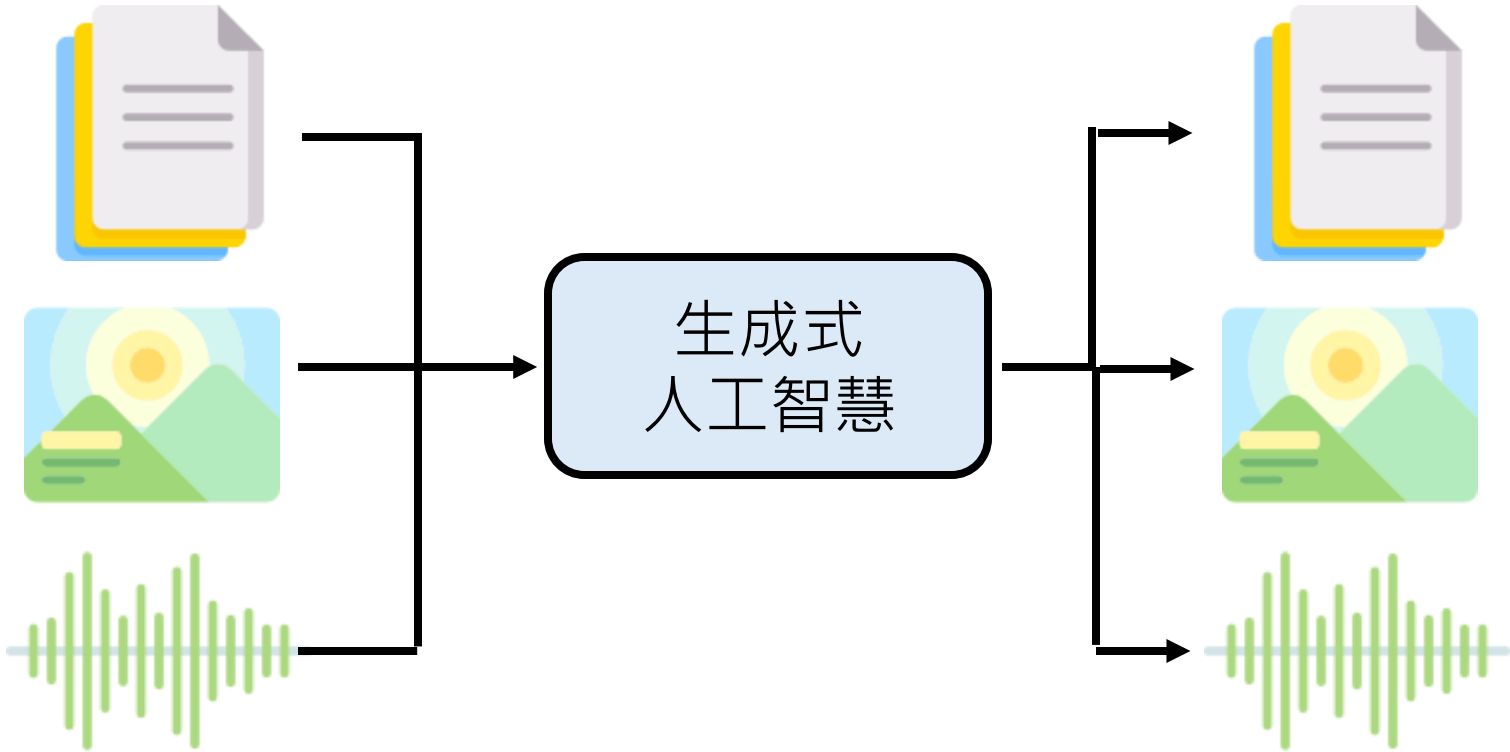
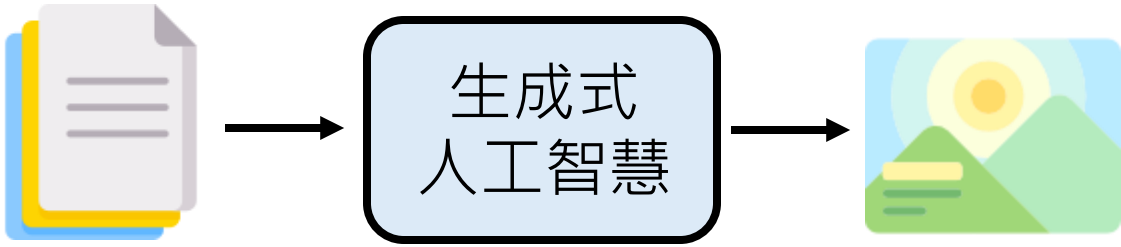
影像和聲音上的生成策略

李宏毅

到目前為止的生成都是文字接龍 ...

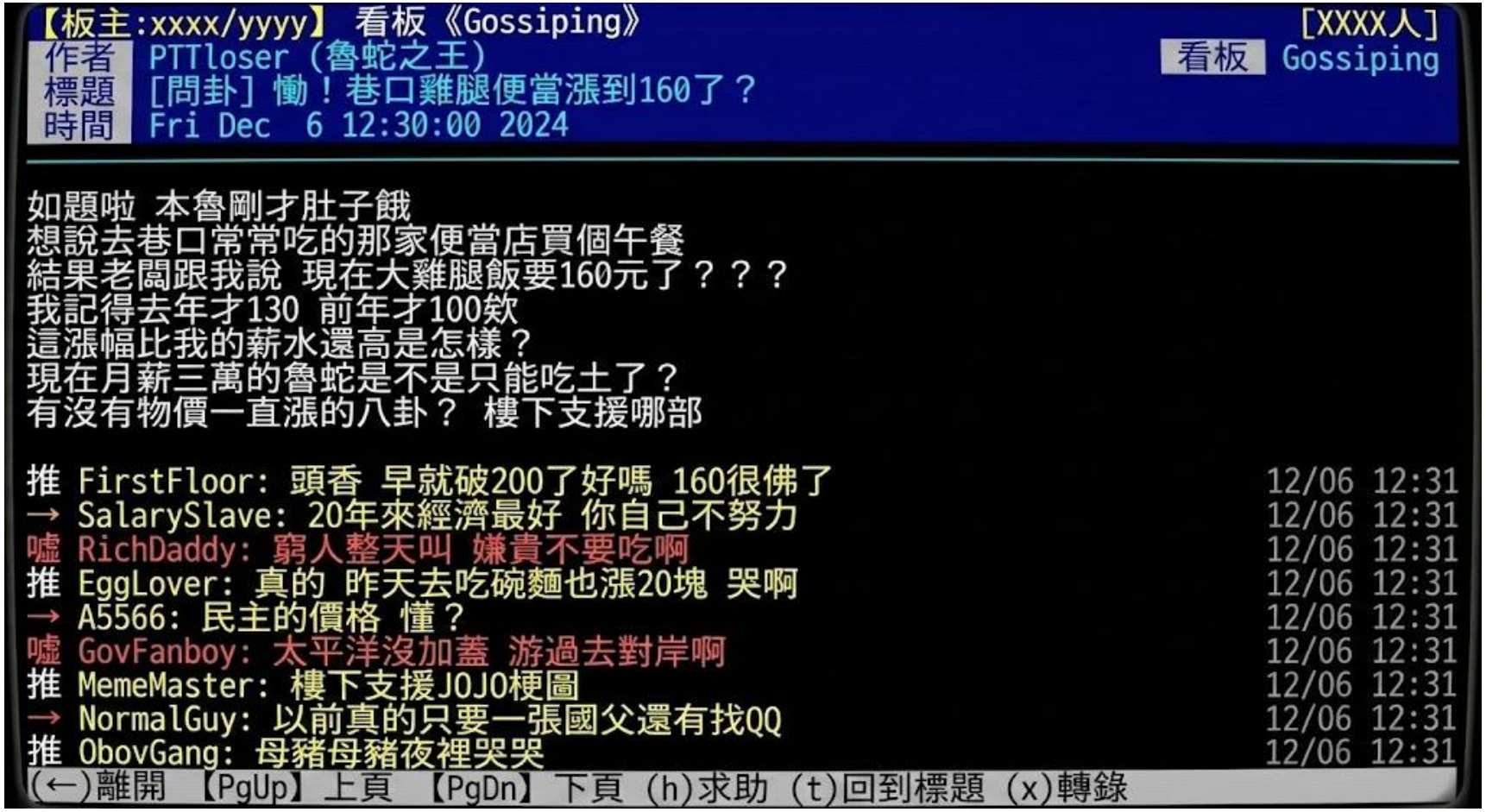


輸入文字、輸出圖片



- Gemini (Nano Banana)

“產生 PTT 八卦版鄉民貼文截圖”



(本圖由 Gemini 完全自動產生，不代表授課教師立場)

Problem 1
(7%) Compute $\det A$, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}.$$

Solution:

Problem 2
(7%) Let V and W be two subspaces of $\mathbb{R}^n$. Define $V + W = \{\mathbf{x} | \mathbf{x} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in W\}$. Prove that $V + W$ is a subspace of $\mathbb{R}^n$.

Solution:

生成式
人工智慧

Gemini (Nano Banana)

給我這份考卷被考試學生完成後的樣子

生成式
人工智慧

Gemini (Nano Banana)

Problem 1
(7%) Compute det A, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Solution: R_2-R_1, R_3-R_1, R_4-R_1, R_5-R_1

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$\det A = 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 = 1$

Problem 2
(7%) Let V and W be two subspaces of $\mathbb{R}^n$. Define $V + W = \{\mathbf{x} | \mathbf{x} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in W\}$. Prove that $V + W$ is a subspace of $\mathbb{R}^n$.

Solution: 1. Zero vector: Since V and W are subspaces, $0 \in V$ and $0 \in W$.
Thus, $0 = 0 + 0 \in V + W$.

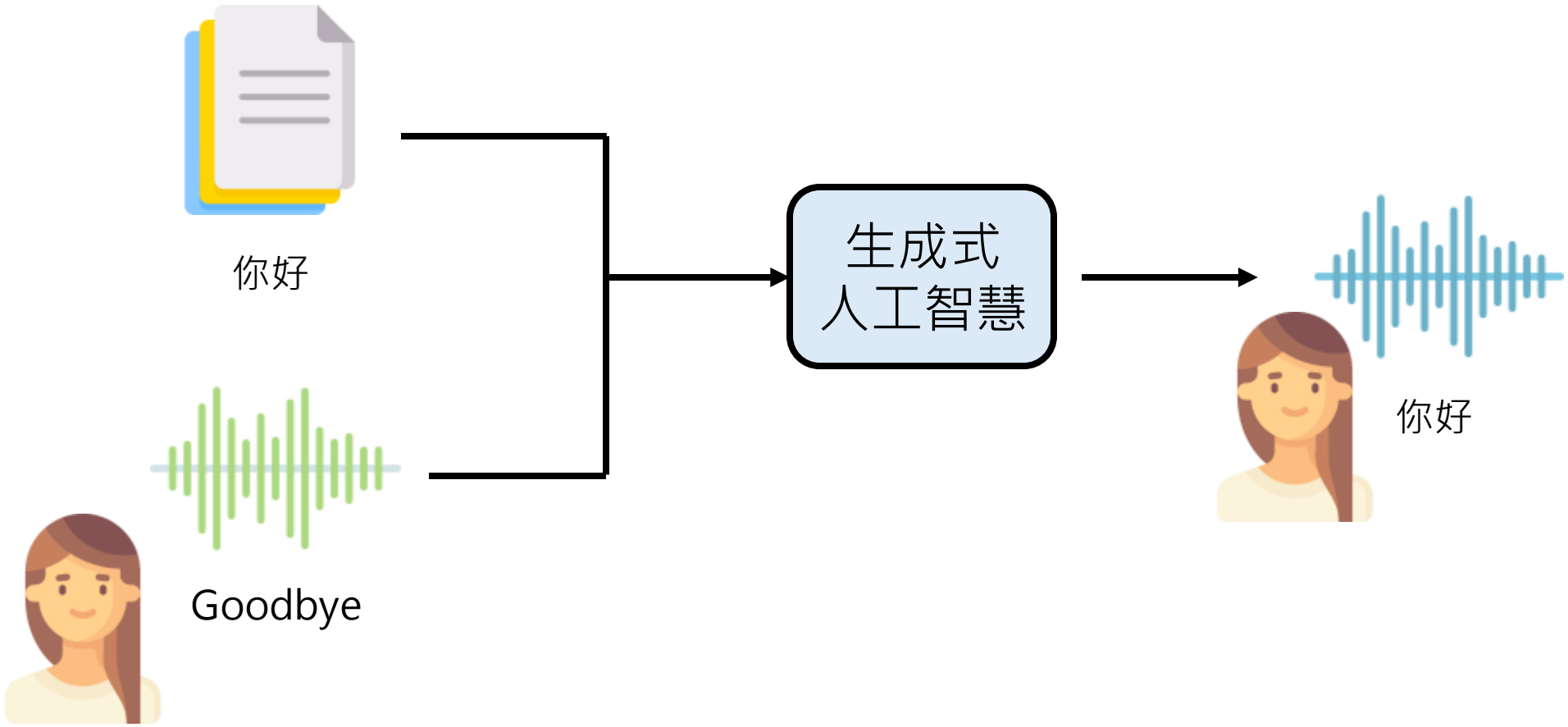
2. Closed under addition: Let $x_1, x_2 \in V+W$. Then $x_1 = v_1 + w_1$, $x_2 = v_2 + w_2$ for some $v_1, v_2 \in V$ and $w_1, w_2 \in W$.

$$x_1 + x_2 = (v_1 + w_1) + (v_2 + w_2) = (v_1 + v_2) + (w_1 + w_2)$$
$$= x_1 + x_2 \in V + W.$$

3. Closed under scalar multiplication: Let $x \in V+W$ and $c \in \mathbb{R}$.
$x = v + w$ for $v \in V$, $w \in W$.
$cx = c(v+w) = cv + cw \Rightarrow cx \in V + W$.

