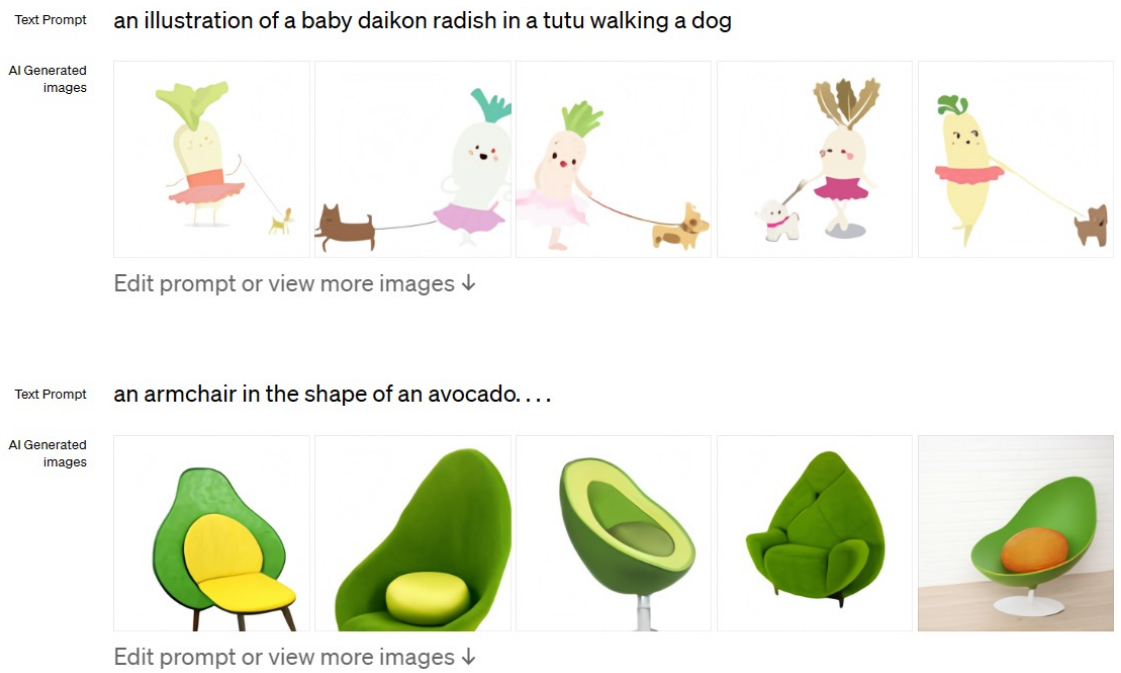
- Diffusion is an image Generative AI approach that has recently (since 2020) become more popular than other approaches due to the quality and tuneability of the images it can produce.
- DALL-E 2, Stable Diffusion, and Midjourney all use diffusion.
- At a high level:
  - The model is trained by progressively destroying each training image by adding noise (forward diffusion).
  - A *denoiser* is trained to undo the noise.



  - To generate a new image from scratch, we can start with pure noise then repeatedly apply denoising
  - Note: The reason the noise is applied iteratively is because it makes the denoiser easier to train. If the original input image was destroyed in one step, it would be hard to train the “undo” process.
- Each execution of such a diffusion model produces essentially a random image.
- However, when starting with a slightly noisy image the model can ‘imagine’ (generate) a new variation.

