

CSCI 5541: Natural Language Processing

Lecture 17: Multi-Modal Models

James Mooney

With slides borrowed from Song Han (MIT)

What are MultiModal Models?

- ❑ Multi-Modal Models are models which incorporate multiple different modalities concurrently
- ❑ Typically, these modalities are speech, images, and text



Why Multi-modal?

- ❑ Advantages of using more than one modality
- ❑ There is untapped data
- ❑ Current text data (high quality) running low - need to explore more of the pool of data



Examples



Bunch of Examples

- ❑ Text-to-image (Dall-E/MidJourney/Imagen/Firefly/Stability/Flux/stablediff)
- ❑ Text-to-speech (Whisper/elevenlabs)
- ❑ Text-to-3d (Dreamcraft3d, openai/shap-e)
- ❑ Video-to-voice (elevenlabs)
- ❑ Text/Image-to-video (make-a-video/runway/openai-sora)



Imagen (Image Generation)



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.



A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.



A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.



Teddy bears swimming at the Olympics 400m Butterfly event.



A cute corgi lives in a house made out of sushi.



A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Ethics of Image Generation



Vision-Language Models (OpenAI)

User

Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



GPT-4

This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.

The text of the meme suggests that the image below is a beautiful picture of the earth from space.

However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.

The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Vision-Language Models (OpenAI)

Visual inputs: VGA charger

Sample 1 of 7

Next sample

User

What is funny about this image? Describe it panel by panel.



Source: [hmmm \(Reddit\)](#)

GPT-4

The image shows a package for a "Lightning Cable" adapter with three panels.

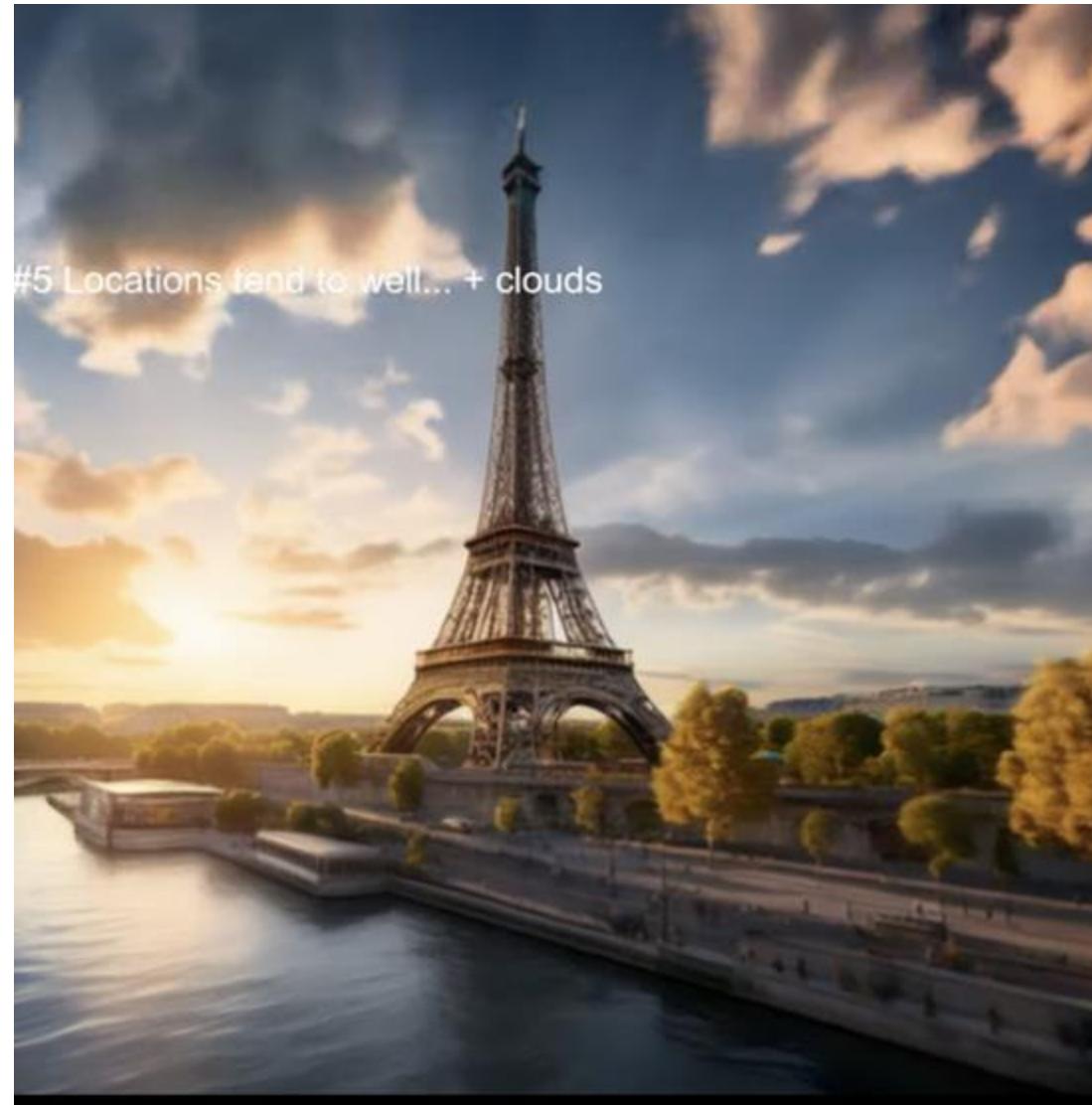
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Image-to-Video (RunwayML)



<https://www.youtube.com/shorts/CXG4vexCwe0>

Text-to-Video (Sora)

<https://www.youtube.com/watch?v=TU1gMl0I0kc>



Prompt: The camera directly faces colorful buildings in burano italy. An adorable dalmation

3D Asset Generation (MeshyAI)



VLM Background



VLM Preliminaries

- ❑ Self-Supervised Learning
- ❑ Vision Modelling
- ❑ VAE/VQ-VAE
- ❑ GANs
- ❑ Diffusion Models
- ❑ Speech Models



VLM Preliminaries

☐ **Self-Supervised Learning**

☐ Vision Modelling

☐ VAE/VQ-VAE

☐ GANs

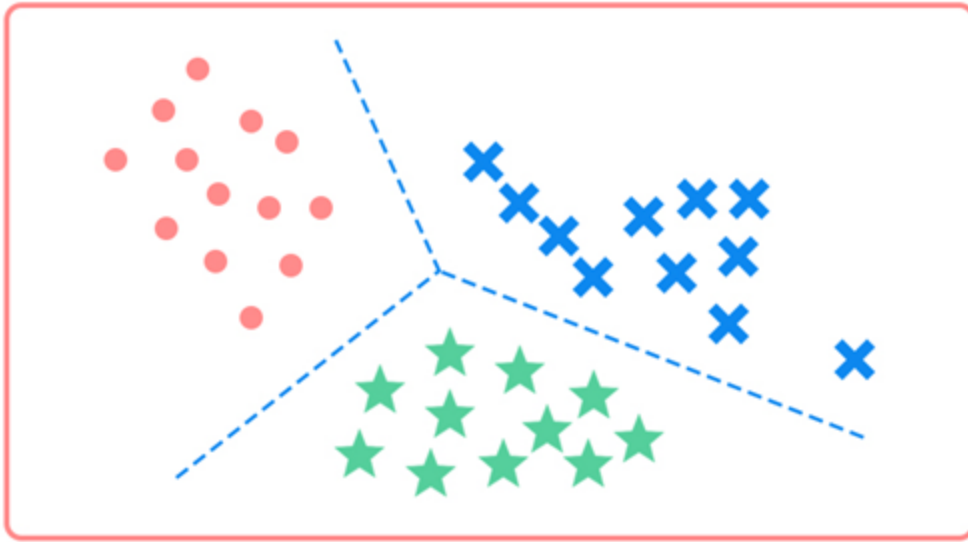
☐ Diffusion Models

☐ Speech Models



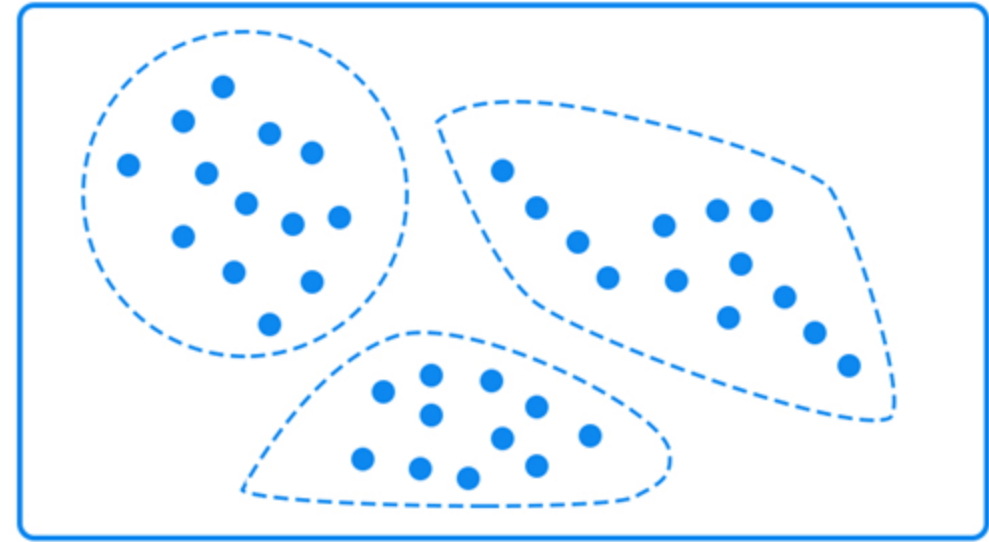
Revisiting Self-Supervised Learning

Classification



Supervised learning

Clustering



Unsupervised learning

Revisiting Self-Supervised Learning

Self Supervised Learning Overview

- ❑ Partition each input into two separate components.
- ❑ One component becomes your input and the other your output
- ❑ Use these generated labels to train your models

Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:
Hannah is a ____

Hannah is a *sister*
Hannah is a *friend*
Hannah is a *marketer*
Hannah is a *comedian*

Masked-language-modeling

The model is given a sequence of words with the goal of predicting a 'masked' word in the middle.

Example
Jacob [mask] reading

Jacob *fears* reading
Jacob *loves* reading
Jacob *enjoys* reading
Jacob *hates* reading



The Utility of Self-Supervised Learning

- ❑ Most data is not explicitly labelled
- ❑ Most data has associations to other data
 - Text to other portions of text
 - Web traffic between websites
 - Images and text from social media websites
- ❑ Very useful in Multi-Modal Settings



VLM Preliminaries

- ❑ Self-Supervised Learning
- ❑ **Vision Modelling**
- ❑ VAE/VQ-VAE
- ❑ GANs
- ❑ Diffusion Models
- ❑ Speech Models



Vision Modelling Background (Convolutions)