Therefore, $V+W$ is a subspace of $\mathbb{R}^n$.

輸入文字 (+語音)、輸出語音



輸入影像 (+語音)、輸出語音

- https://index-tts.github.io/index-tts2.github.io/

Demo from Index TTS2

https://arxiv.org/abs/2506.21619

輸入文字 (+圖片)、輸出影片

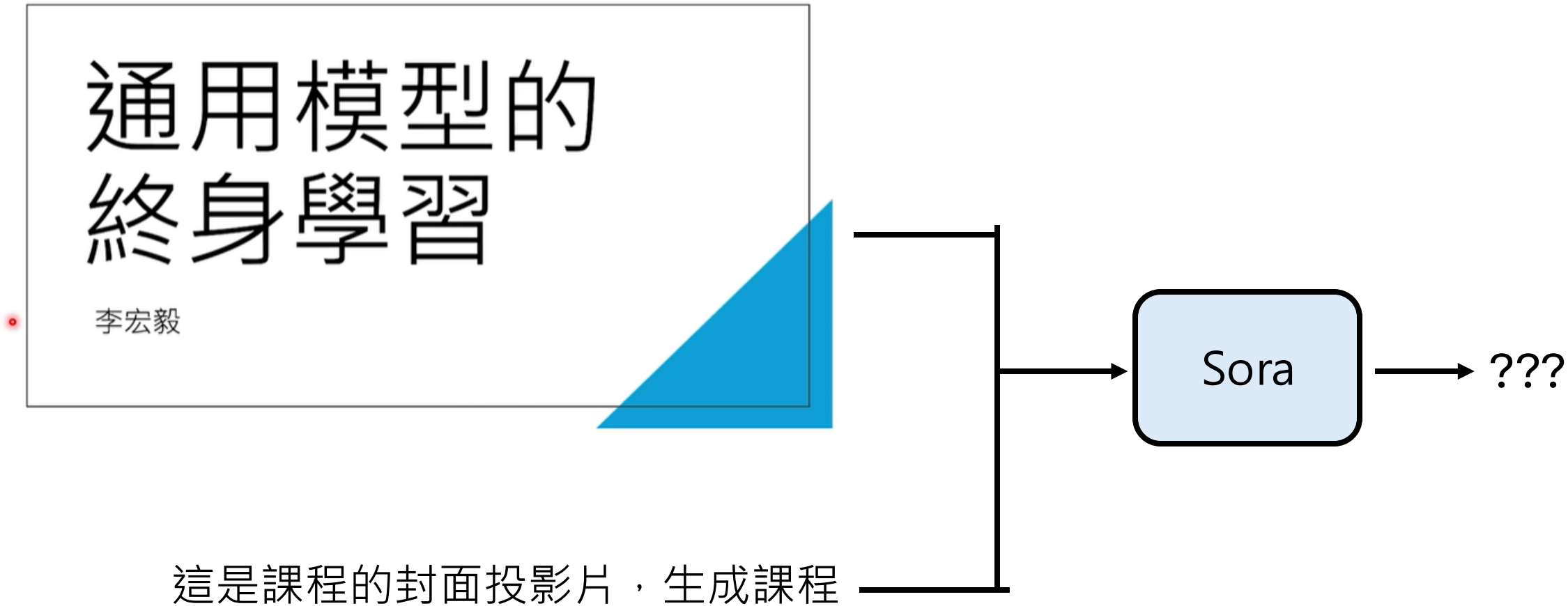
Sora

???

解開這個遊戲的過程


YOU HAVE
ONLY 2 MOVES
IQ: 155
HERO WARS
YOU HAVE
ONLY 2 MOVES
IQ: 155
HERO WARS

輸入文字 (+圖片)、輸出影片



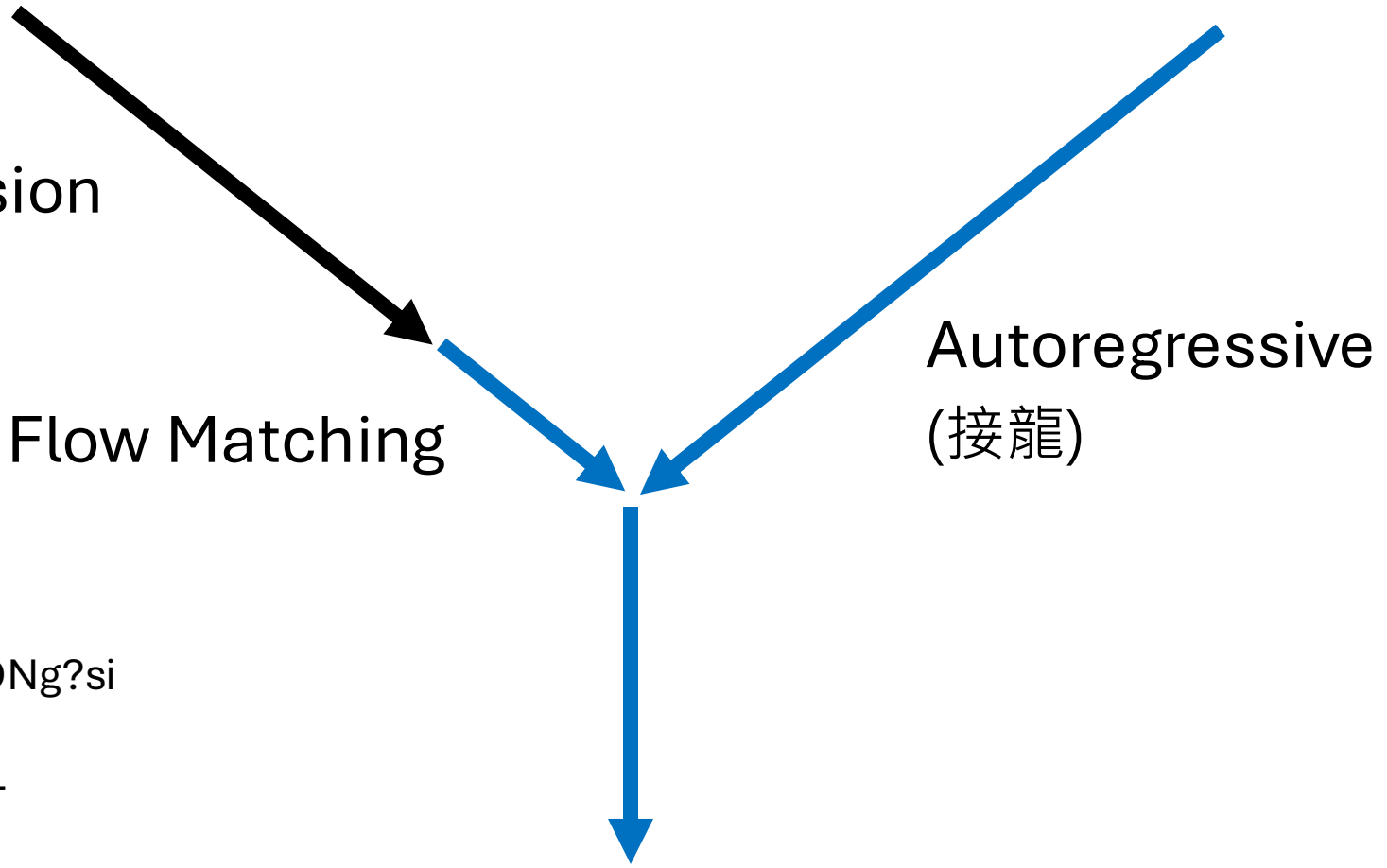
通用模型的終身學習

李宏毅

通用模型的終身學習

李宏毅

技術路線



《生成式人工智慧導論 2024》

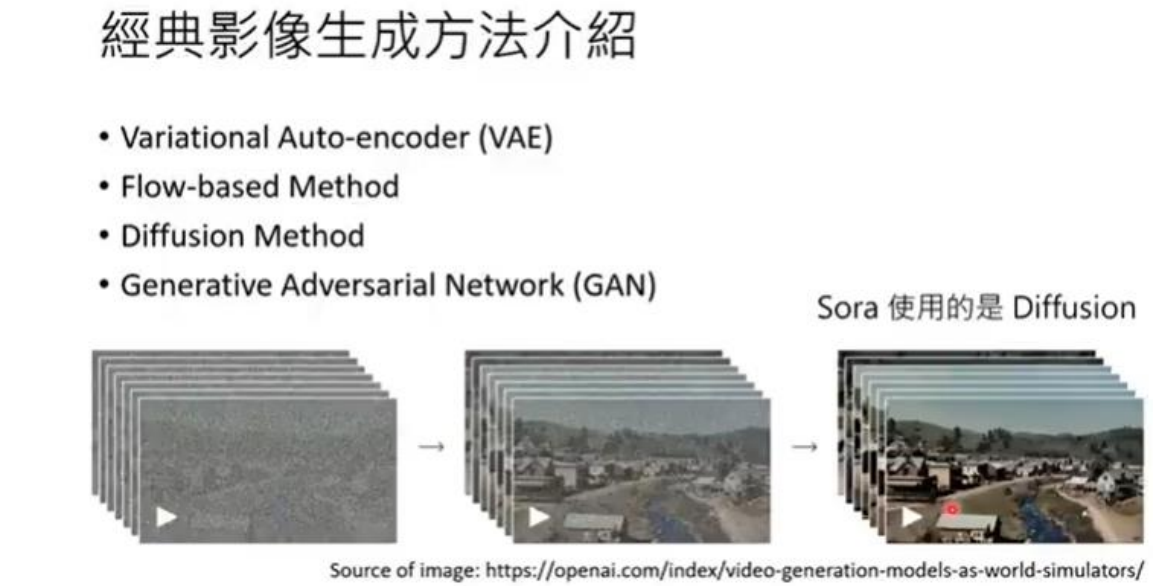
- https://youtu.be/5H2bVEmYDNg?si=_GKFSqSSz-8SZukA
- https://youtu.be/OYN_GvAqv-A?si=WxWP3OeUukfXrchm