Text Conditioning

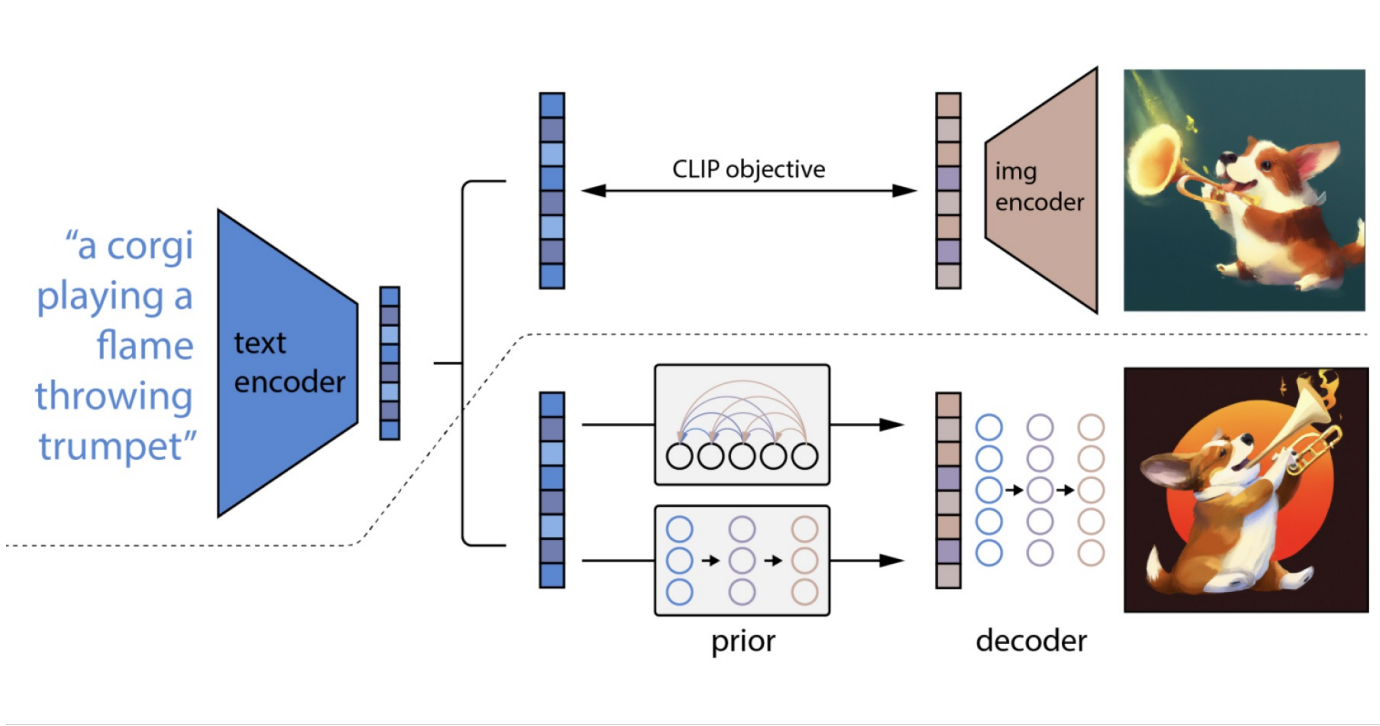
- The problem with diffusion is ‘guiding’ denoising to produce something more than a random image.
- Diffusion models can also be *text conditioned.*
- Text-to-image diffusion use text input (the prompt) to guide both forward diffusion and denoising.
- The training data set are labelled (captioned) images.
- During training, in addition to receiving images degraded by noise, they receive a text description of the undegraded image.
- After training, they can generate images from pure noise + text prompt.
- The technical details of this are well beyond the scope of this course, but if you’ve got some background in deep learning then you might find this blog post interesting: https://eugeneyan.com/writing/text-to-image/



Some examples from the DALL-E 2 paper.

CLIP

- An alternative to text conditioning that is used by Stable Diffusion models (and most other approaches) is an architecture called CLIP to ‘embed’ the text prompt into the diffusion architecture*
- This embedding effectively means that the prompt stays constant on each denoising loop so that each iteration slowly resolves into an image matching the prompt.
- The nature of this text embedding is also why behaviours (discussed later) are exhibited by stable diffusion models.



* We’re glossing over a whole bunch of detail using the terminology we listed earlier.

Finetuning

- Stability AI, who created Stable Diffusion, originally released various “base models” that were trained on a *huge* amount of captioned images.
- The amount of compute involved means it’s impractical to train your own model from scratch at home.
- However, a base model can be “finetuned” by providing a small set of extra captioned images (as few as 10) and training it for an hour or so.
- This is possible as the models are so large that only a small portion of each model is responsible for any one style, object, layout, “prompt element”, etc.
- So only a very small portion needs to be trained.
- The same concept exists for language models.

RMIT UNIVERSITY

Finetuning

RMIT UNIVERSITY

LoRA (Low-Rank Adaptation)

- Image generation Diffusion models can be fined-tuned to particular art-styles, or specific applications.
- LoRA models are very lightweight models that restrict image generation to only the fined-tuned purpose.
- People have made finetunes for photorealism, anime, pixel art, low-quality photography, etc.
- You can download many LoRA models, such as: https://civitai.com