2	4	9	1	4
2	1	4	4	6
1	1	2	9	2
7	3	5	1	3
2	3	4	8	5

Image

x

1	2	3
-4	7	4
2	-5	1

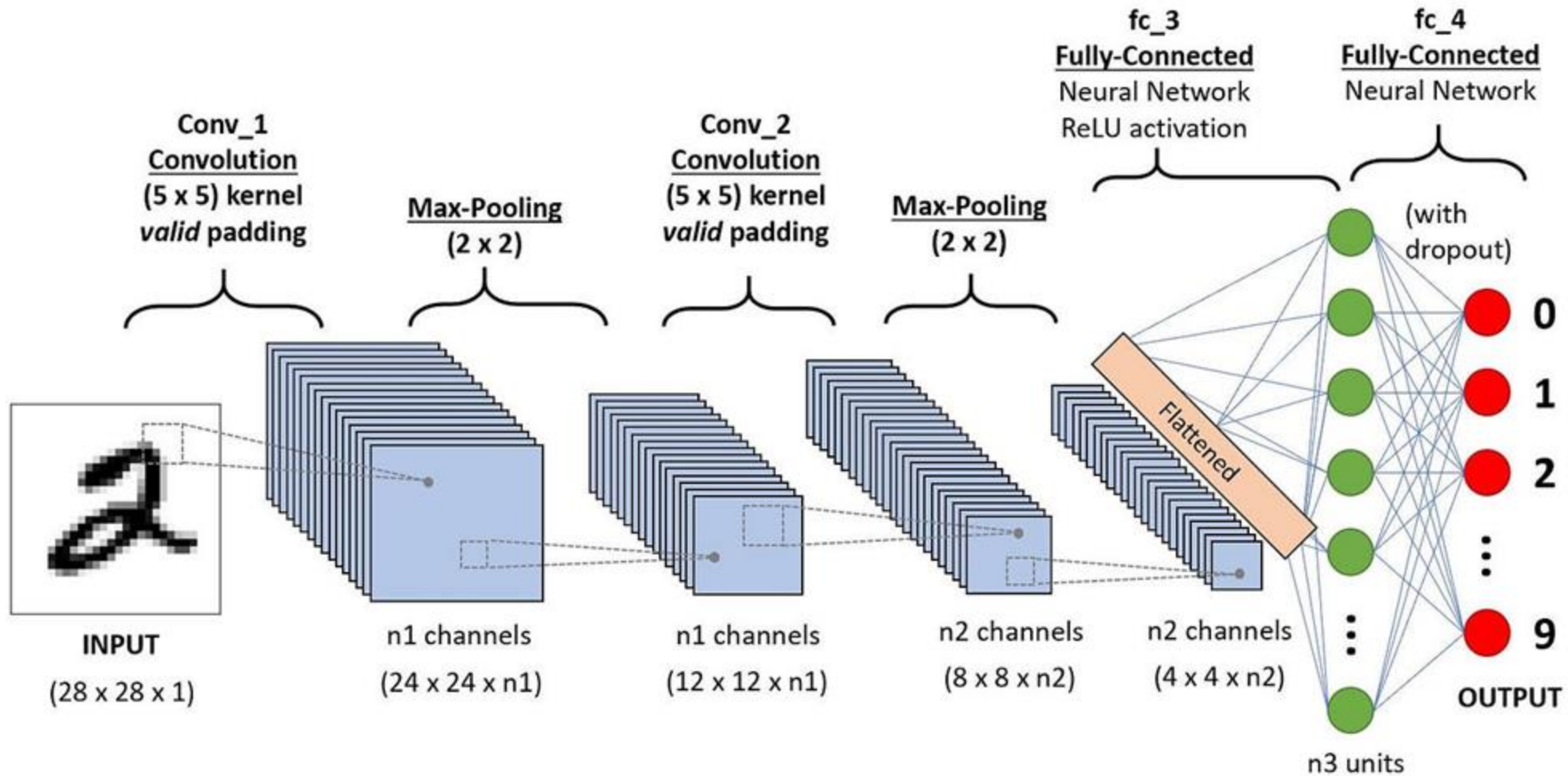
Filter /
Kernel

=

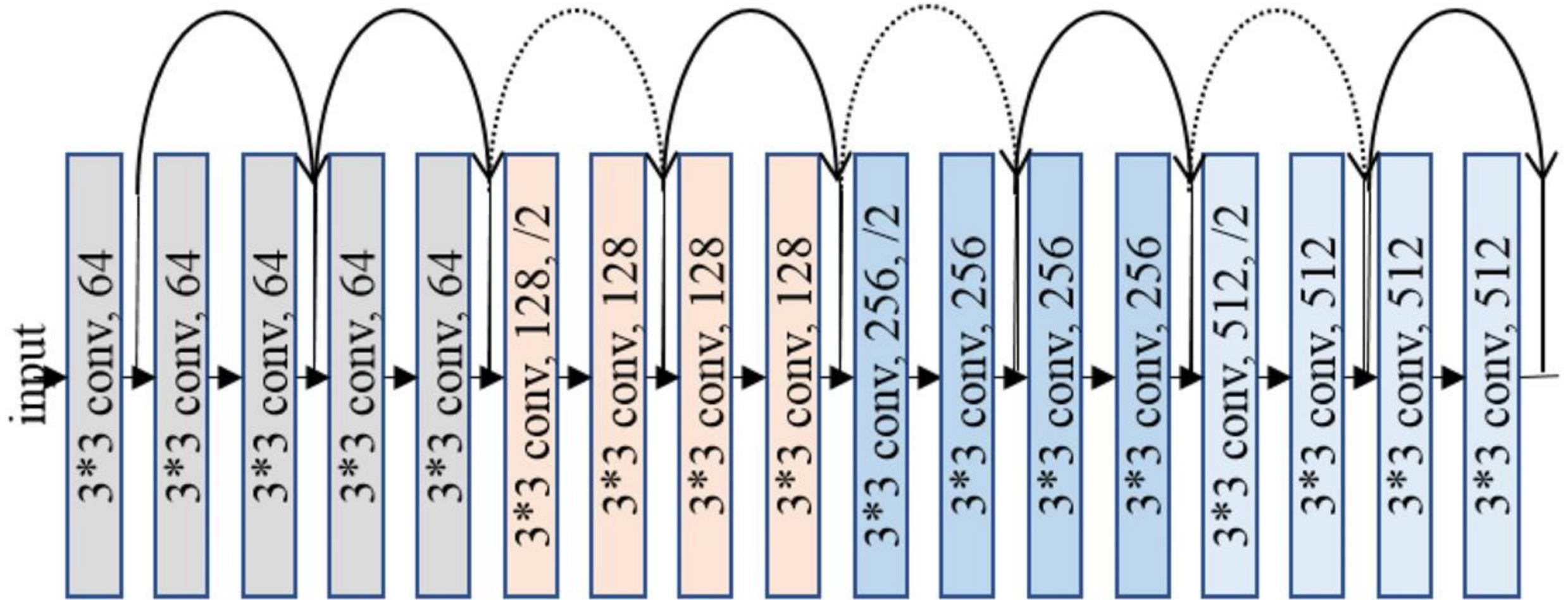
51	66	

Feature

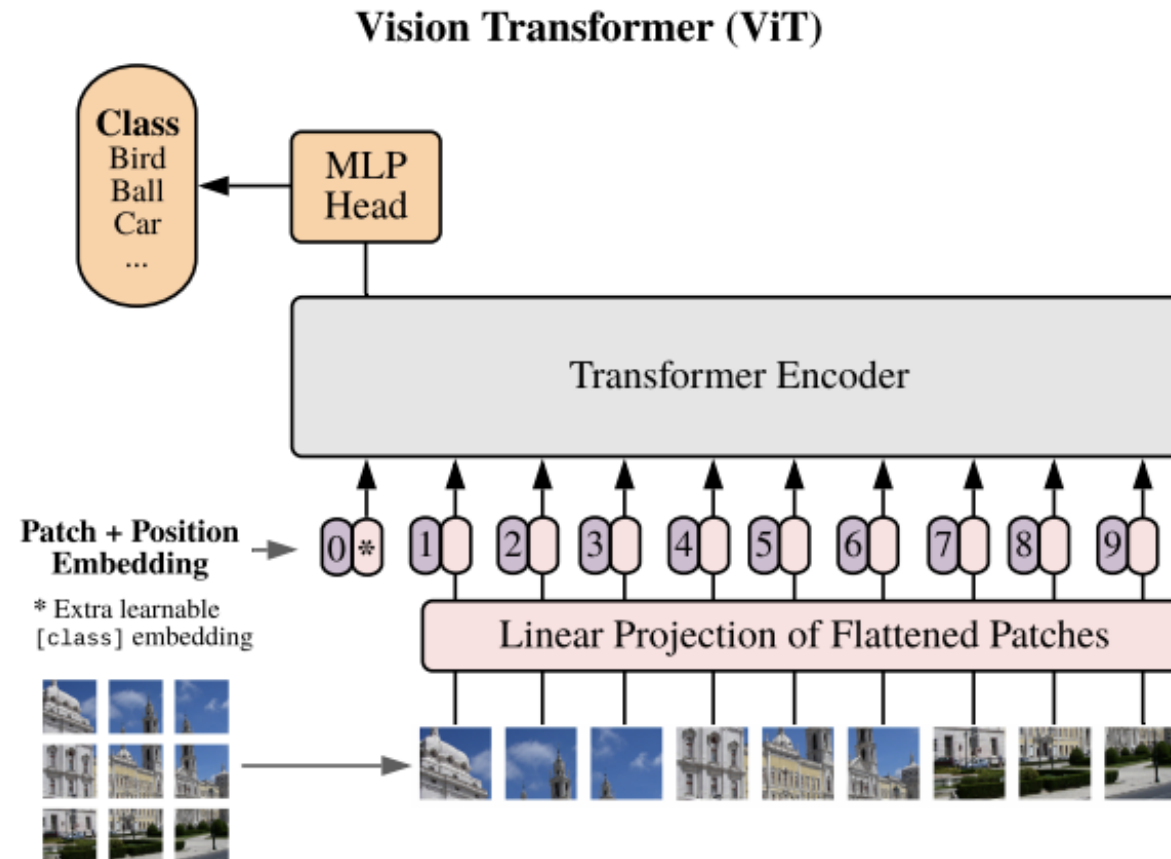
Vision Modelling Background (AlexNet)



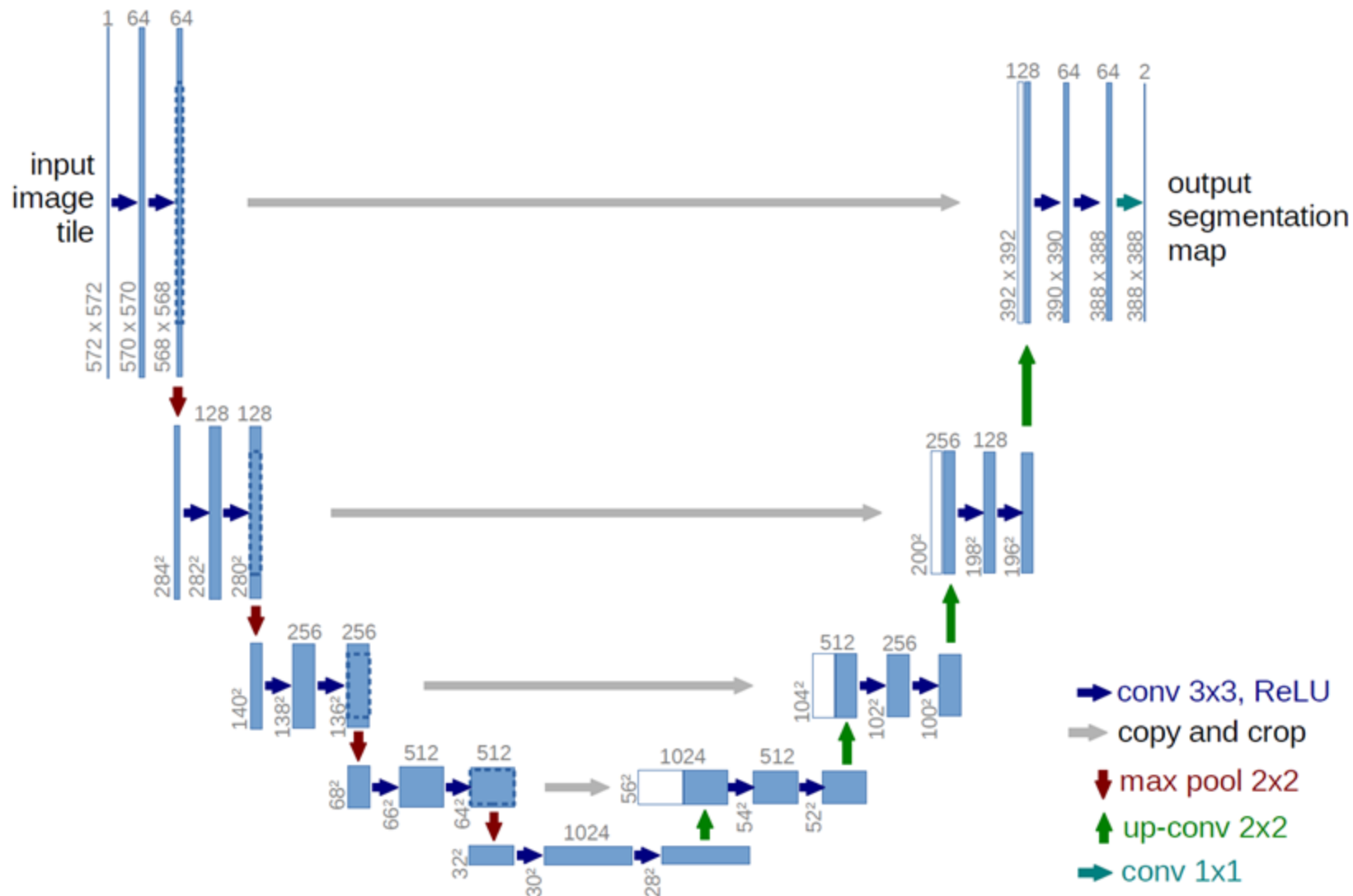
Vision Modelling Background (ResNet)