在 2025 年兩條路線如何走在一起

Diffusion Model 的世界線

【生成式AI導論 2024】第17講：有關影像的生成式AI (上) — AI 如何產生圖片和影片 (Sora 背後可能用的原理)

https://youtu.be/5H2bVEmYDNg?si=_GKFSqSSz-8SZukA



【生成式AI導論 2024】第18講：有關影像的生成式AI (下) — 快速導讀經典影像生成方法 (VAE, Flow, Diffusion, GAN) 以及與生成的影片互動

https://youtu.be/OYN_GvAqv-A?si=WxWP3OeUukfXrchm

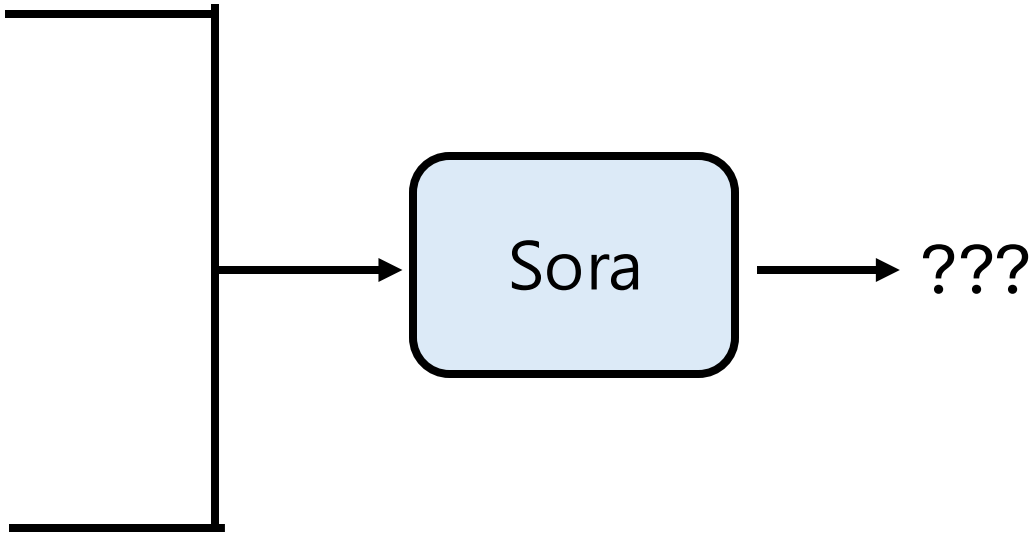
聲音和影像的基本單位是什麼？

構成影像的基本單位

Personalized Generation

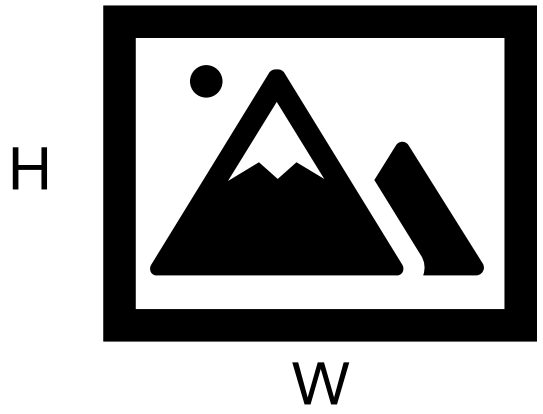
@hungyilee

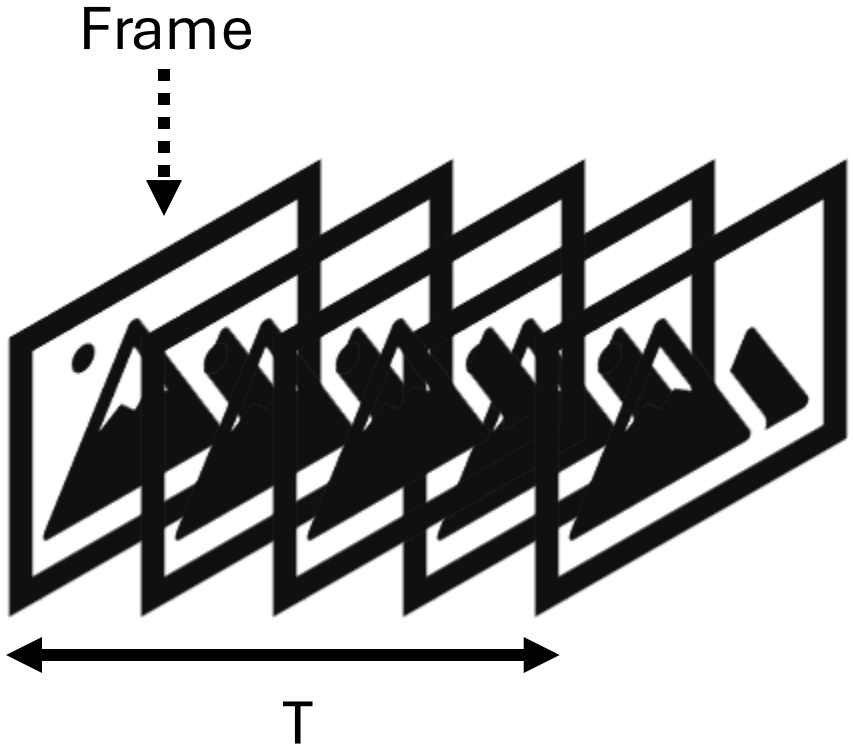
創作一個 @hungyilee 由講解的教學影片，畫一個二次元人物，然後展示他是由像素構成 .....



構成影片的基本單位？

- 影片就是一連串的圖片

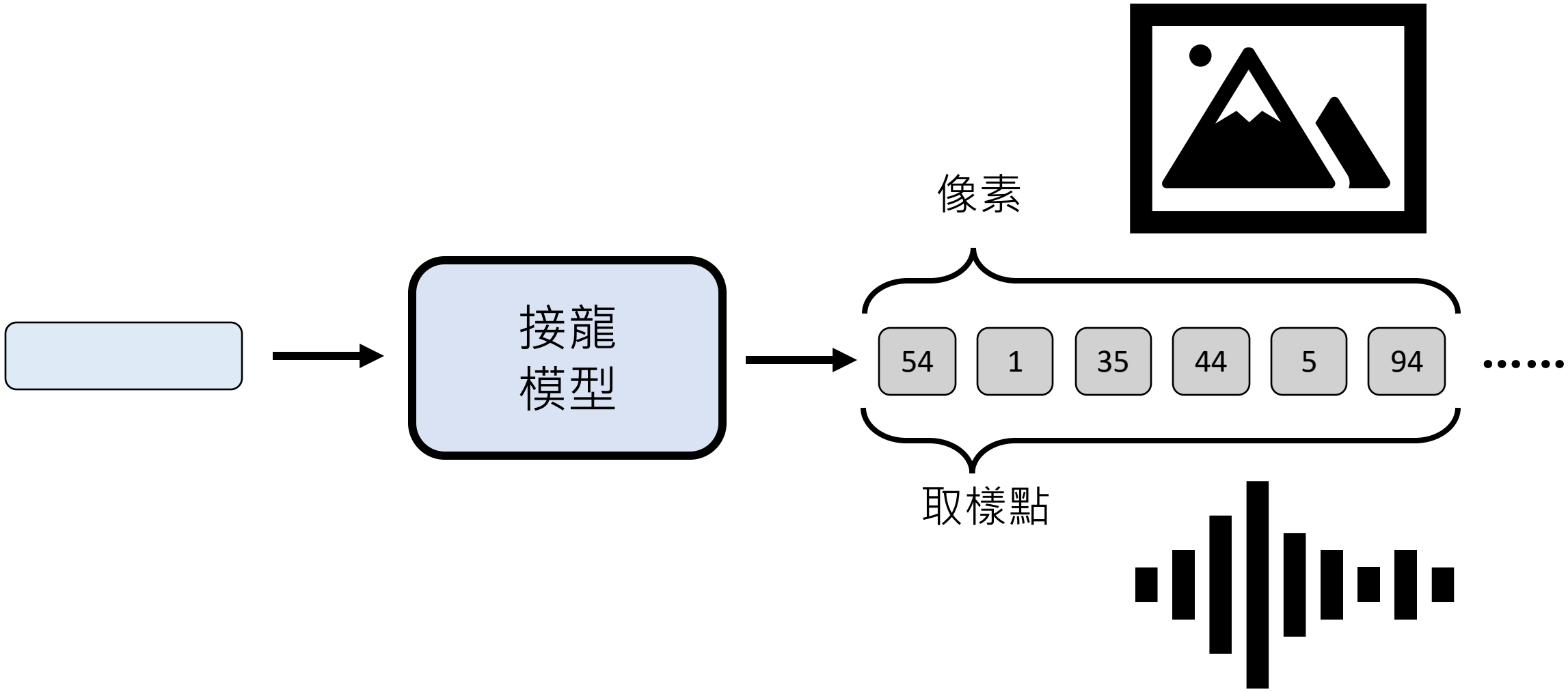




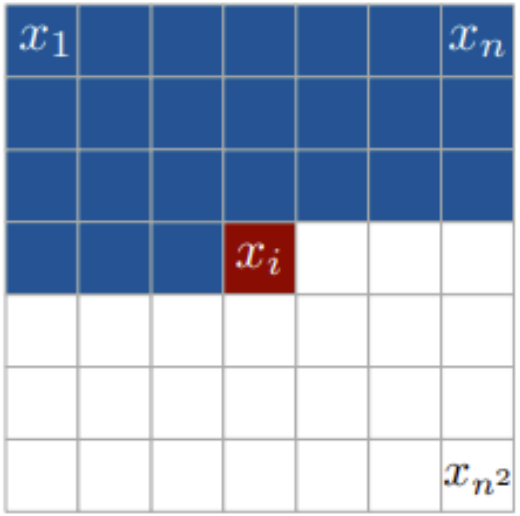
Hygen: 創作由 @hungyilee 講解的教學影片 ，影片說明聲音訊號是由取樣點構成的。遠看是聲音訊號，拉近後則可以看到一個一個的取樣點。一秒鐘有多少取樣點，取決於取樣率（sampling rate）；例如，16kHz 的取樣率意味著一秒鐘有 16,000 個取樣點。而每個取樣點有多少種可能的數值，則取決於取樣解析度（bit resolution）；常見的 16-bit 解析度有 65,535 種可能的數值。

(此處輸入指令有誤 ，應該是 65,536 的數值，但是 AI 自行生成了正確的數字)

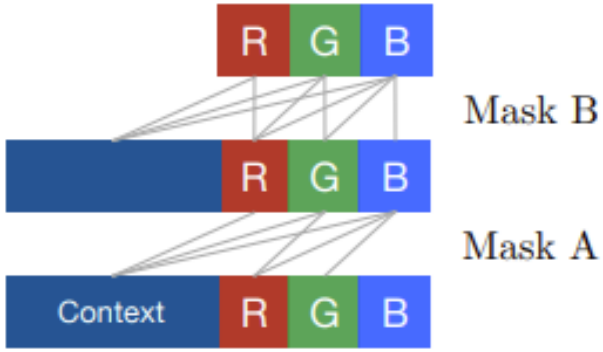
影像/語音生成就是像素/取樣點接龍？



影像生成就是像素接龍？

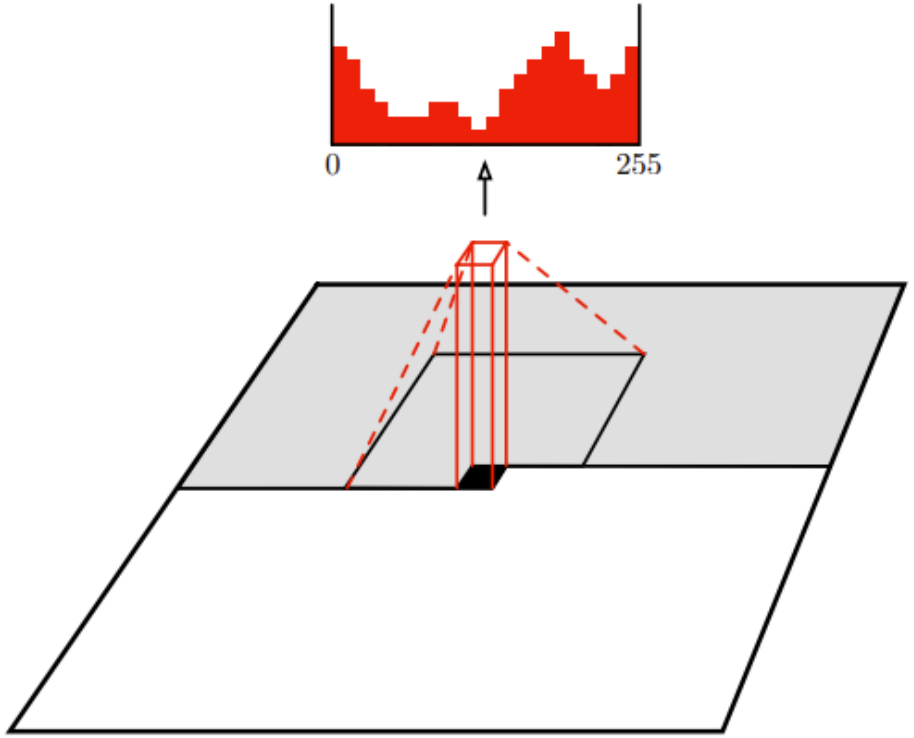


Context



Pixel RNN

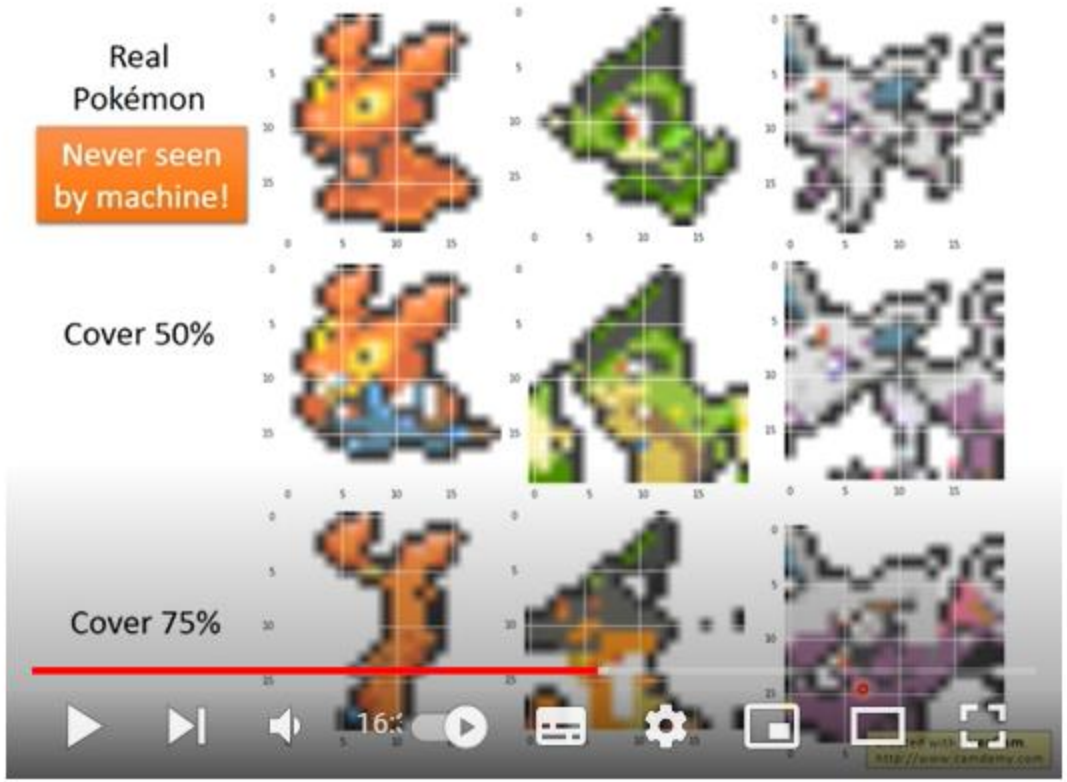
https://arxiv.org/abs/1601.06759



Pixel CNN

https://arxiv.org/abs/1606.05328

影像生成就是像素接龍？



ML Lecture 17: Unsupervised Learning - Deep Generative Model (Part I)