"Not Mario" as a Disney princess

"Not Mario" pixel art sprites

RMIT UNIVERSITY

Using Stable Diffusion

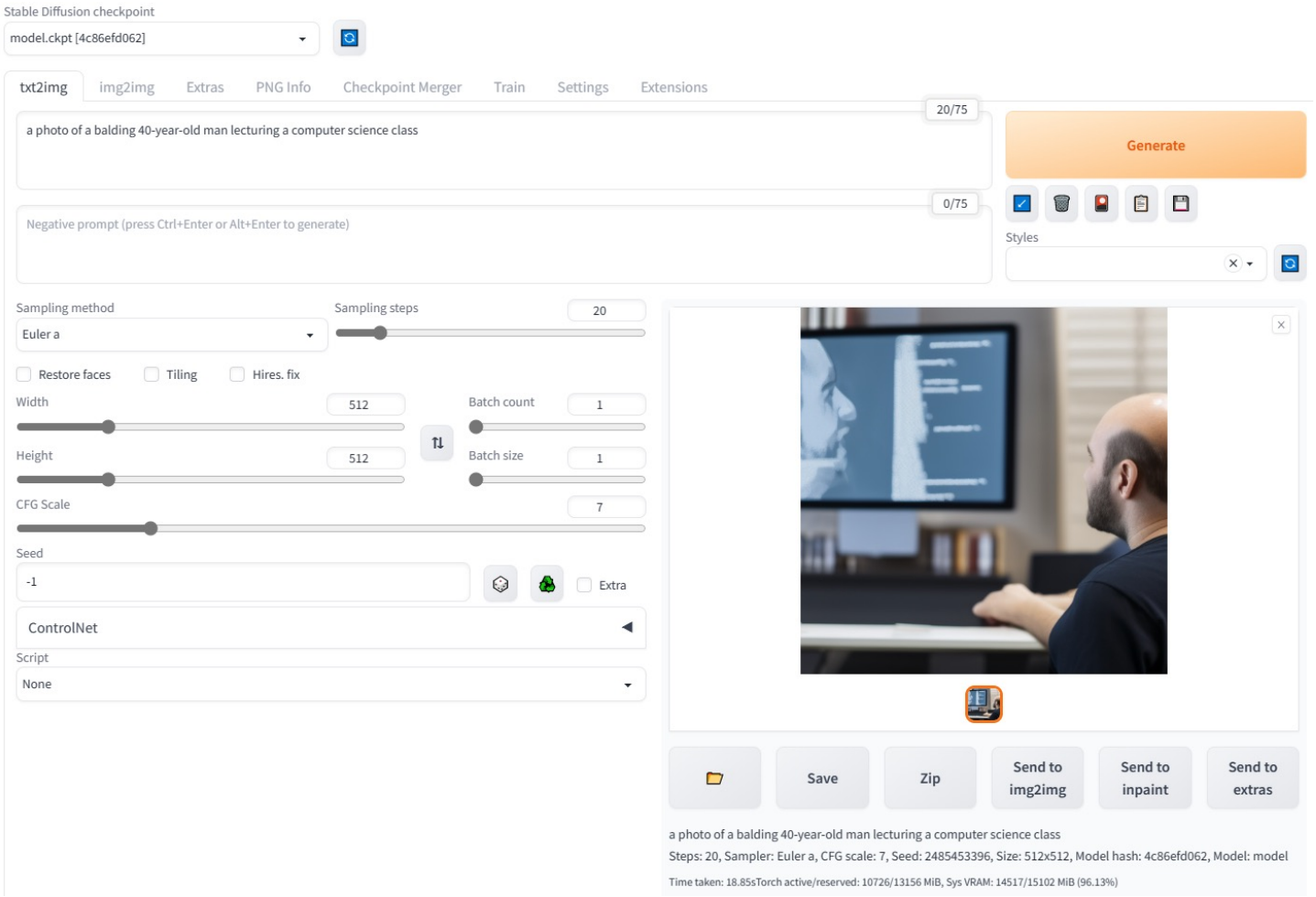
RMIT UNIVERSITY

Getting Started

- For a detailed beginner's guide, see https://diffusion-news.org/how-to-start. Alternatives:
  - https://stable-diffusion-art.com
  - https://www.reddit.com/r/StableDiffusion/wiki/guide/
  - or just Google.
- While it's possible to use Stable Diffusion from the command line, it's not fun.
  - AUTOMATIC1111's Stable Diffusion web UI is very popular → used in these examples.
  - A more advanced option with many more features is ComfyUI, improved a lot over 2024.
  - Many tools (both free and paid) are continually coming out.
- There are a few different options for running AUTOMATIC1111:
  - **Local Install**. Recommended if your PC can handle it. Ideally >= 8gb VRAM for SDXL.
  - **Kaggle or Google Colab**. Free GPU computing in the cloud. With Kaggle, you get 30 GPU hours per week. To use Colab you'll need to create a personal Google account to access it. (Your student account will be blocked from using Collab, I believe.) Free-tier users will occasionally get kicked off.
  - **Web Clients**. For example, FineTuned Diffusion. This is the easiest but least powerful option. Some advanced options will likely be missing, and image generation may be slow.

Generating an Image

Prompt: “a photo of a balding 40-year-old man lecturing a computer science class”.