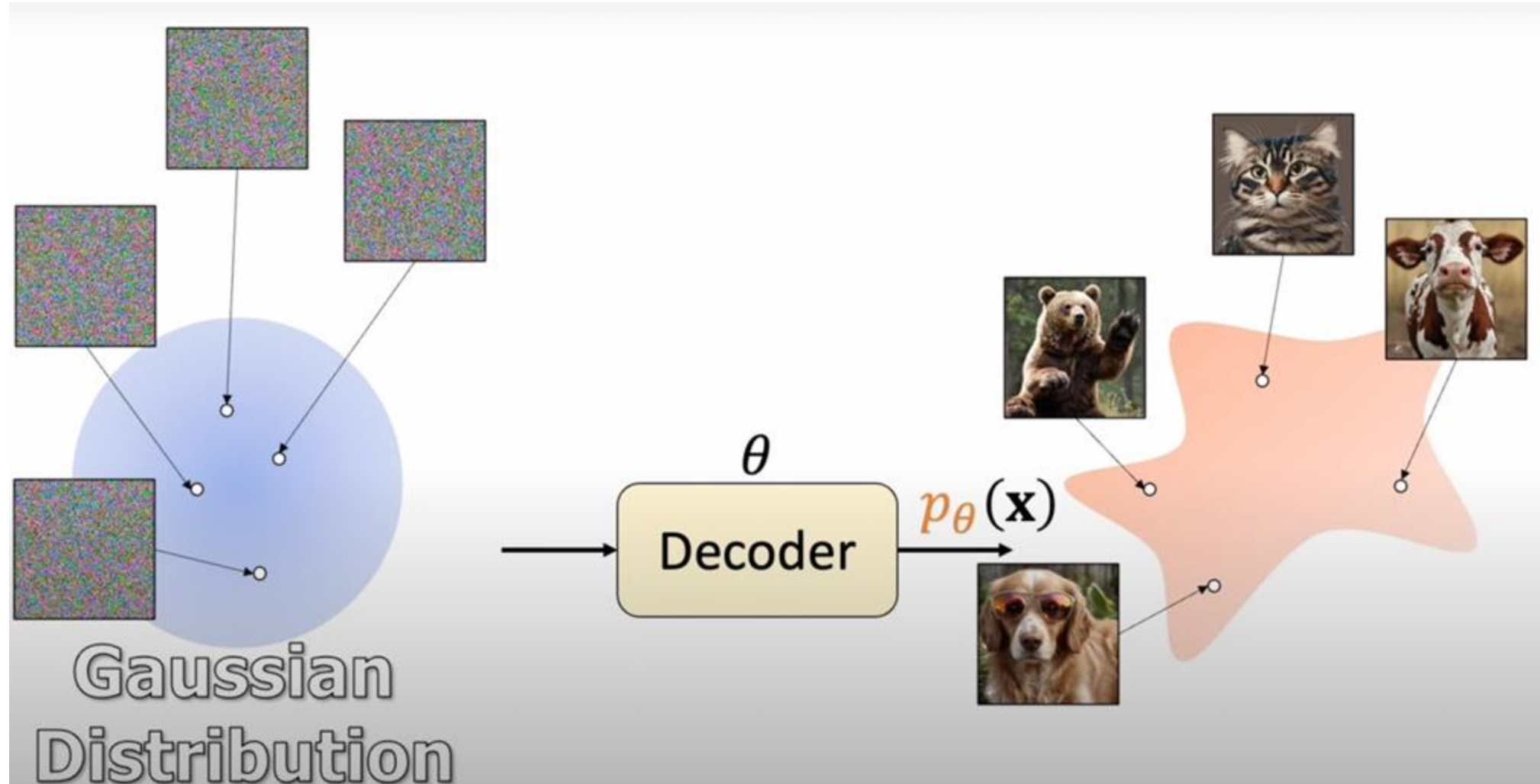
Vision Modelling Background (ViT)



Vision Modelling Background (UNet)



Vision Modelling Background (Generation)

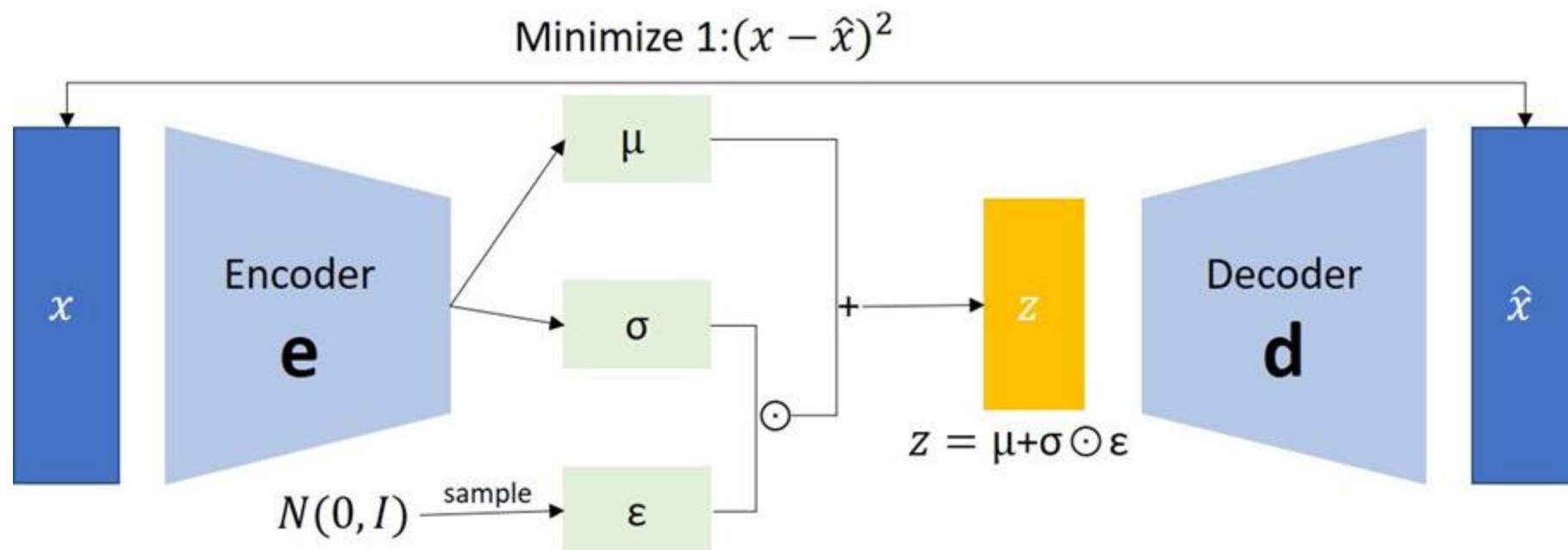


VLM Preliminaries

- ❑ Self-Supervised Learning
- ❑ Image Modelling
- ❑ **VAE/VQ-VAE**
- ❑ GANs
- ❑ Diffusion Models
- ❑ Speech Models



VAE



Minimize 2: $\frac{1}{2} \sum_{i=1}^N (\exp(\sigma_i) - (1 + \sigma_i) + \mu_i^2)$