https://youtu.be/YNUek8ioAJk?t=537

(2016 年秋季班《機器學習》上課錄影)

Little Demo

https://speech.ee.ntu.edu.tw/~hylee/mlds/2017-spring.php

input text: black hair blue eyes

input text: pink hair green eyes

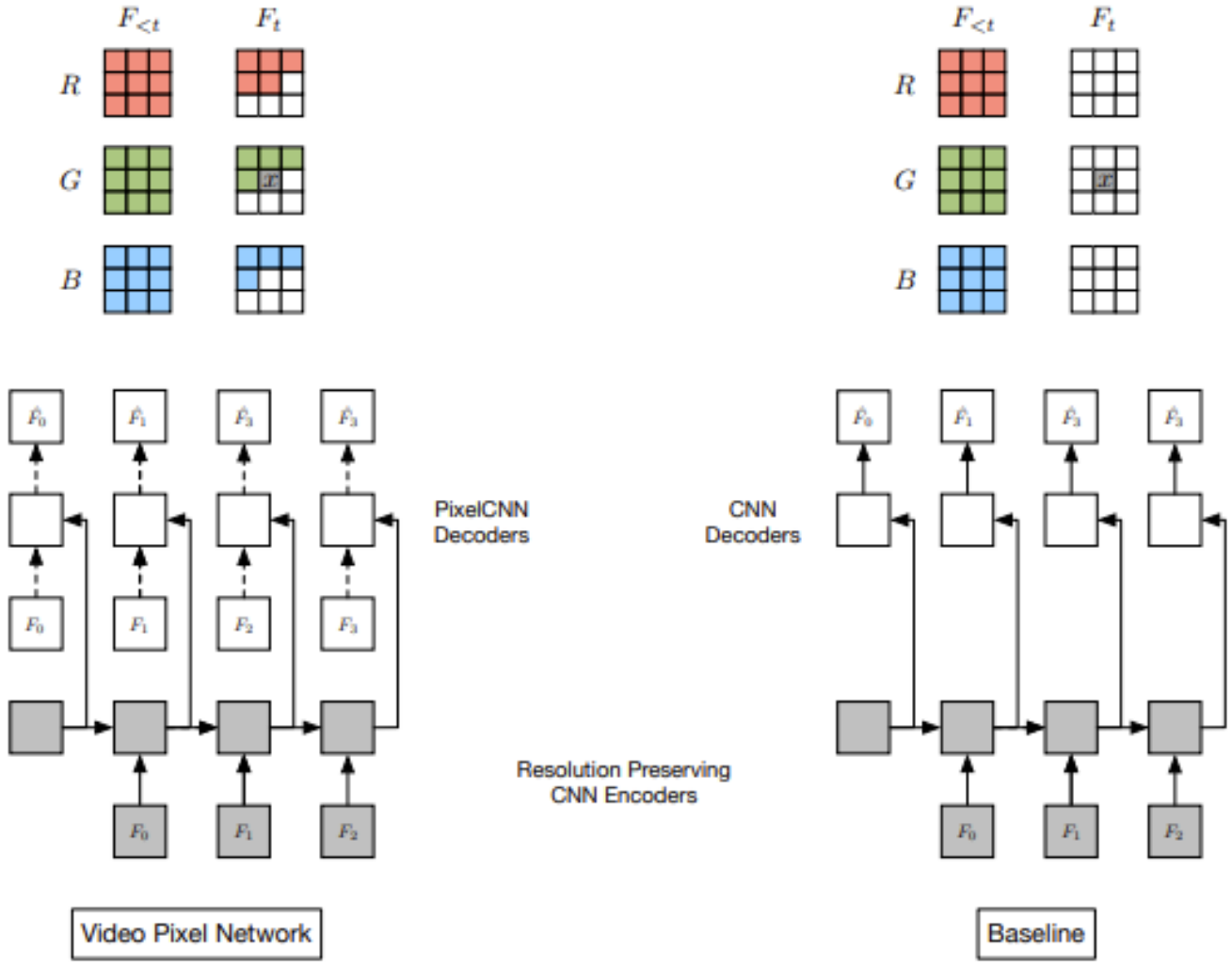
input text: green hair green eyes

input text: blue hair red eyes

(2017 年春季班《機器學習及其生成與結構化》作業三)

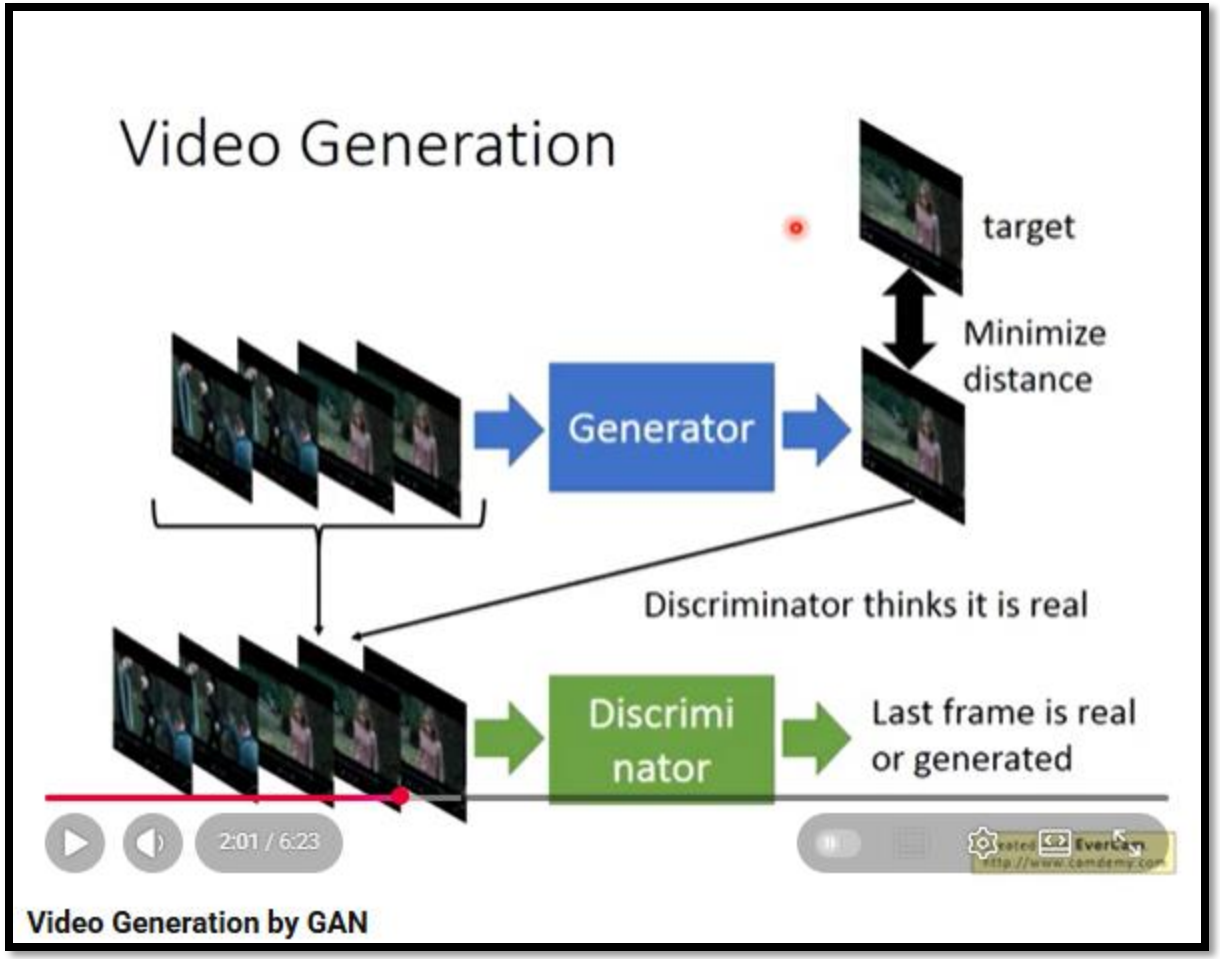
把一張一張圖片產生出來就是影片

- **Video Pixel Networks**



https://arxiv.org/abs/1610.00527

把一張一張圖片產生出來就是影片



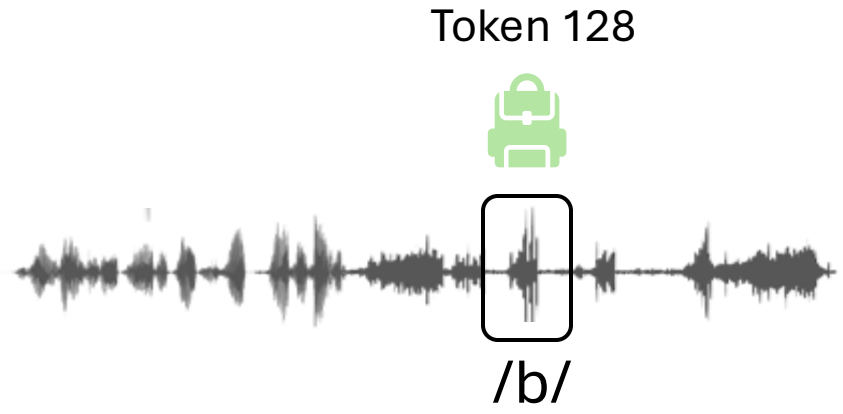
2017 年春季班《機器學習及其生成與結構化》

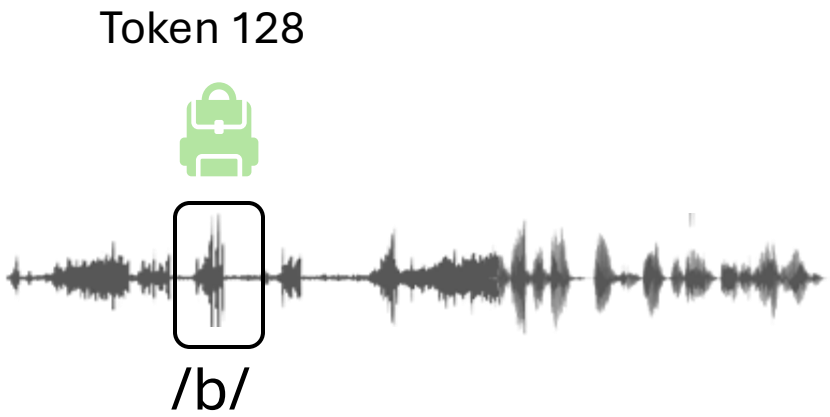
https://www.youtube.com/watch?v=TN8cJiomk_k