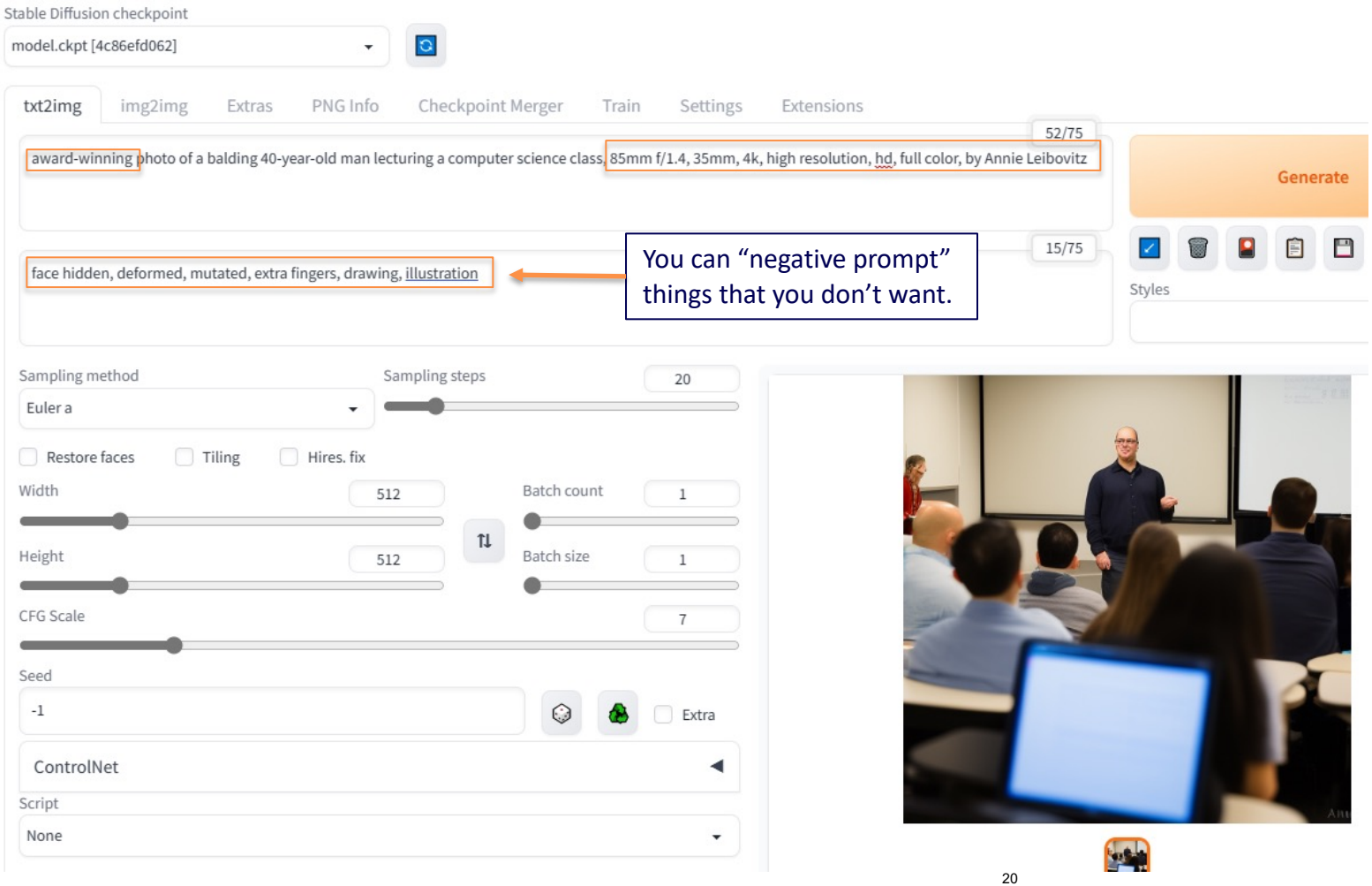


Notes:

- This is from 2 years ago, on SD1.5.
- The model was trained on 512x512 images, so varying the resolution may degrade the results.
- Roughly speaking, the “CFG Scale” controls how closely the image generator tries to stick to the prompt. Leaving this in the 3 – 10 range is good.
- The sampling method can have a big impact on the results. *Euler a* is fast to run, but *DPM++ SDE Karras* often gives much better results.
- You can fix the seed to regenerate an image exactly (useful for testing the impact of slight configuration changes).

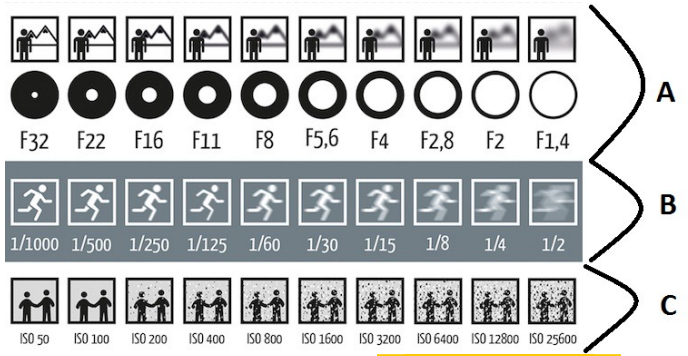
Prompt Engineering

Figuring out the best prompt to use is a bit of an art form! Sometimes terms like “award-winning” and “high resolution” can help, but there’s no real consensus on this.



Tips:

- As with most things, the best way to learn is by experimenting a lot.
- It’s very hard to tell from a single sample whether your text prompt is good.
- “Cherry picking” the best result from 5-10 generations typically yields a large improvement.
- While not really our focus, if you want to generate realistic-looking photos, it can help to learn some photography terminology:

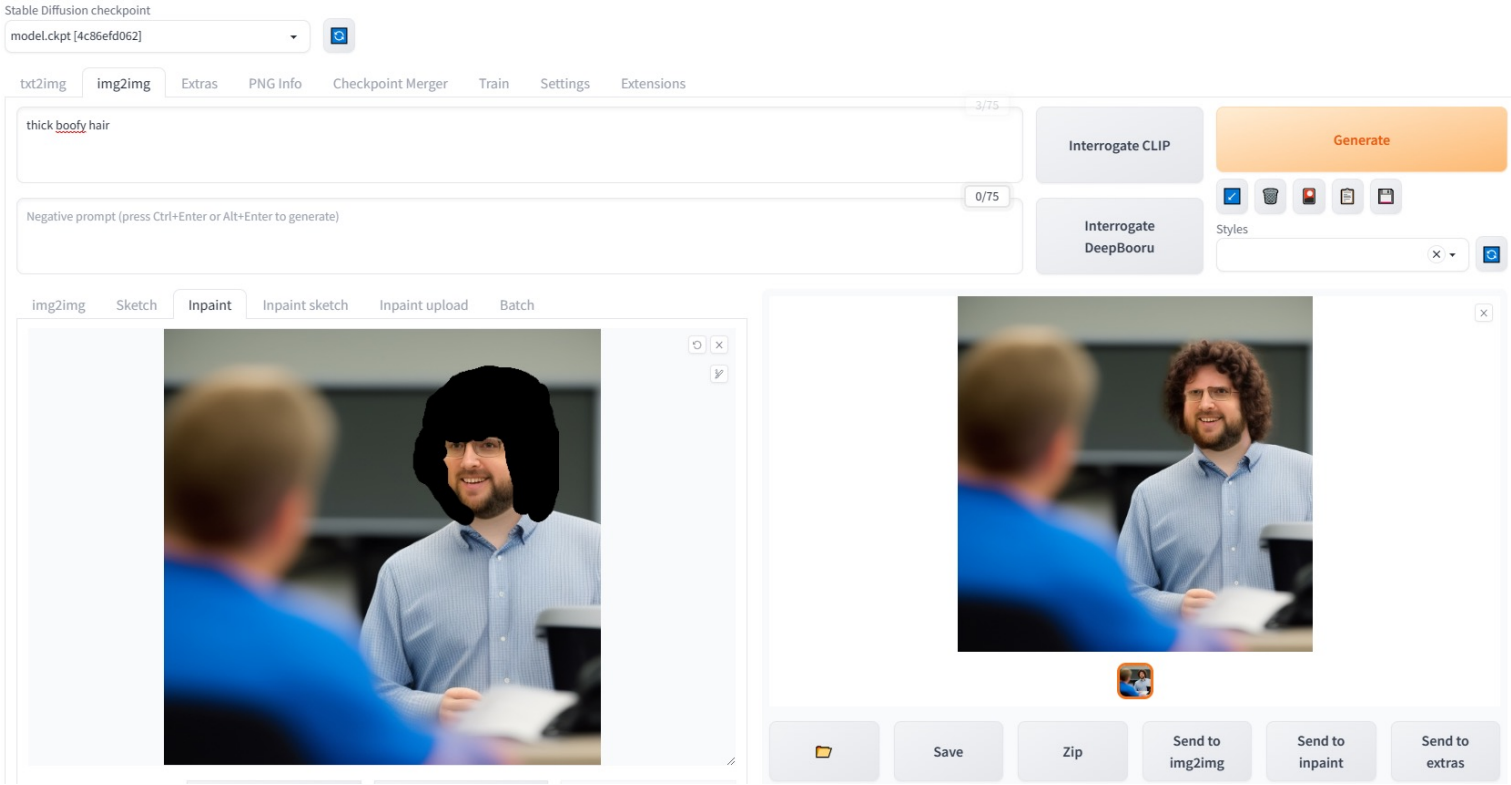


Inpainting

A common problem is that 80-90% of an image looks good, but a few things look wonky (especially hands and eyes, although newer models are improving on this).

Inpainting is a useful tool that allows you regenerate part of an image that is specified by a mask. Generally, you'll get better results if you alter the prompt based on the masked part of the image you're trying to edit.

Below, the inpainting prompt is "thick boofy hair"



Note: Any trained model can be used for inpainting, but some models have been specifically trained to be good at it. Usually they'll have "inpainting" at the end of their names.