VQ-VAE

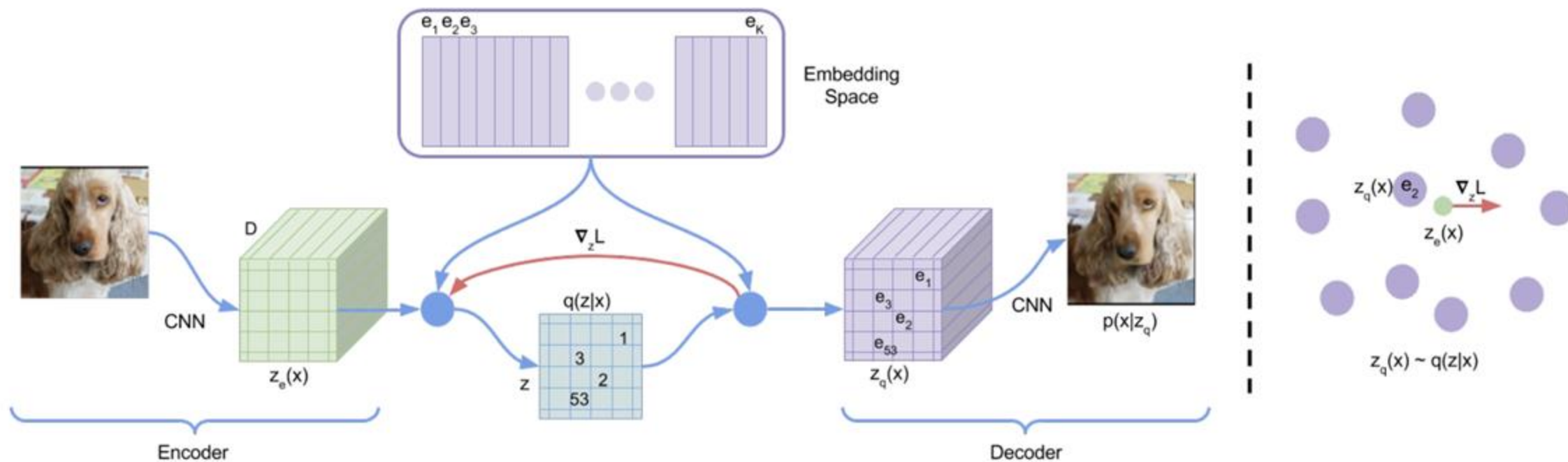


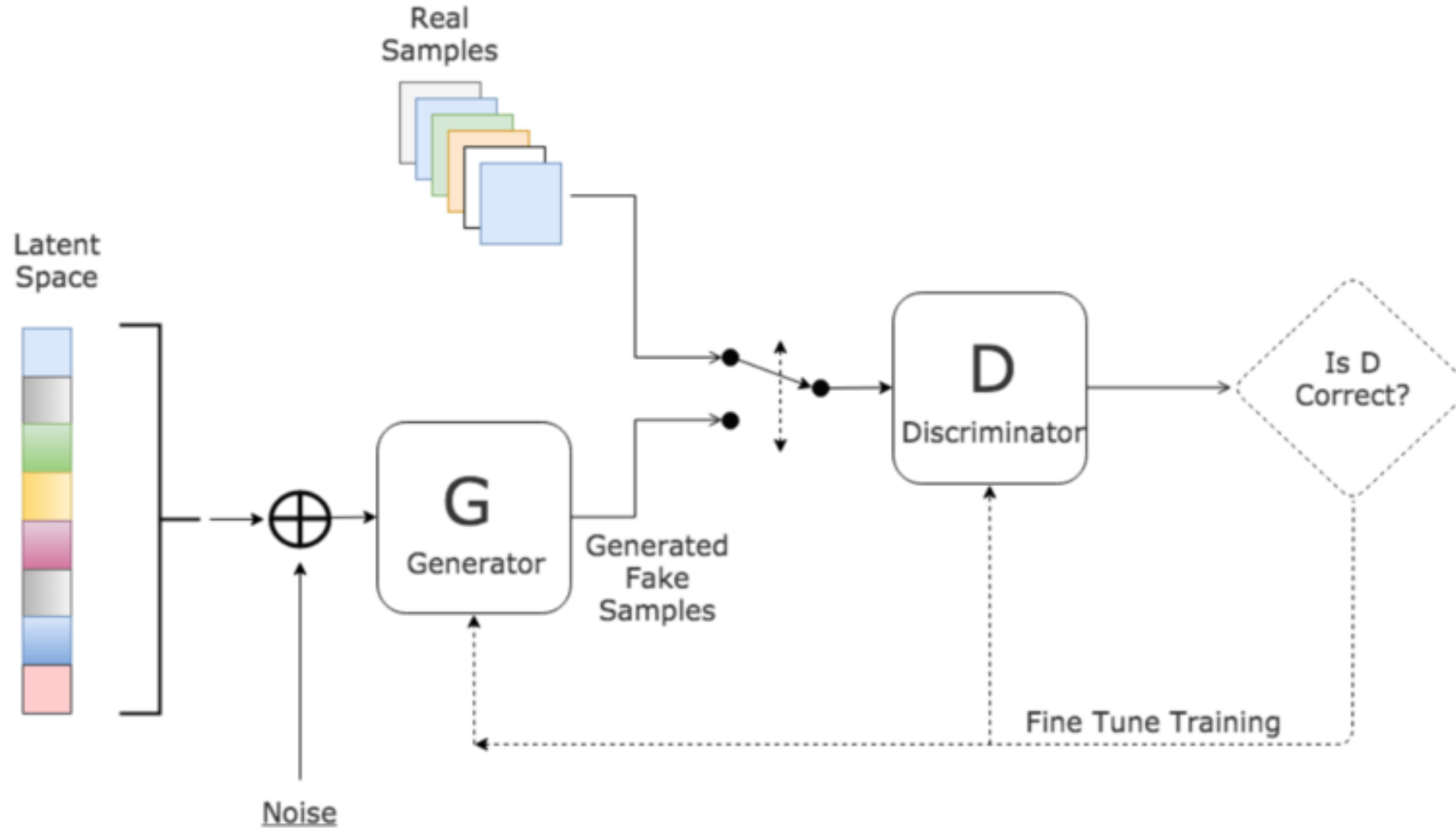
Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2 . The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

VLM Preliminaries

- ❑ Self-Supervised Learning
- ❑ Image Modelling
- ❑ VAE/VQ-VAE
- ❑ **GANs**
- ❑ Diffusion Models
- ❑ Speech Models



Generative Adversarial Network

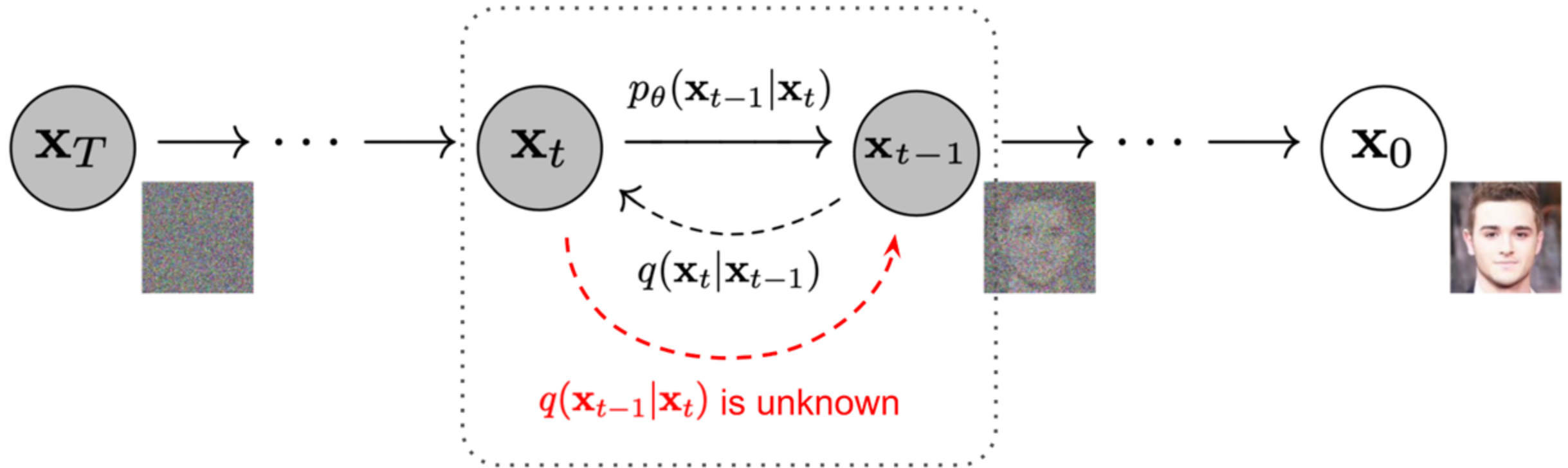


VLM Preliminaries

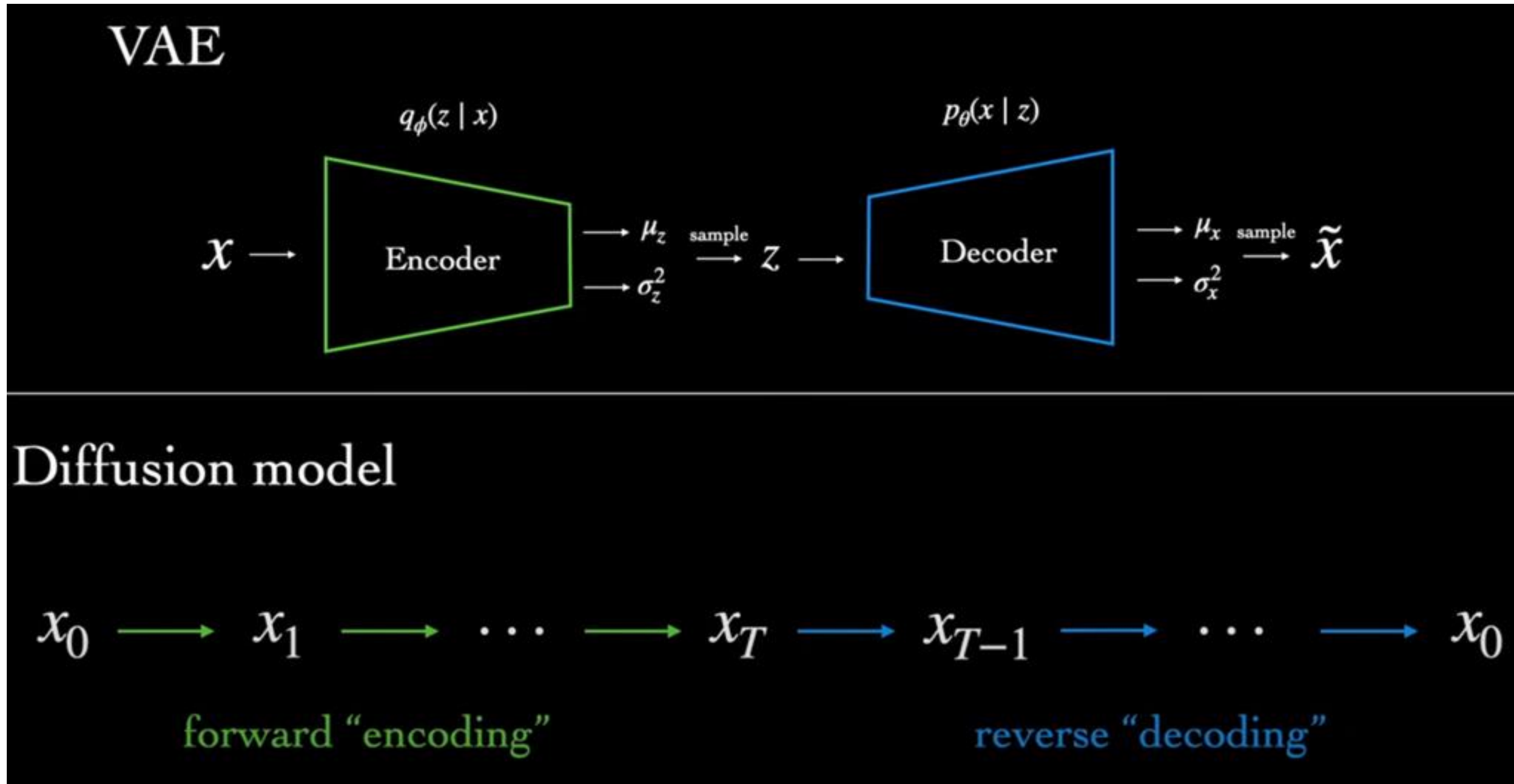
- ☐ Self-Supervised Learning
- ☐ Image Modelling
- ☐ VAE/VQ-VAE
- ☐ GANs
- ☐ **Diffusion Models**
- ☐ Speech Models