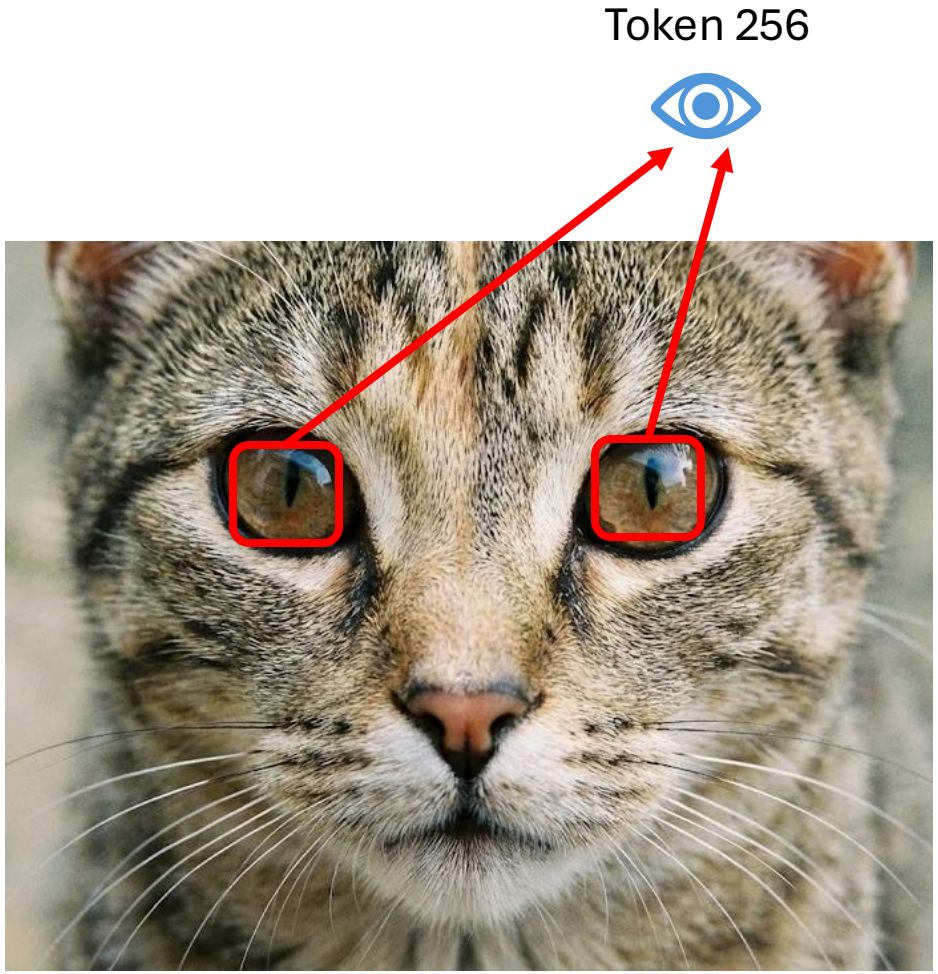
需要適合生成的基本單位：語音和影像的 Token

雖然構成萬物的基本單位是「原子」，
但是我們通常說構成人體的基本單位是「細胞」

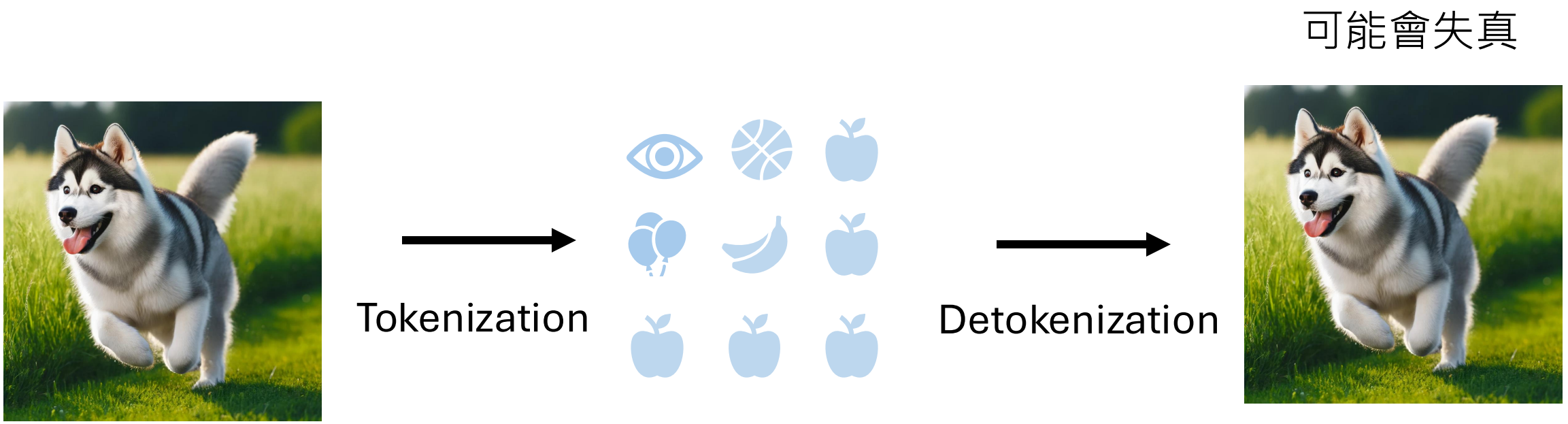
語音和影像的 Token



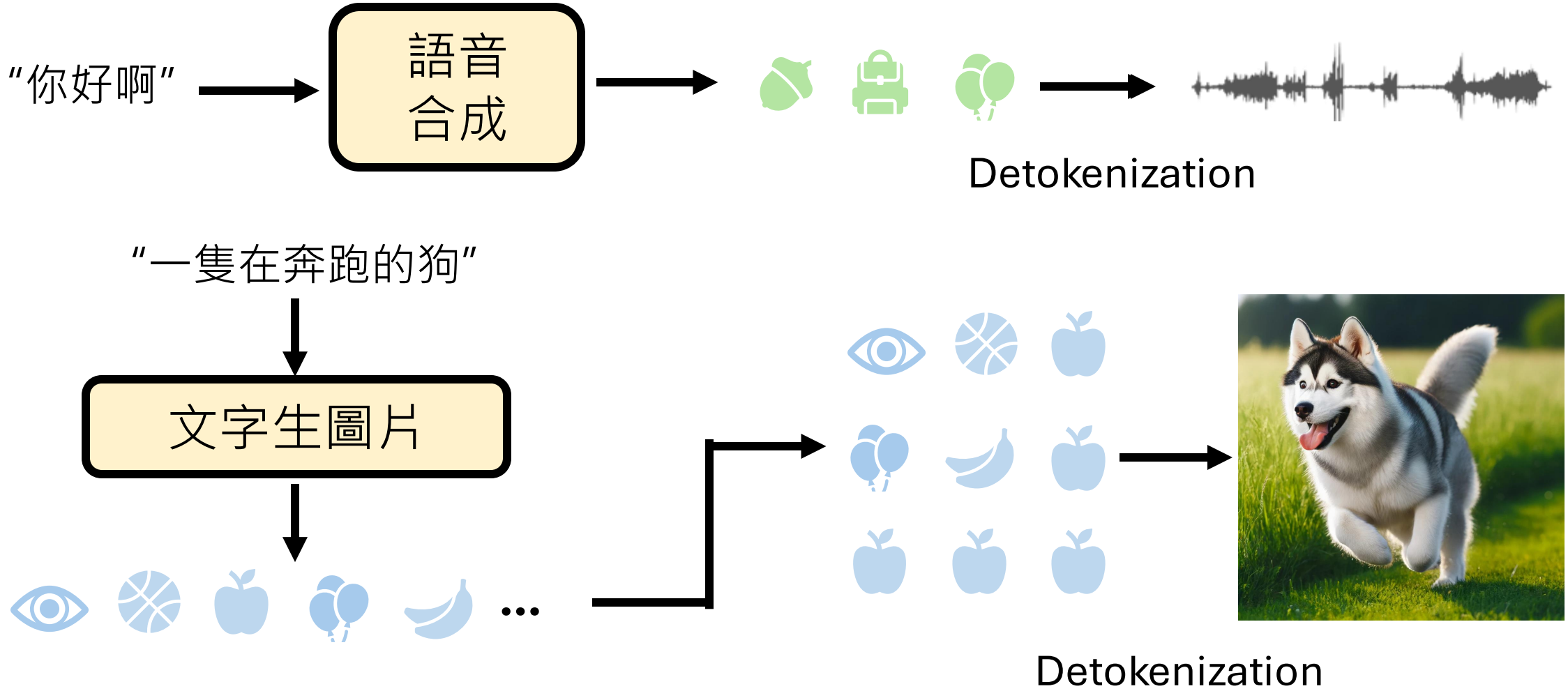




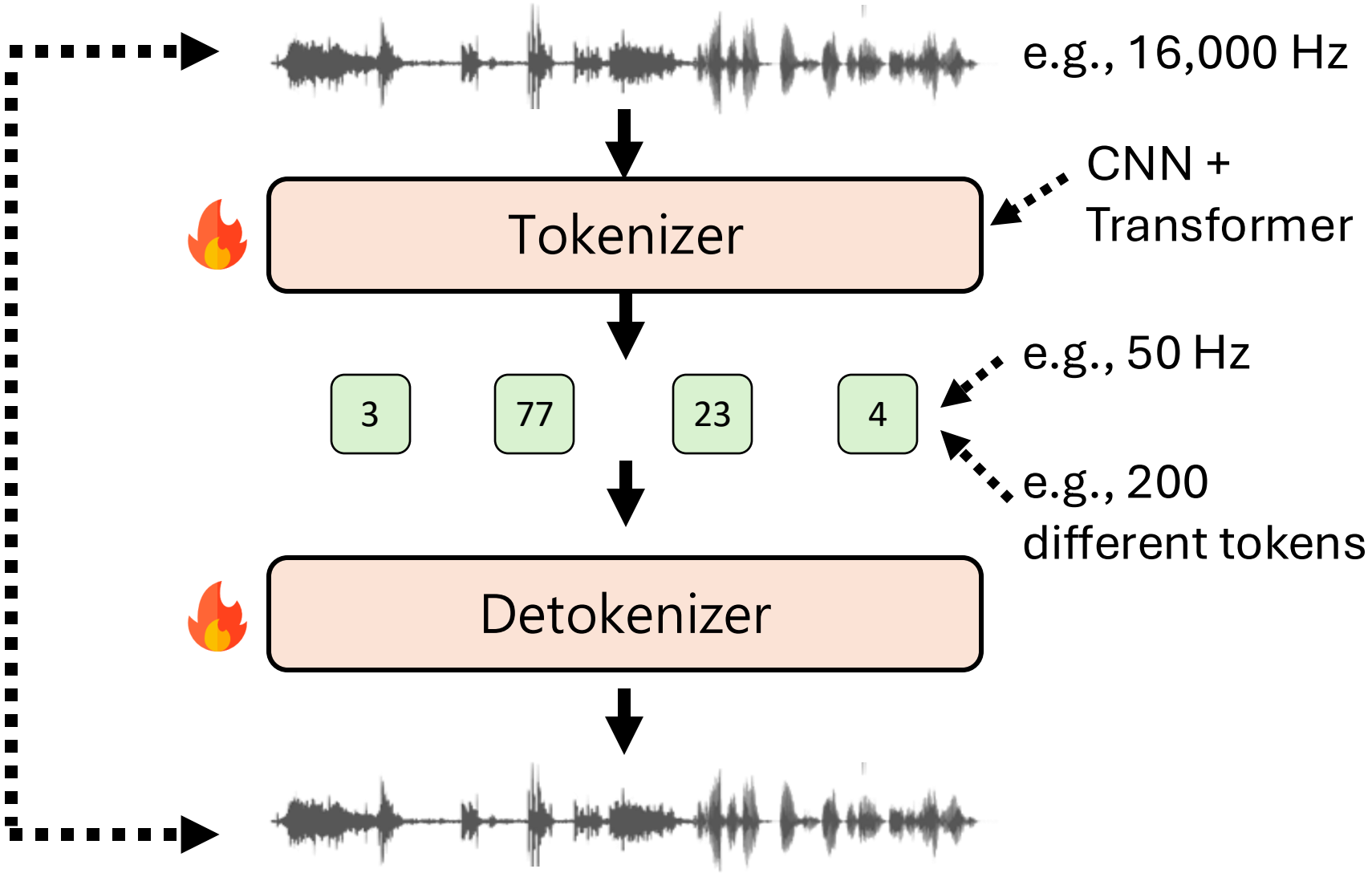
語音和影像的 Token

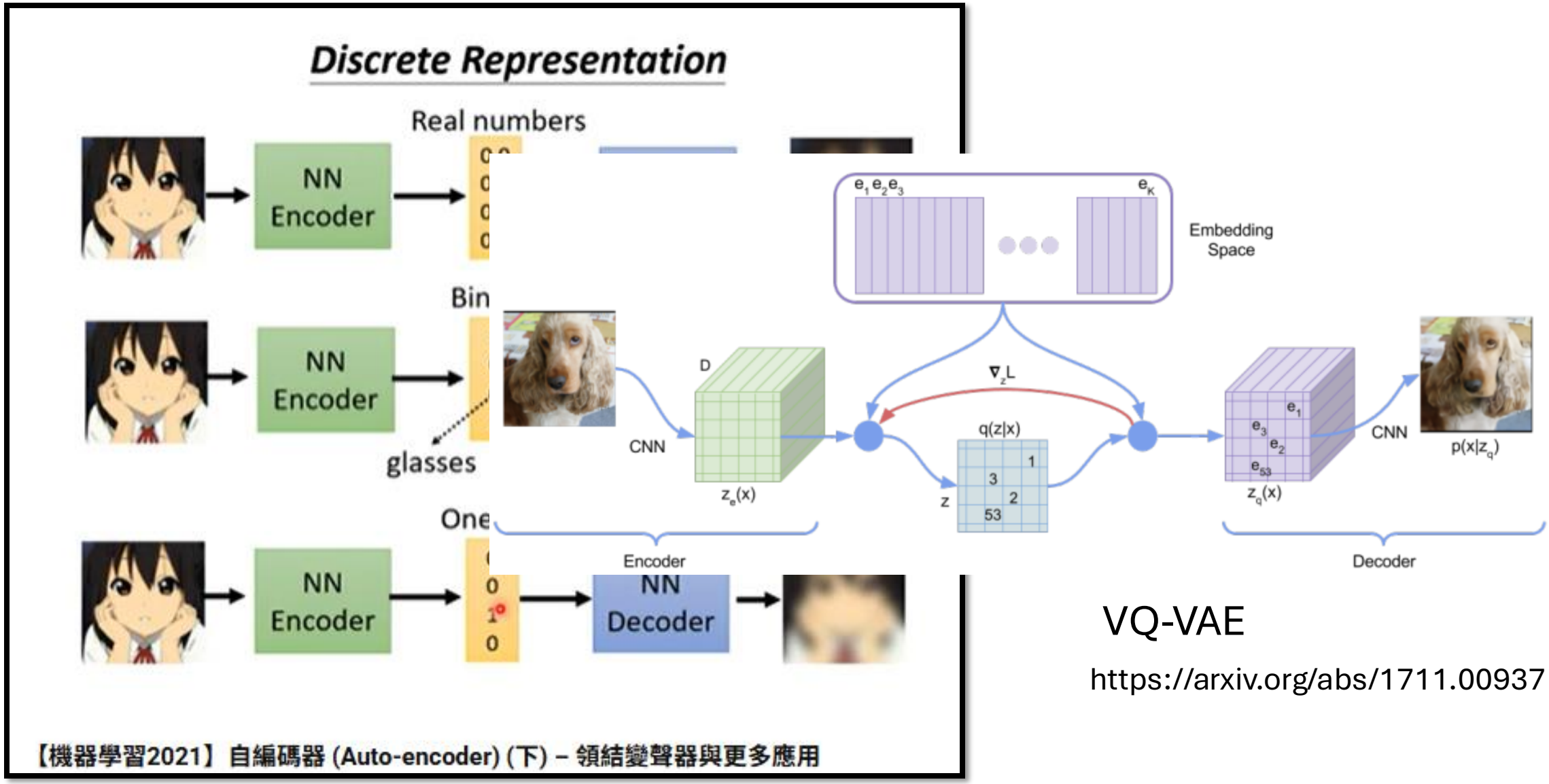


語音、影像仍然是 Token 接龍



如何找出語音和影像的Token？





VQ-VAE

https://arxiv.org/abs/1711.00937

https://youtu.be/JZvEzb5PV3U?si=EZaWgvTl2Um4vrpJ&t=580

(為了方便上課說明，本頁投影片的表達並不精確，
真正的做法詳見論文)

Tokenizer 1

3 77 23 4 200

Tokenizer 2

31 99 13 71 200

200^2

(類似進位的概念)

Residual Vector Quantization (RVQ)

SoundStream	https://arxiv.org/abs/2107.03312
EncoDec	https://arxiv.org/abs/2210.13438
Mimi	https://arxiv.org/abs/2410.00037

Detokenizer

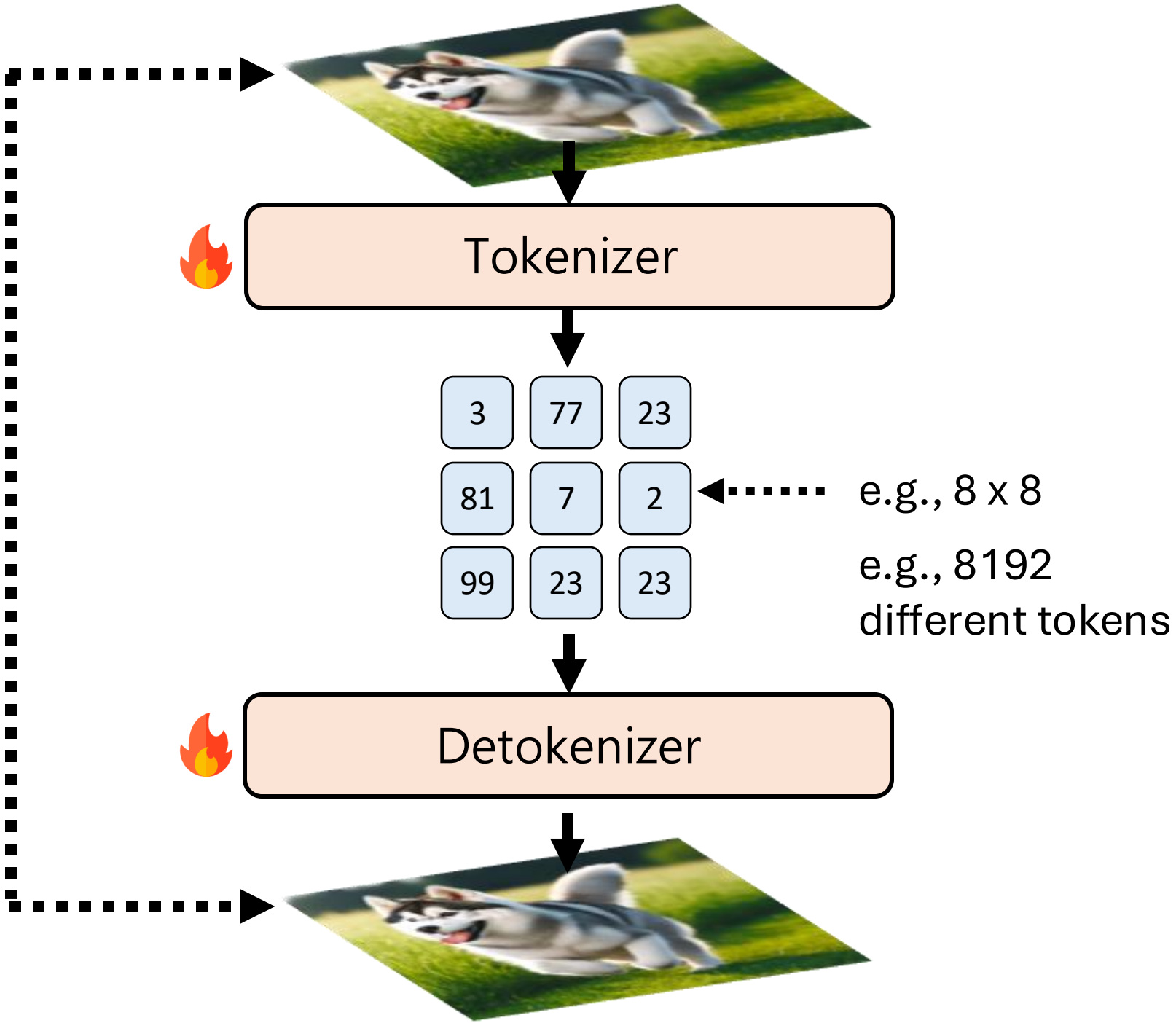
Discrete Audio Tokens: More Than a Survey!

Pooneh Mousavi*†[1,2], Gallil Maimon*[3], Adel Moumen*[4], Darius Petermann*[5], Jiatong Shi*[6], Haibin Wu*[7], Haici Yang*[5], Anastasia Kuznetsova*[5], Artem Ploujnikov[8,2], Ricard Marxer[9], Bhuvana Ramabhadran[10], Benjamin Elizalde[11], Loren Lugosch[11], Jinyu Li[7], Cem Subakan[12,2,1], Phil Woodland[4], Minje Kim[14], Hung-yi Lee[13], Shinji Watanabe[6], Yossi Adi[3], Mirco Ravanelli[1,2,8]