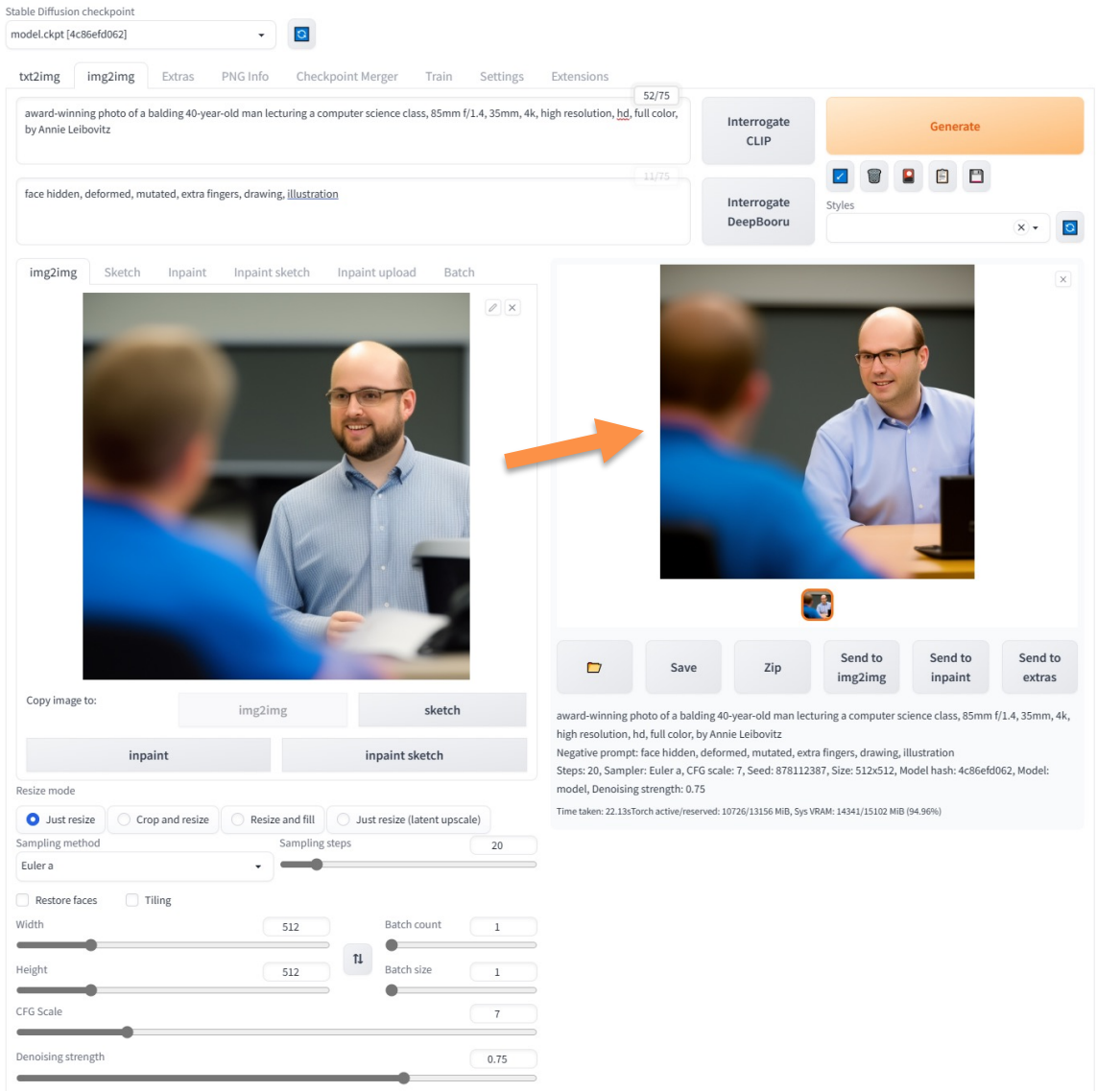
Image-to-image

Another useful tool is image-to-image, which allows you to provide an input image in addition to a text prompt.

The “denoising strength” controls how much the new image varies from the input image.

- Denoising strength = 0: The new image will look exactly the same as the input image.
- Denoising strength = 1: The new image will potentially look very different from the input image.

This can be useful for performing a “realism pass” after inpainting, making the joins less noticeable.



RMIT UNIVERSITY

Post Production

For best quality results, it's often helpful to do some post-production using image editing software.

Adobe Photoshop is good, but expensive. Krita is free, which is available on Windows, MacOS and Linux.

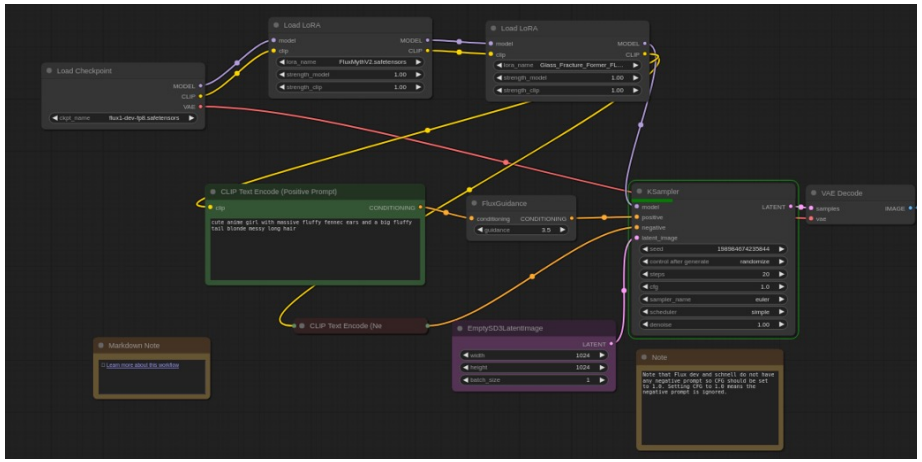
For example, a useful approach is transparency masks combined with the gradient tool. This lets you blend two images together seamlessly.

RMIT UNIVERSITY

Flux & ComfyUI Workflows

An alternative to the post-production workflows is using Built-in utilities and adjustments with ComfyUI. This is a complex tools designed to connect many of the features into a plug-and-play system. Installing ComfyUI is straight-forward, but some research may be required to successfully configure a good workflow.

Flux1.Dev is currently one of the best models (but very large at approx. 17GB). ComfyUI is the best way to run Flux models.



ComfyUI examples with different flux models and example prompts

RMIT UNIVERSITY

Limitations of Generation

Don’t believe the hype - Generative AI doesn’t “invent” new things. It composes from previously seen data. This means it struggles with concepts, and especially object meaning and relationships.

Question: Why do you think Generative models struggle with these types of prompts?