Diffusion Models



Diffusion Models

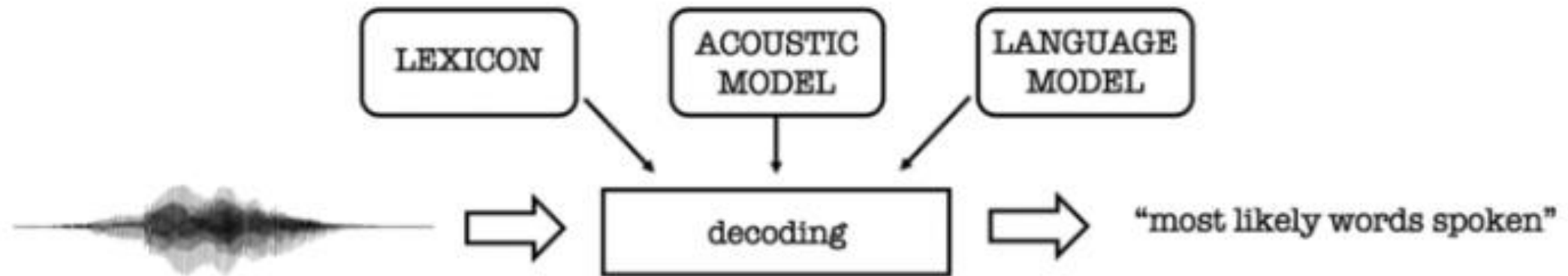


VLM Preliminaries

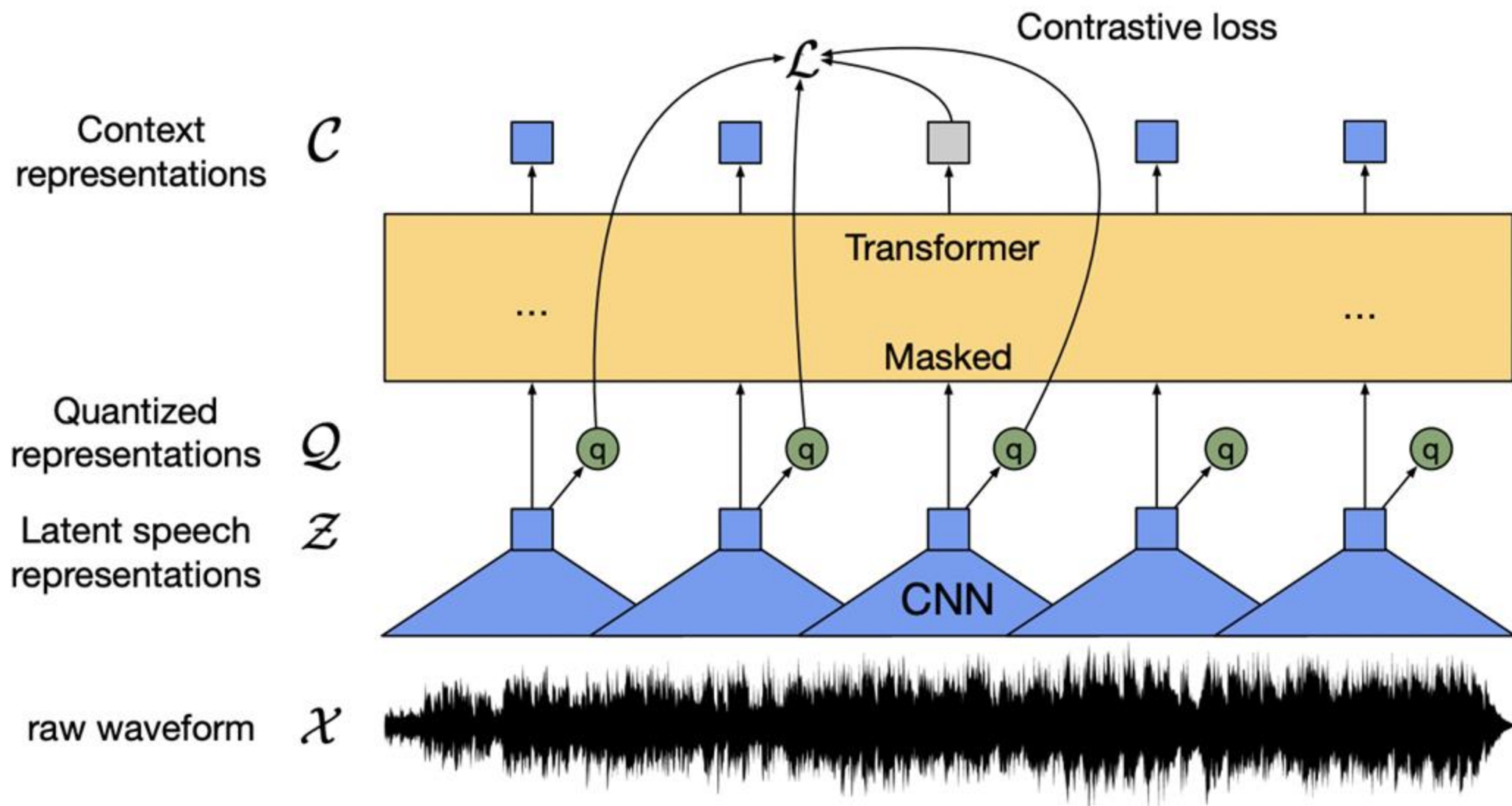
- ☐ Self-Supervised Learning
- ☐ Image Modelling
- ☐ VAE/VQ-VAE
- ☐ GANs
- ☐ Diffusion Models
- ☐ **Speech Models**



Automatic Speech Recognition

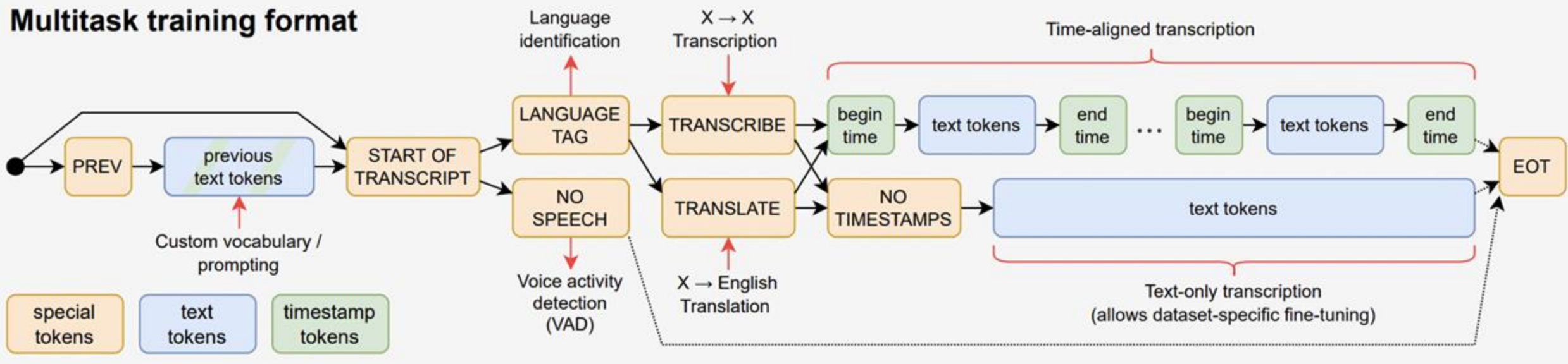


Wav2Vec



Whisper

Multitask training format



Whisper

Multitask training data (680k hours)

English transcription

🗣️ "Ask not what your country can do for ..."

📄 Ask not what your country can do for ...

Any-to-English speech translation

🗣️ "El rápido zorro marrón salta sobre ..."

📄 The quick brown fox jumps over ...

Non-English transcription

🗣️ "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."

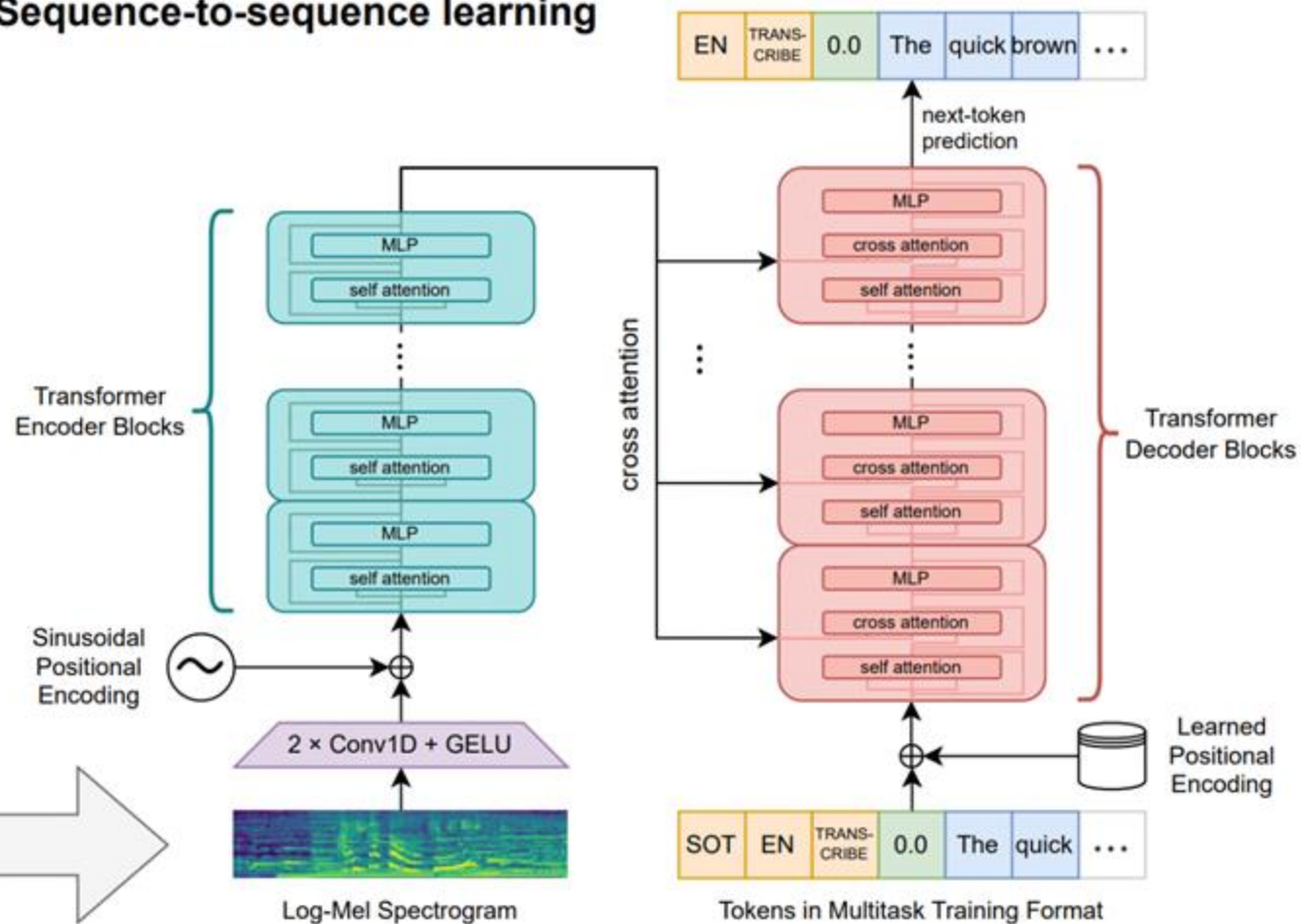
📄 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

No speech

🔊 (background music playing)