[1]Concordia University, [2]Mila-Quebec AI Institute, [3]The Hebrew University of Jerusalem, [4]University of Cambridge, [5]Indiana University, [6]Carnegie Mellon University, [7]Microsoft, [8]Université de Montréal, [9]Université de Toulon, [10]Google, [11]Apple [12]Laval University, [13]National Taiwan University, [14]University of Illinois at Urbana-Champaign

https://arxiv.org/abs/2506.10274

如何找出語音和影像的Token？



As “close” as possible

OpenAI DALL·E

https://arxiv.org/abs/2204.06125

Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

A blue Porsche 356 parked in front of a yellow brick wall

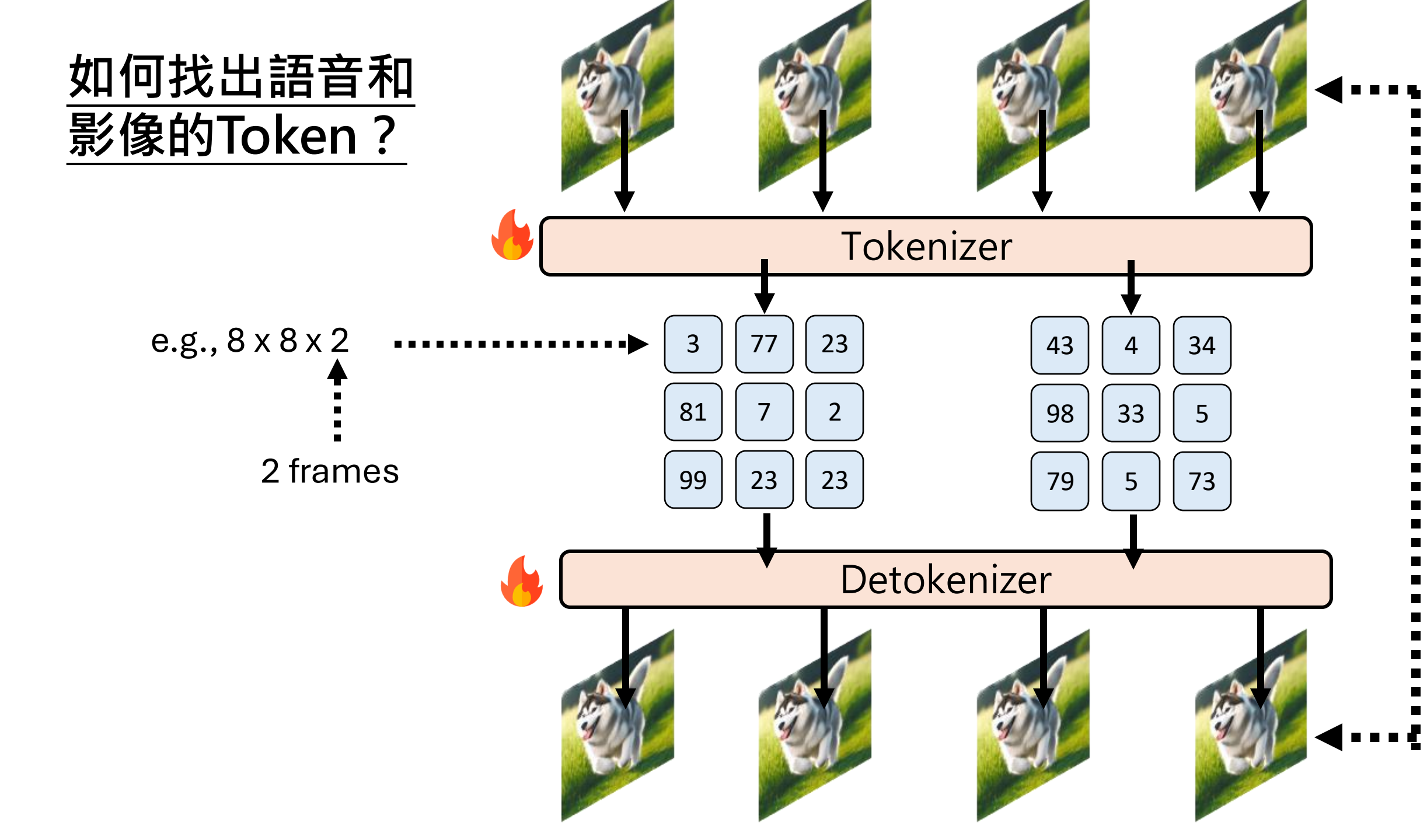
A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.

a photograph of the mona lisa drinking coffee as she has her breakfast. her plate has an omelette and croissant

A portrait of a metal statue of a pharaoh wearing steampunk glasses and a leather jacket over a white t-shirt that has a drawing of a space shuttle on it.