Prompt: *a collection of analogue wrist watches that are all showing the time as 12.03pm*

Prompt: *a picture of a person writing on a notebook with their left hand, using a pencil.*

Limitations in Making Game Art

The below are generated with SDXL pixel-art finetune with the prompt:
“An 8-bit sprite of a swordsman in the style of Legend of Zelda, white background”.

Question: What are some of the practical difficulties here for video game use?

RMIT UNIVERSITY

Student Work Showcase

Some past students managed to generate animations, which is a fair bit harder!

Their full presentations are on the Week 4 canvas module, in case you're interested.

You'll get to do some cool stuff for the next assignment too :)

RMIT UNIVERSITY

Ethics & Legality

RMIT UNIVERSITY

Inpainting + Image-to-Image + Krita

To give you some idea of the potential for fake news, etc…

Do you recognise the people on the left?

RMIT UNIVERSITY

Copyright and IP

- Models will generate near-identical looking outputs to existing copyrighted material.
- Some tools (such as Midjourney's discord system) may have filter restrictions to avoid such situations, but generally the actual models have no concept of material that shouldn't be generated.
- It can still be a copyright violation, even if material is generated if the final output is too similar.
- There are also major debates over the training data used for models, and whether copyrighted material is used.

RMIT UNIVERSITY

Content Filters

- Like copy-righted material, image generation models do not have specific filters for “objectional” material.
- Generation Tools may post-filter content.
- Some providers of models avoid objectional content by not training their models using that material in the first place, but it’s not possible to block everything. It’s also unclear exactly what internal connections might be established in a model from the training data set, and thus what is can generate.
- Be especially careful of LoRA models. LoRA models are much easier to finetune, and quite small. Thus, the internet being what it is, unfortunately many freely available LoRA models can have a tendency towards objectional material.

RMIT UNIVERSITY

Considerations for this Course

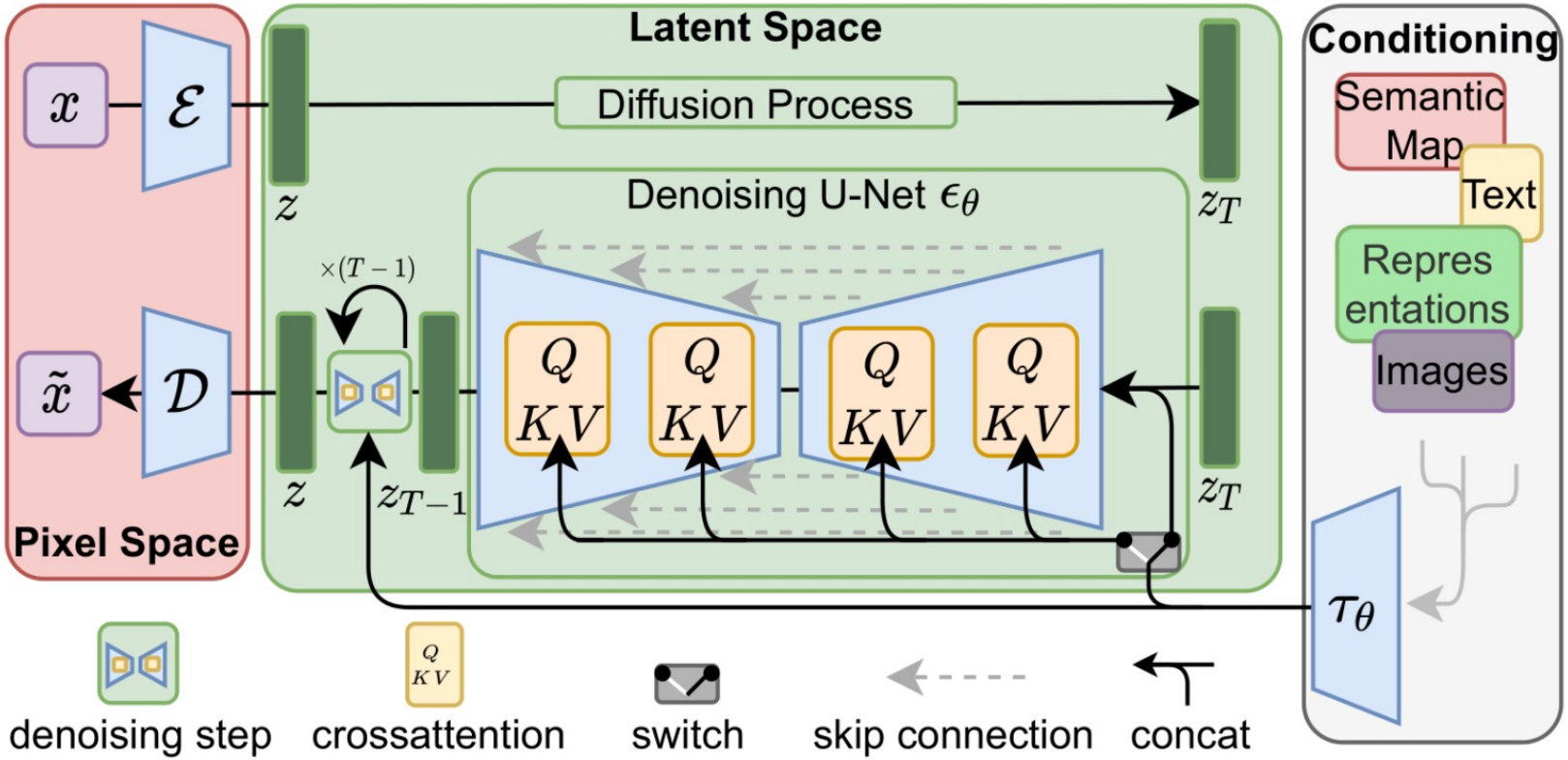
- We are happy for you to use Generative AI tools for creating unique art for your assessments. This can be beneficial over art libraries, so you don’t violate image copyright use.
- Acknowledge all use of Generative AI tools (and AI tools in general) in course assessments.
- However, also pay attention to violating copyrighted images and characters. You must have the rights to use any material in your assessments.
- We expect that all content is G-rated, and follows RMIT’s policies of respect and inclusion.