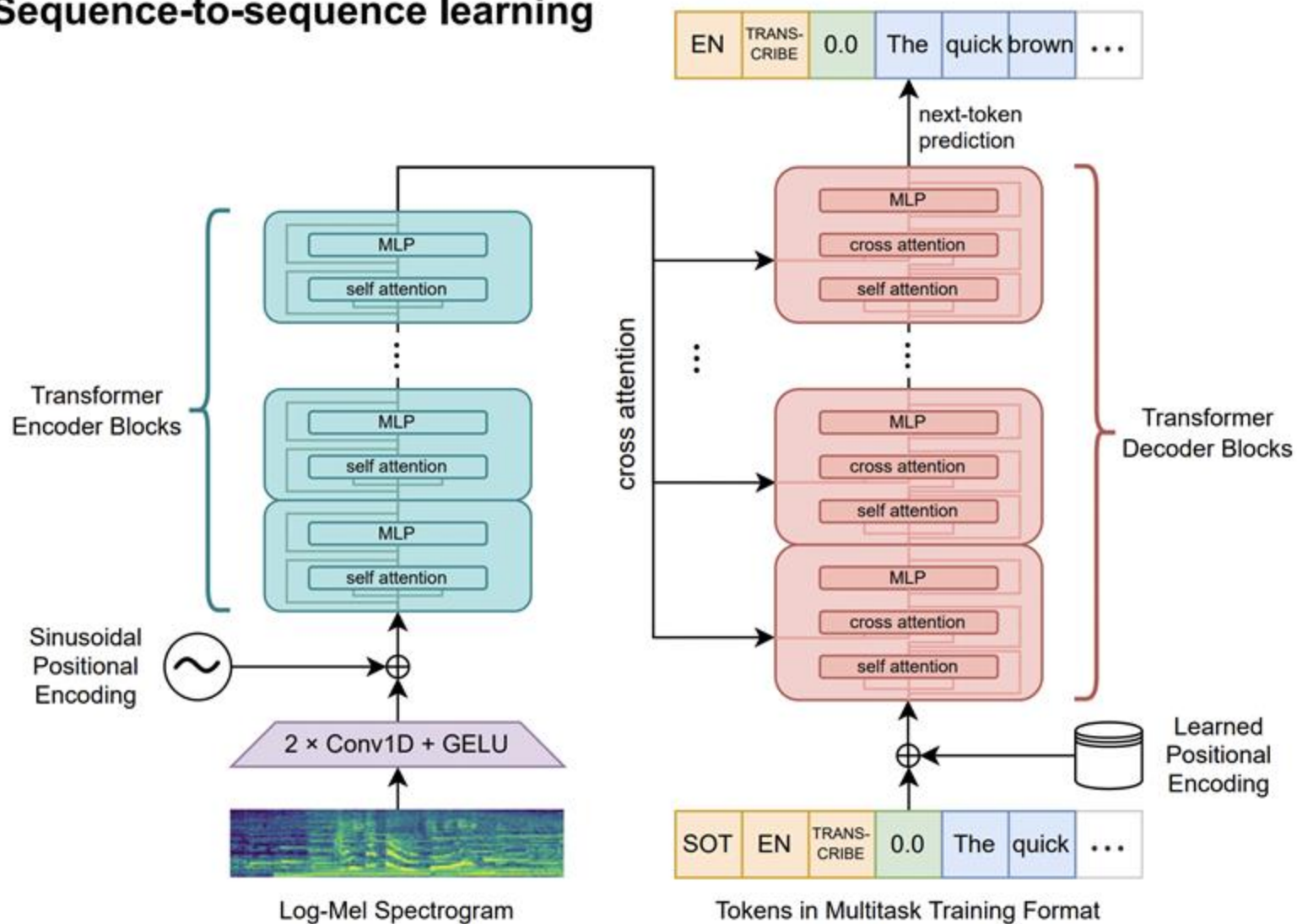
📄 ∅

Sequence-to-sequence learning



Whisper

Sequence-to-sequence learning



Vision Language Models



VLM Categories

- ☐ Contrastive
- ☐ Masking
- ☐ Generation



VLM Categories

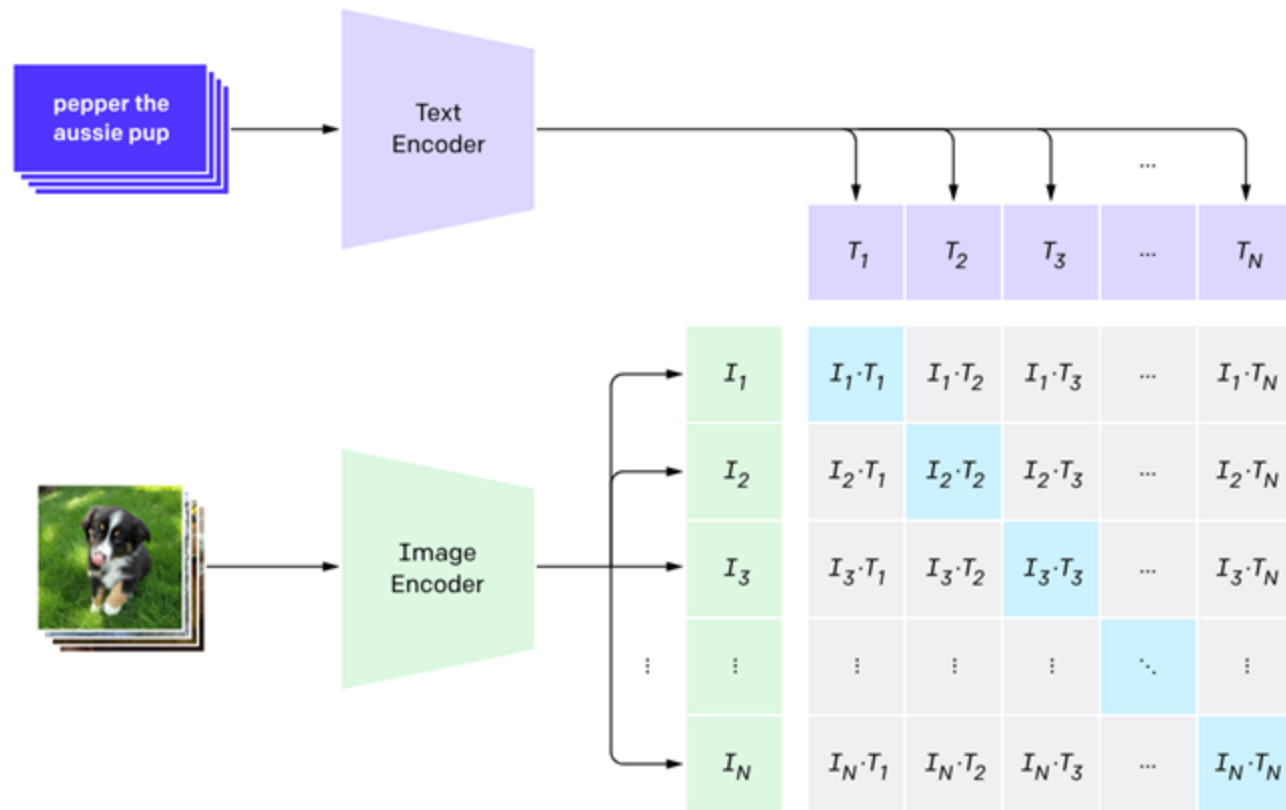
- ☐ **Contrastive**
- ☐ Masking
- ☐ Generation



Contrastive VLM

- ❑ We have access to both images and corresponding text
- ❑ Feed a batch of N images and their corresponding N texts through an image and text encoder, separately
- ❑ Calculate similarity between each combination of encoded images/texts

1. Contrastive pre-training

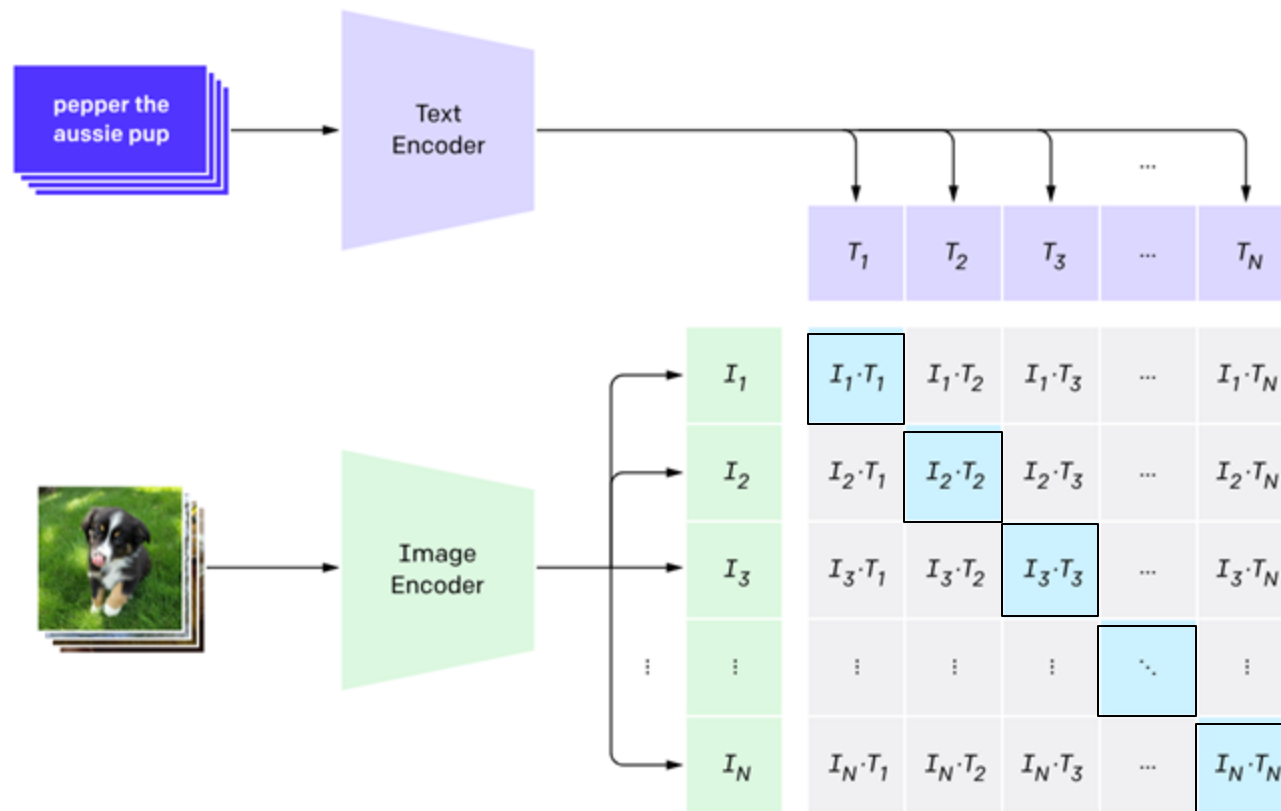


Contrastive VLM

- ❑ We have access to both images and corresponding text
- ❑ Feed a batch of N images and their corresponding N texts through an image and text encoder, separately
- ❑ Calculate similarity between each combination of encoded images/texts

Maximize Diagonal Entries

1. Contrastive pre-training

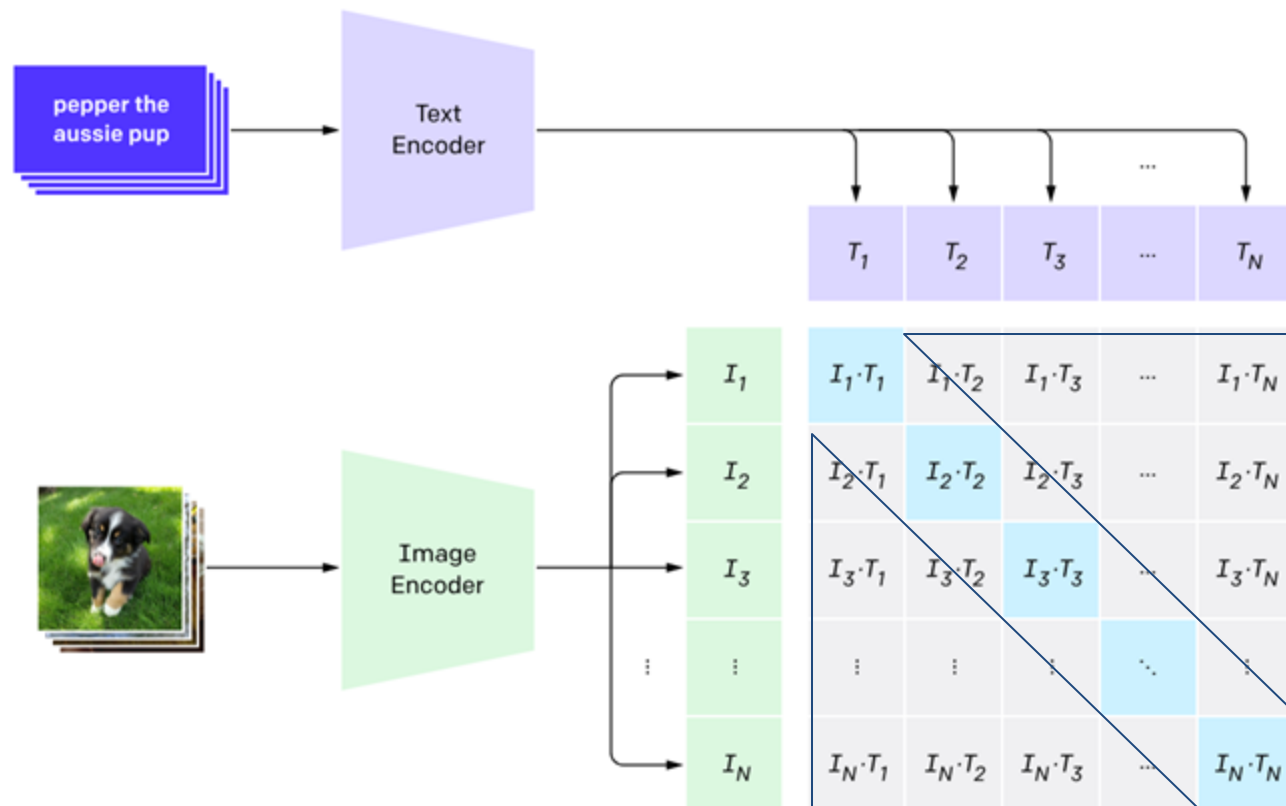


Contrastive VLM

- ❑ We have access to both images and corresponding text
- ❑ Feed a batch of N images and their corresponding N texts through an image and text encoder, separately
- ❑ Calculate similarity between each combination of encoded images/texts