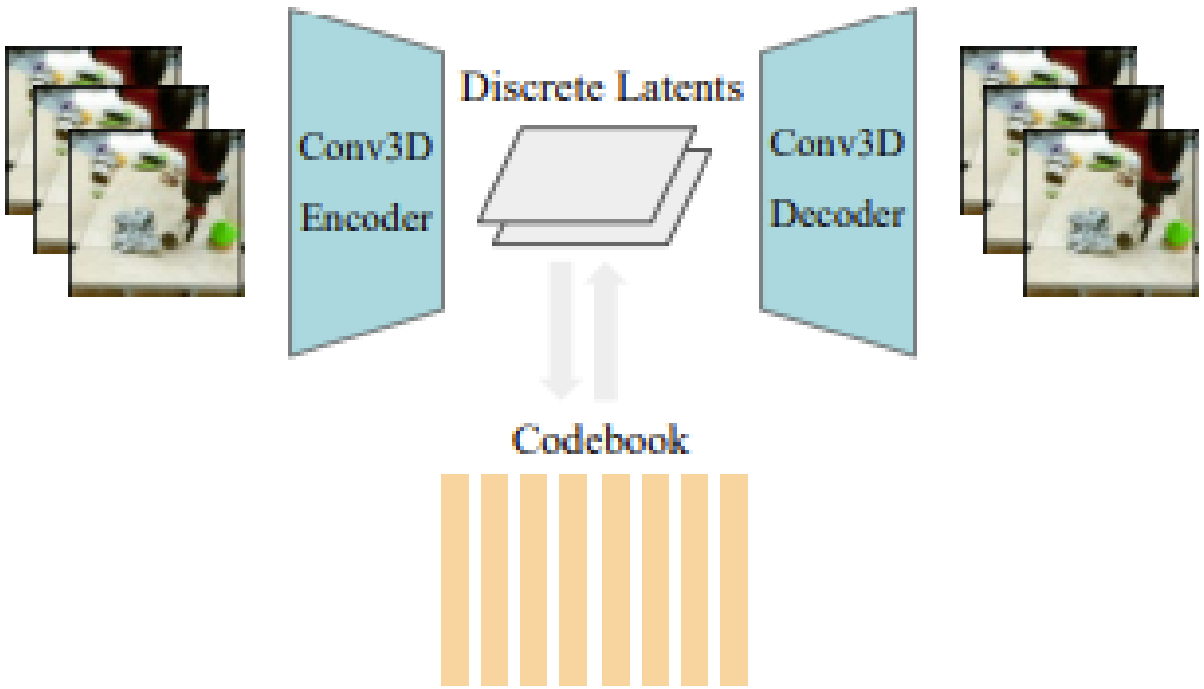
https://arxiv.org/abs/2406.06525

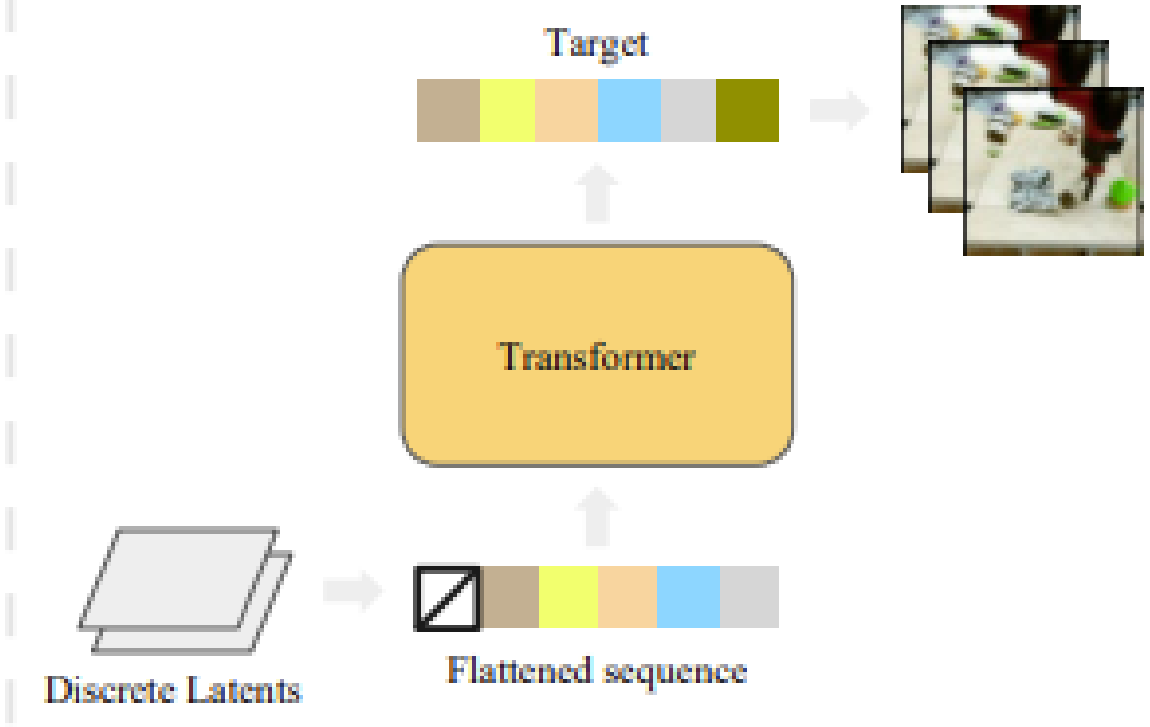
如何找出語音和影像的Token？



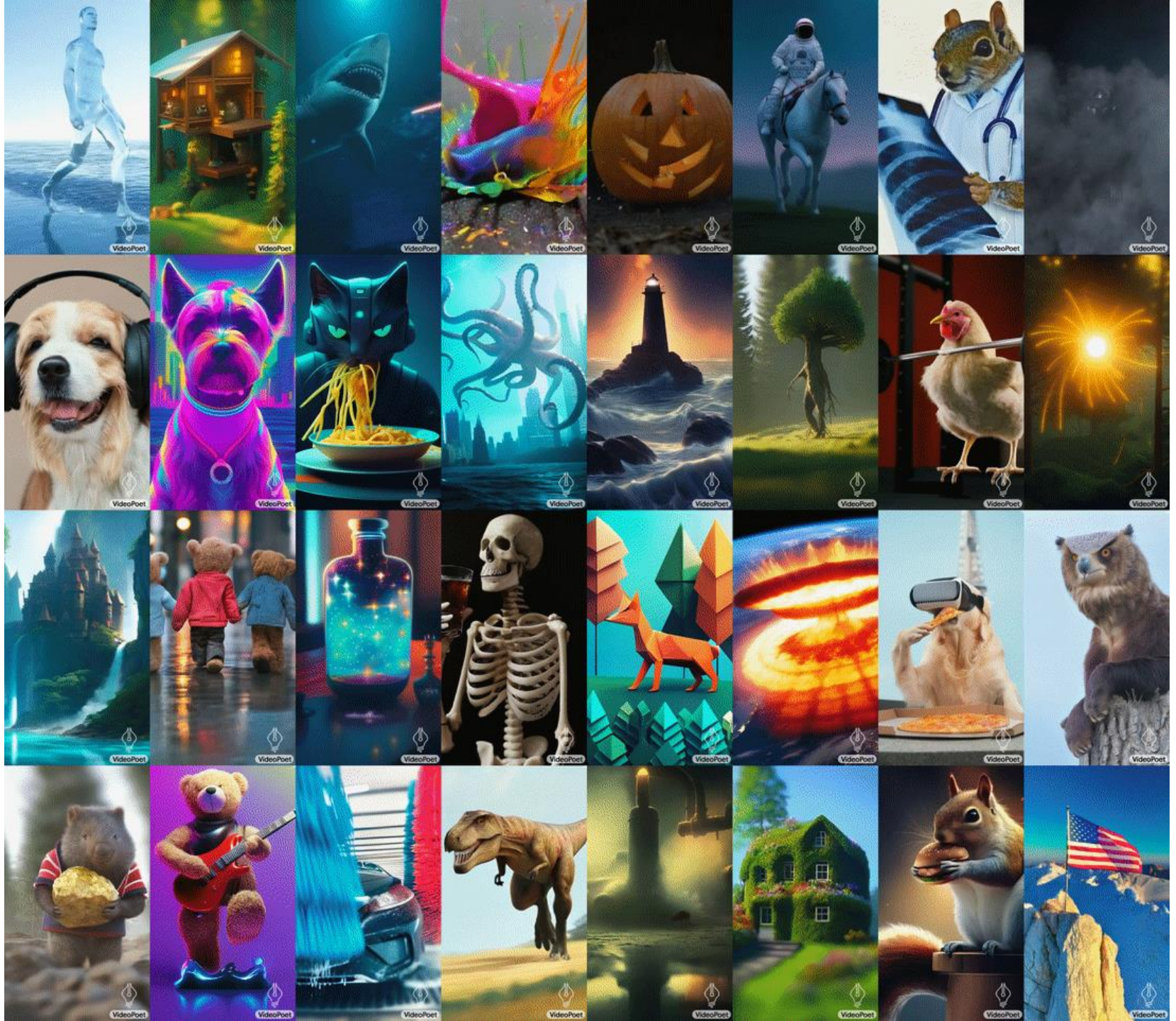
影像生成 Video GPT

https://arxiv.org/pdf/2104.10157
https://wilsonyan.com/videogpt/index.html





Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation



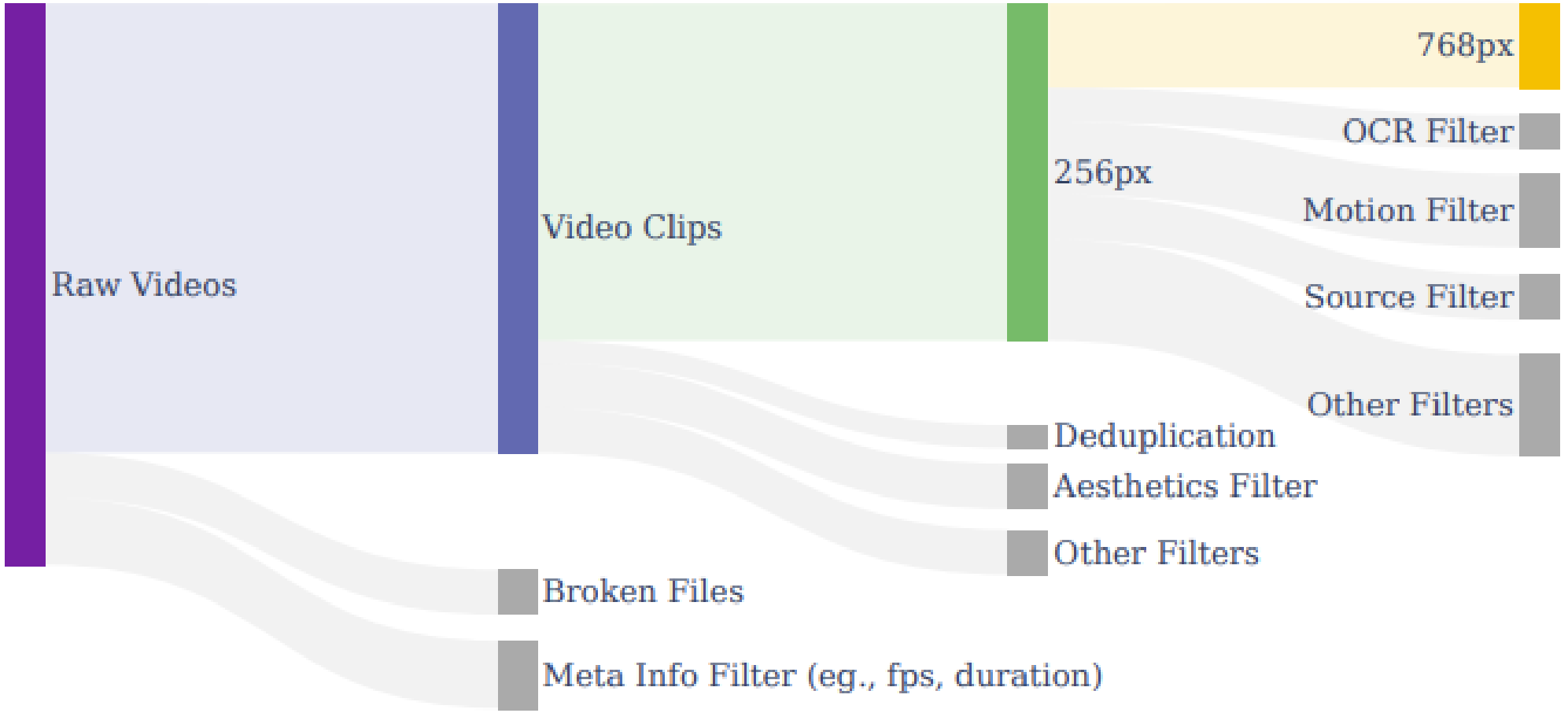
https://arxiv.org/abs/2310.05737

Training Data?