RMIT UNIVERSITY

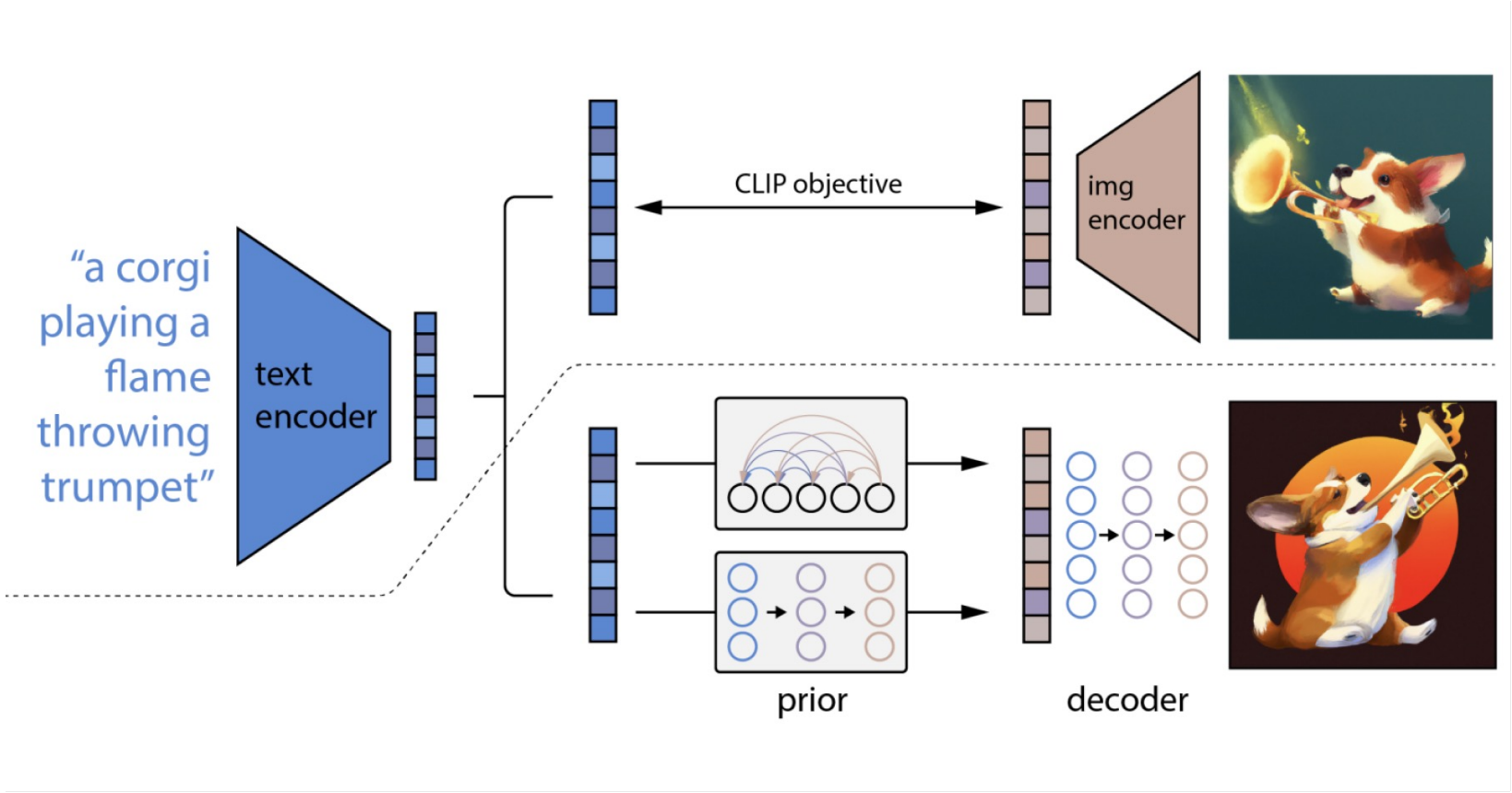
The Detail

RMIT UNIVERSITY

Stable Diffusion Architecture



Stable Diffusion Architecture



Week 5: Path Planning

RMIT UNIVERSITY