Minimize Non-Diagonal Entries

1. Contrastive pre-training



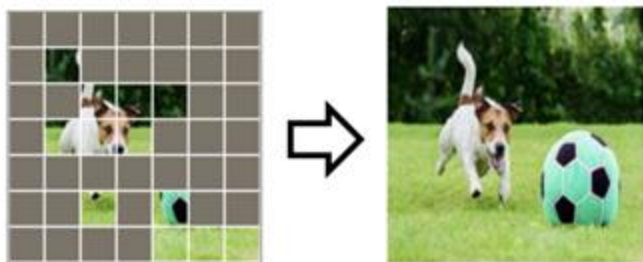
VLM Categories

- ☐ Contrastive
- ☐ **Masking**
- ☐ Generation

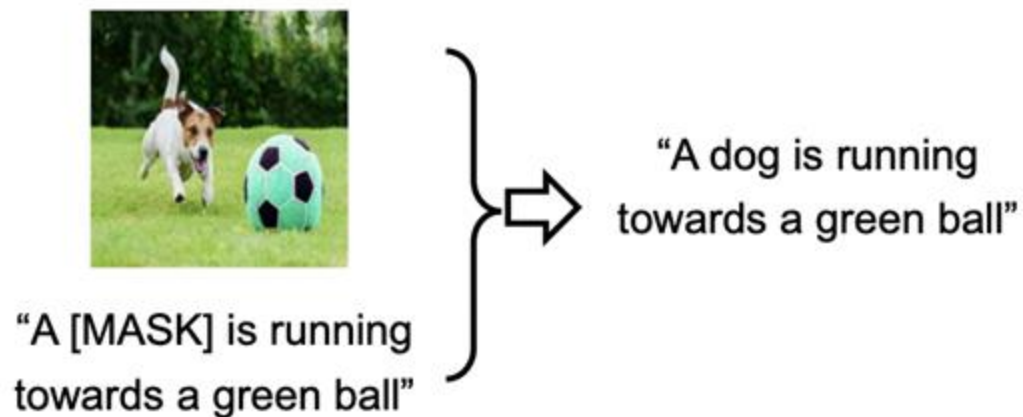


Masked VLM

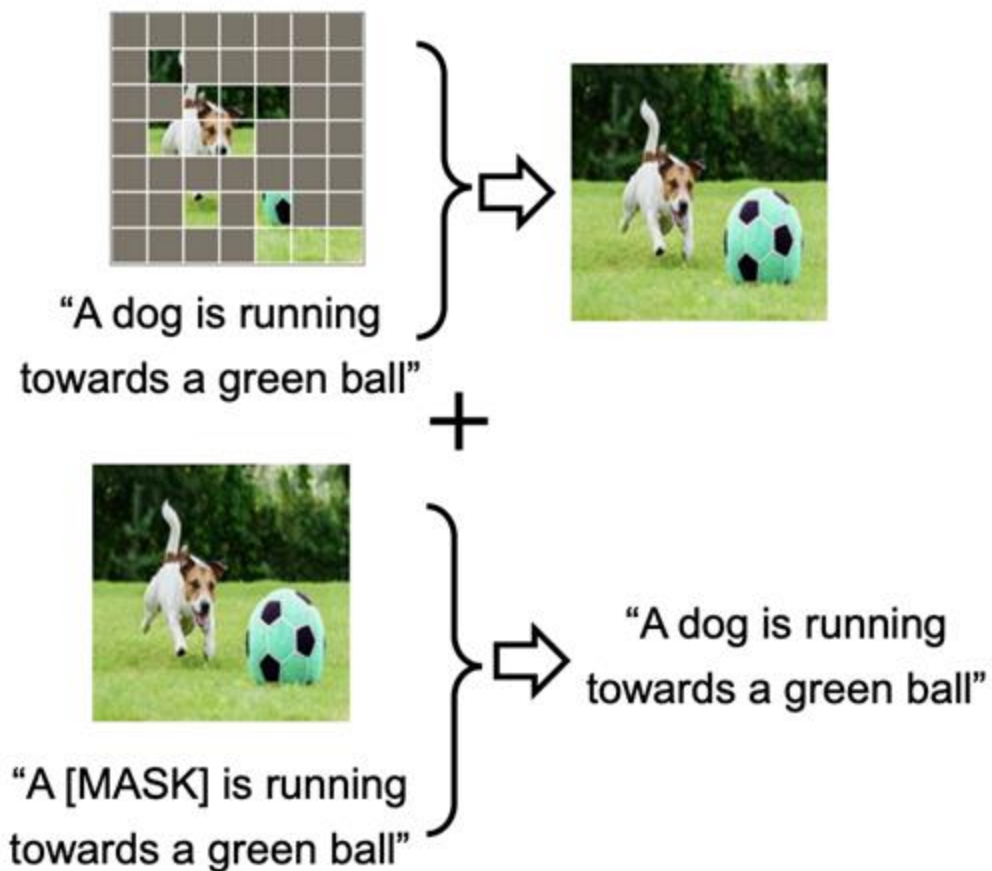
Masked Image Modeling



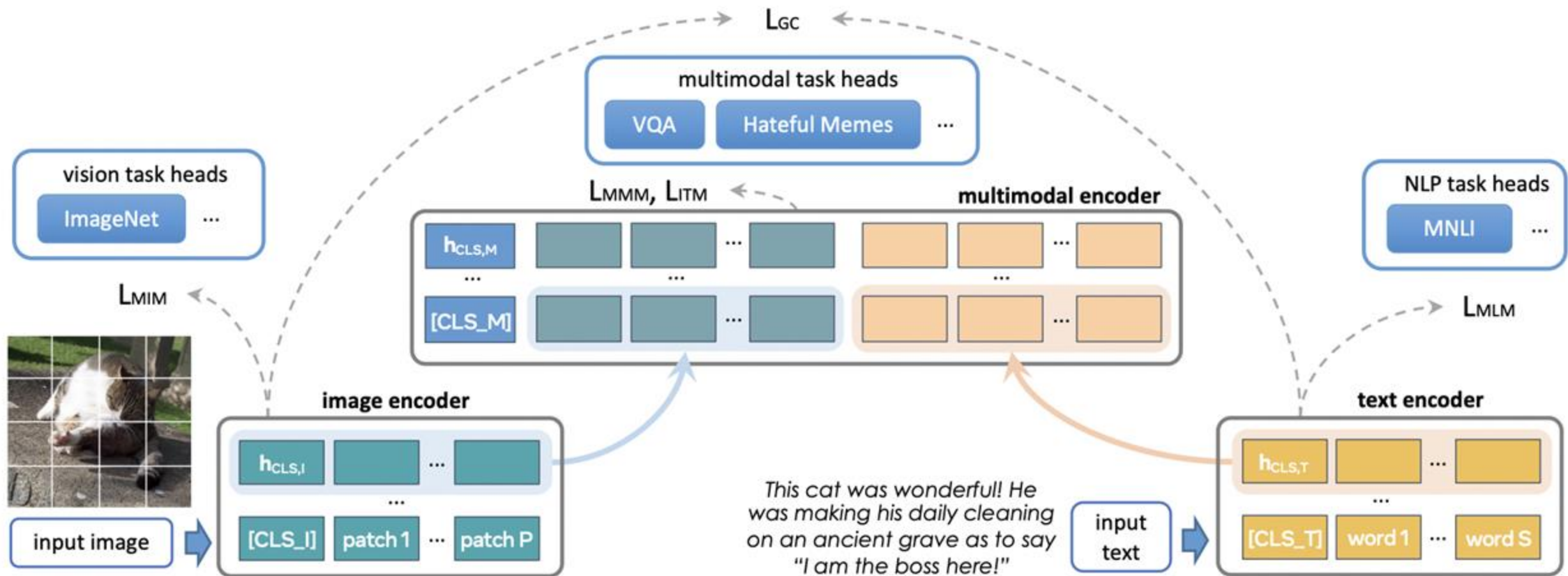
Masked Language Modeling in V+L Learning



Masked Vision and Language Modeling



Masked VLM

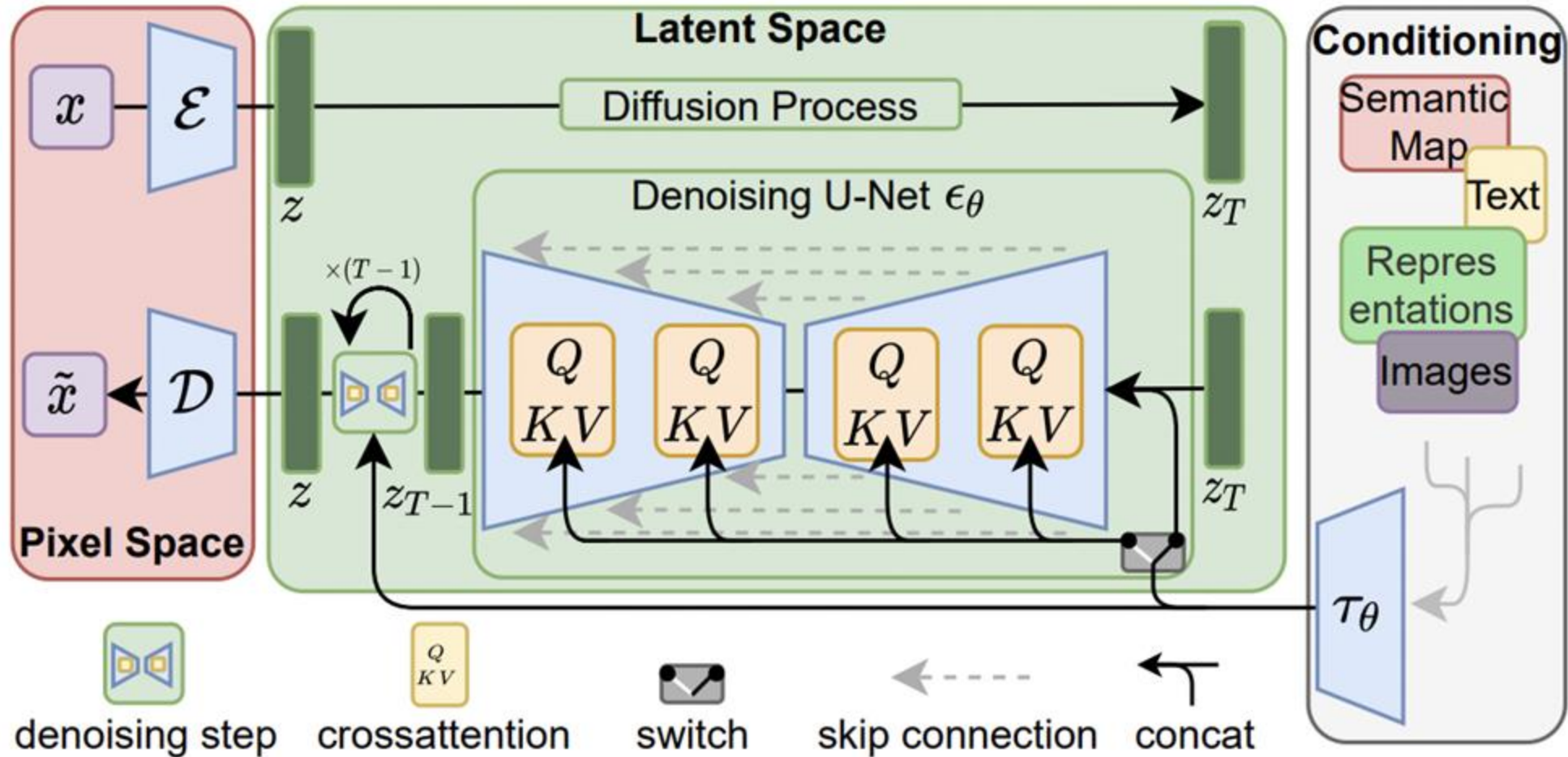


VLM Categories

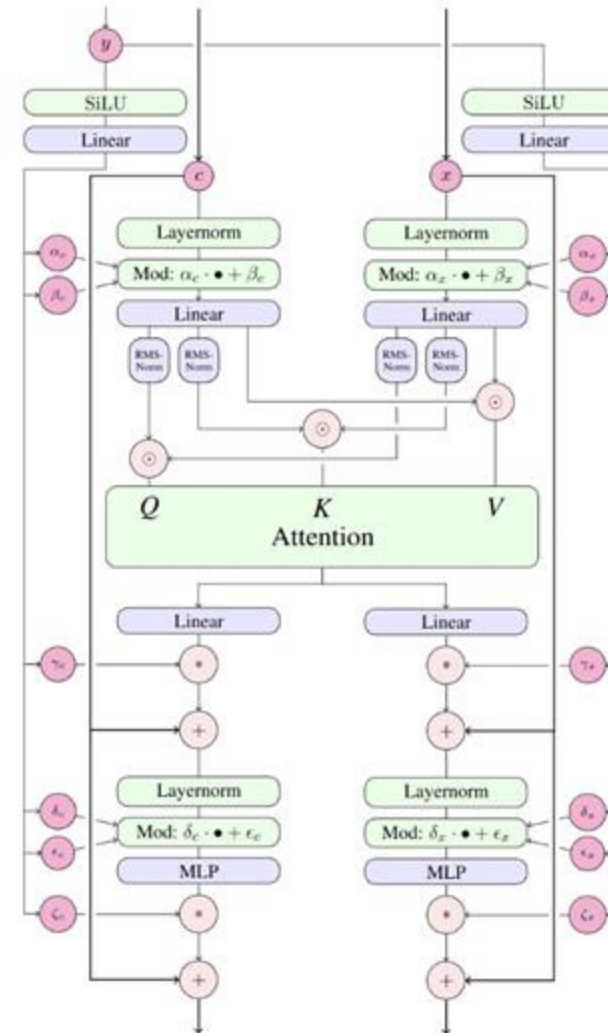
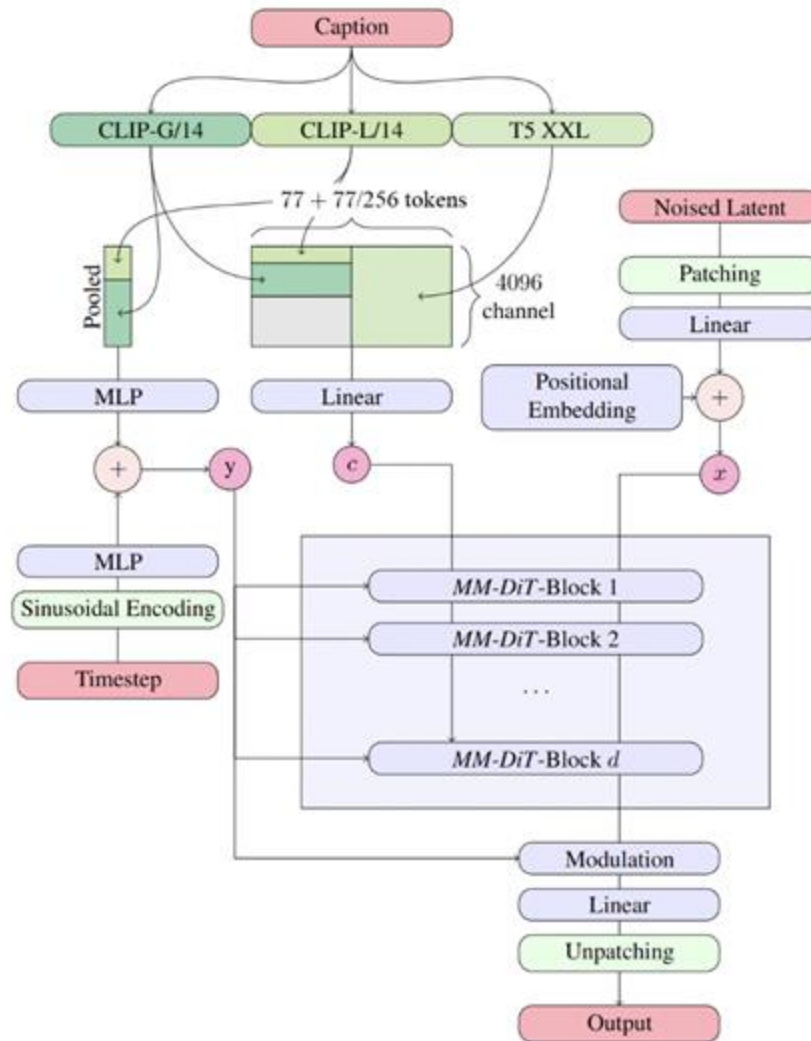
- ☐ Contrastive
- ☐ Masking
- ☐ **Generation**



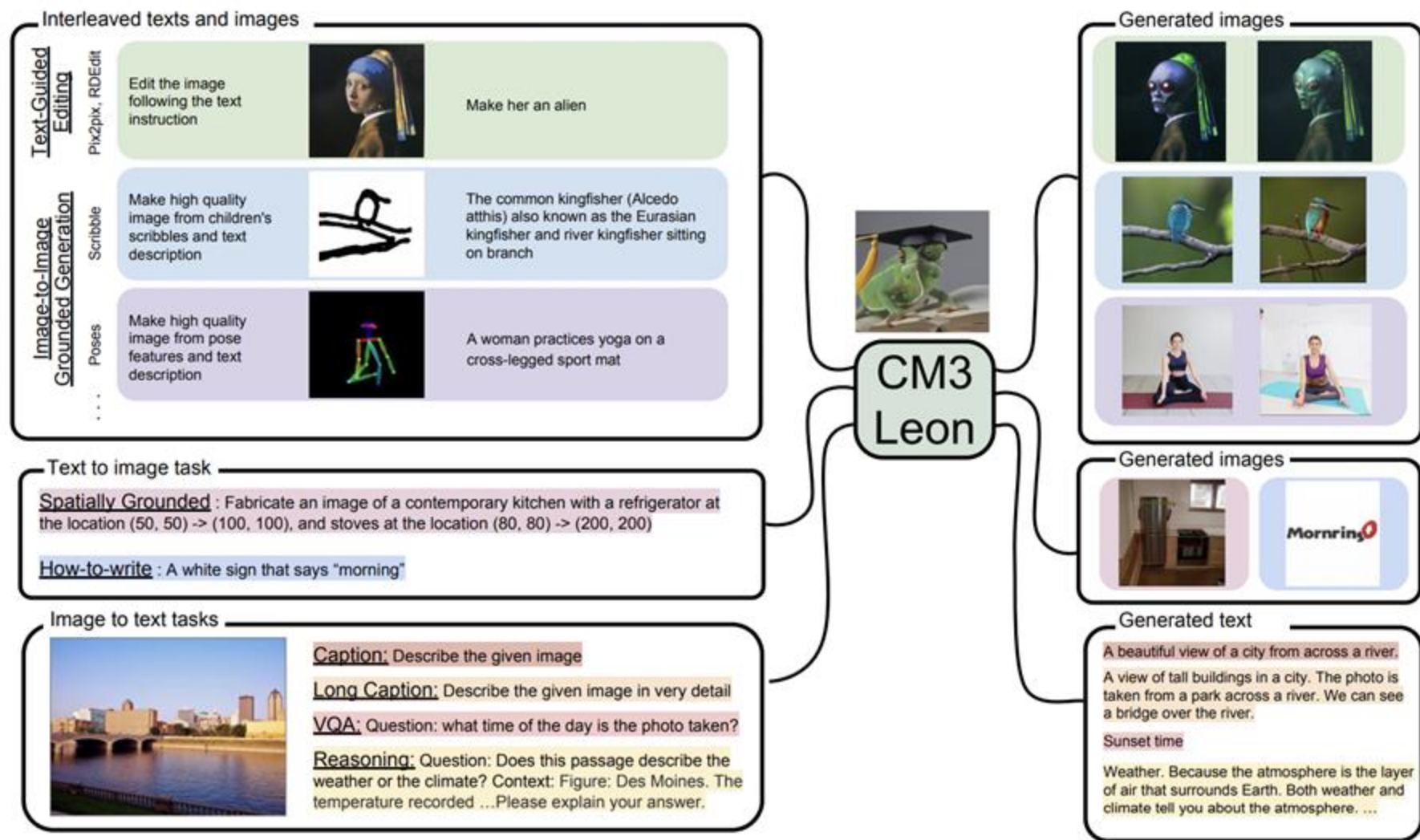
VLMs for Generation



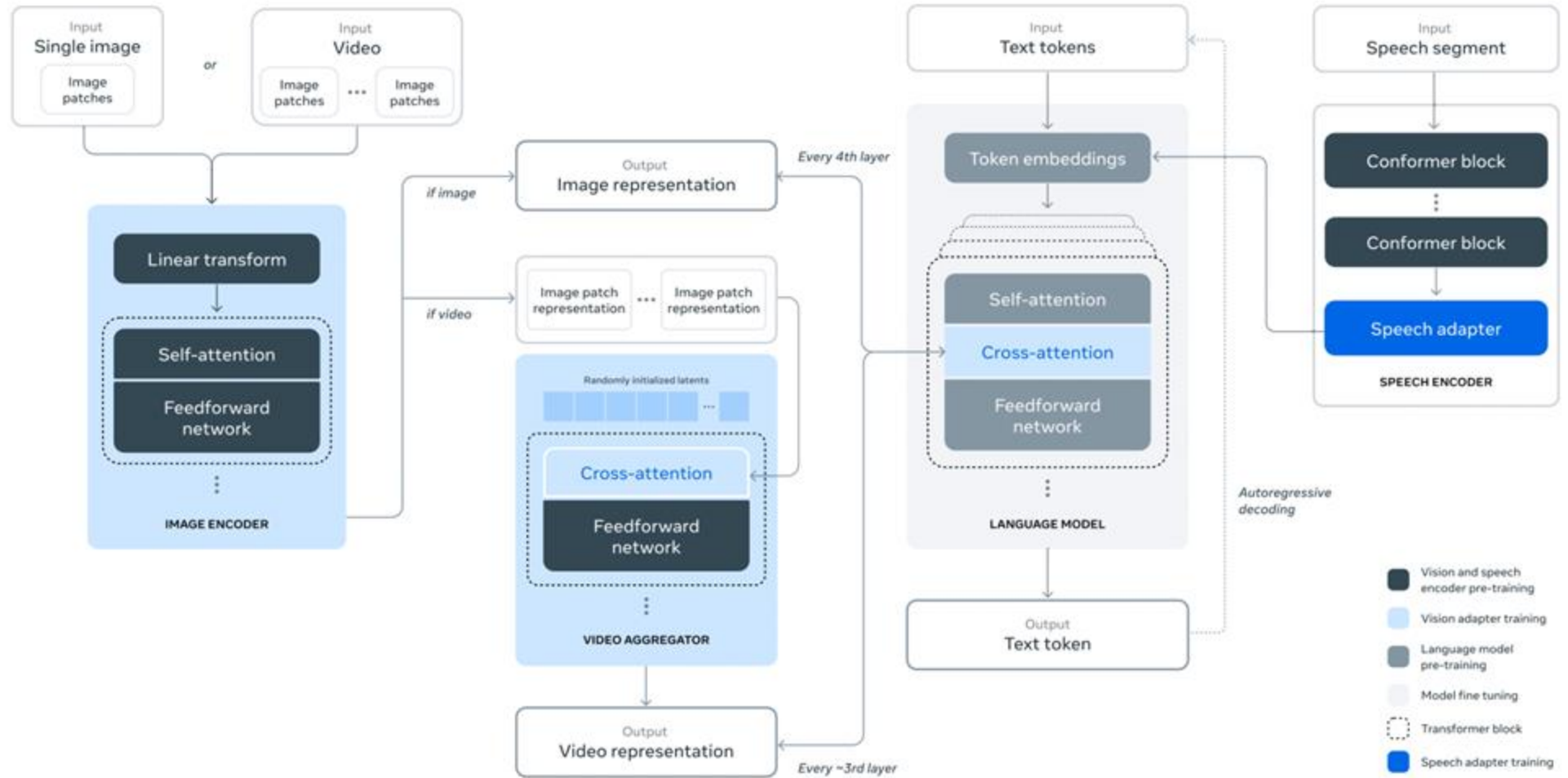
VLMs for Generation



VLMs for Generation



VLMs for Generation



Summary



Summary

- ❑ Multi-modal models can provide use across speech, vision, and text
- ❑ Different modalities and domains have, so far, required different approaches to get the best results
- ❑ There is **a lot** of data that is left unused in vision and speech that will go into training many of these models in the coming years
- ❑ On more complex tasks such as video generation, we are limited in the outputs and the gap with respect to the real world is currently larger - this will be an area where researchers will be investigating deeply in the coming years