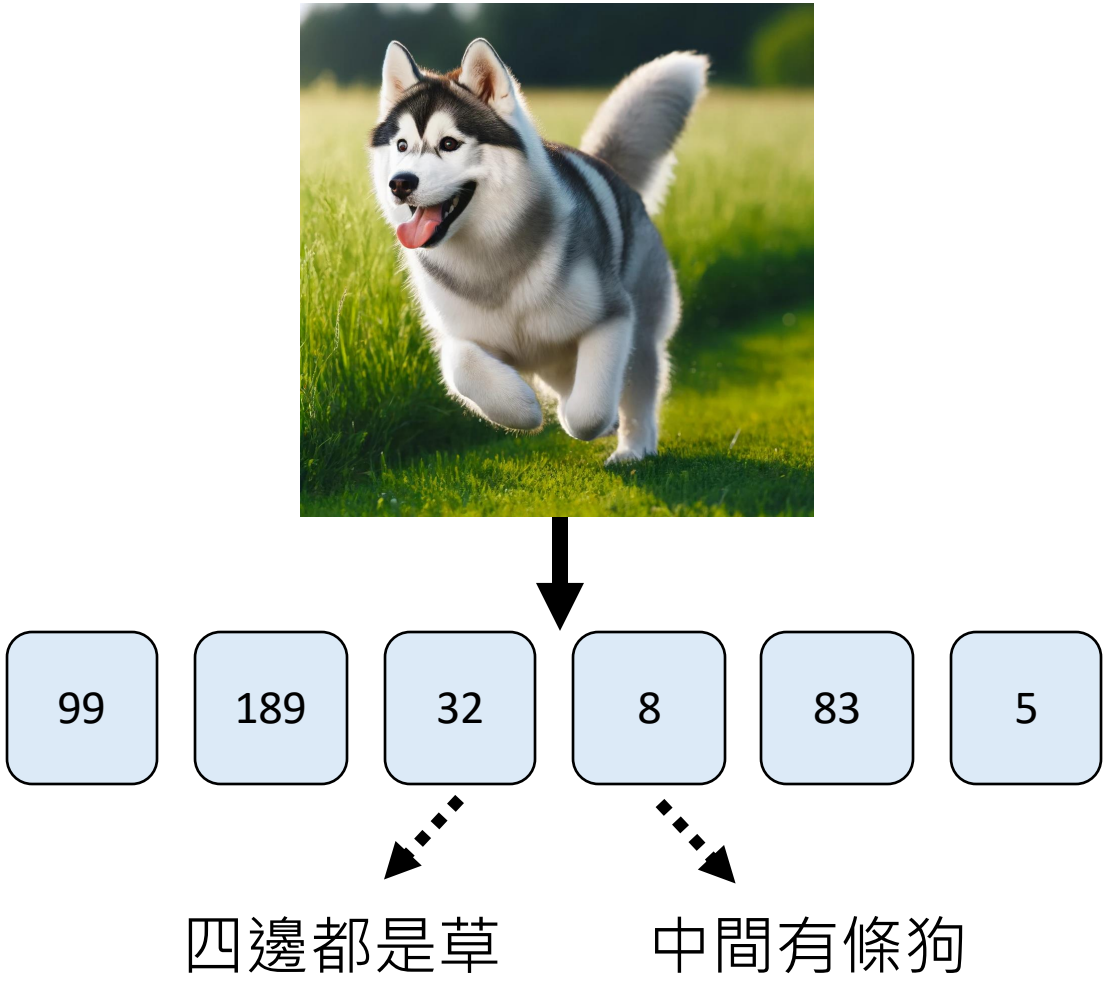
Open Sora
https://arxiv.org/abs/2412.20404
https://arxiv.org/abs/2503.09642

The dataset used is all open-sourced to make the model training fully reproducible. In total, 30M video clips ranging from 2s to 16s are generated, with a total duration of 80k hours. **Webvid-10M** [1] contains 10M video-
watermark. **Panda-**
high-quality subset
generated by BLIP-
high-quality dataset
densely annotated d
4K resolution.

In addition, we get f
websites are of high

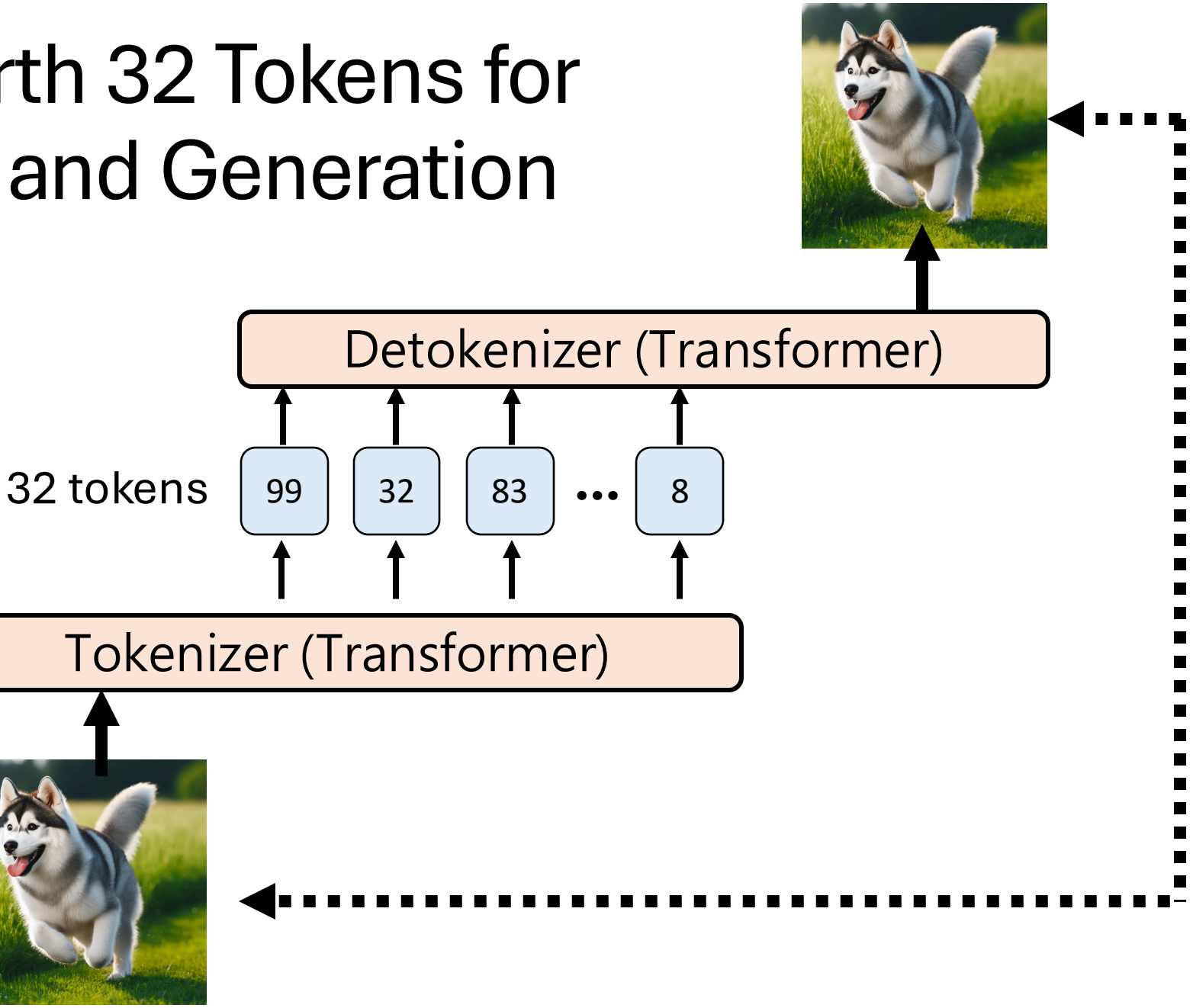


影像的 Token 一定要以 2-D 的方式排列？

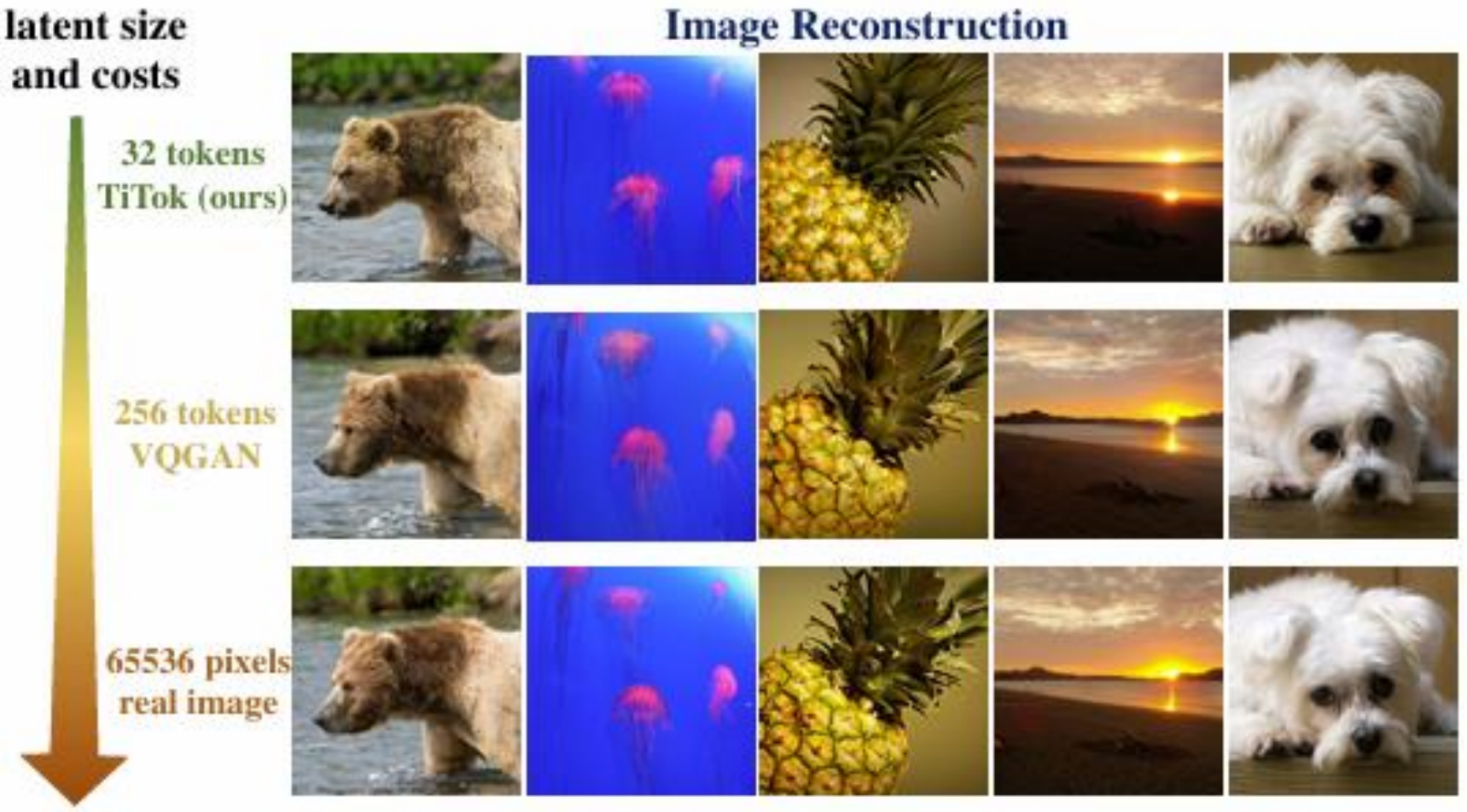


An Image is Worth 32 Tokens for Reconstruction and Generation

https://arxiv.org/abs/2406.07550



An Image is Worth 32 Tokens for Reconstruction and Generation



清晰
(但略
有改變)

較模糊

原圖

甚麼叫做「像」

表面上的像

112, 118, 135, 160, 192, 228, 260, 281, -6, -22, -29, -25, ...

As “close”
as possible

……

Regression loss

135, 150, 152, 140, 115, 80, 40, 0, -32, -55, -65, -62, ...

甚麼叫做「像」

表面上的像

112, 118, 135, 160, 192, 228, 260, 281, -6, -22, -29, -25, ...

As “close”
as possible

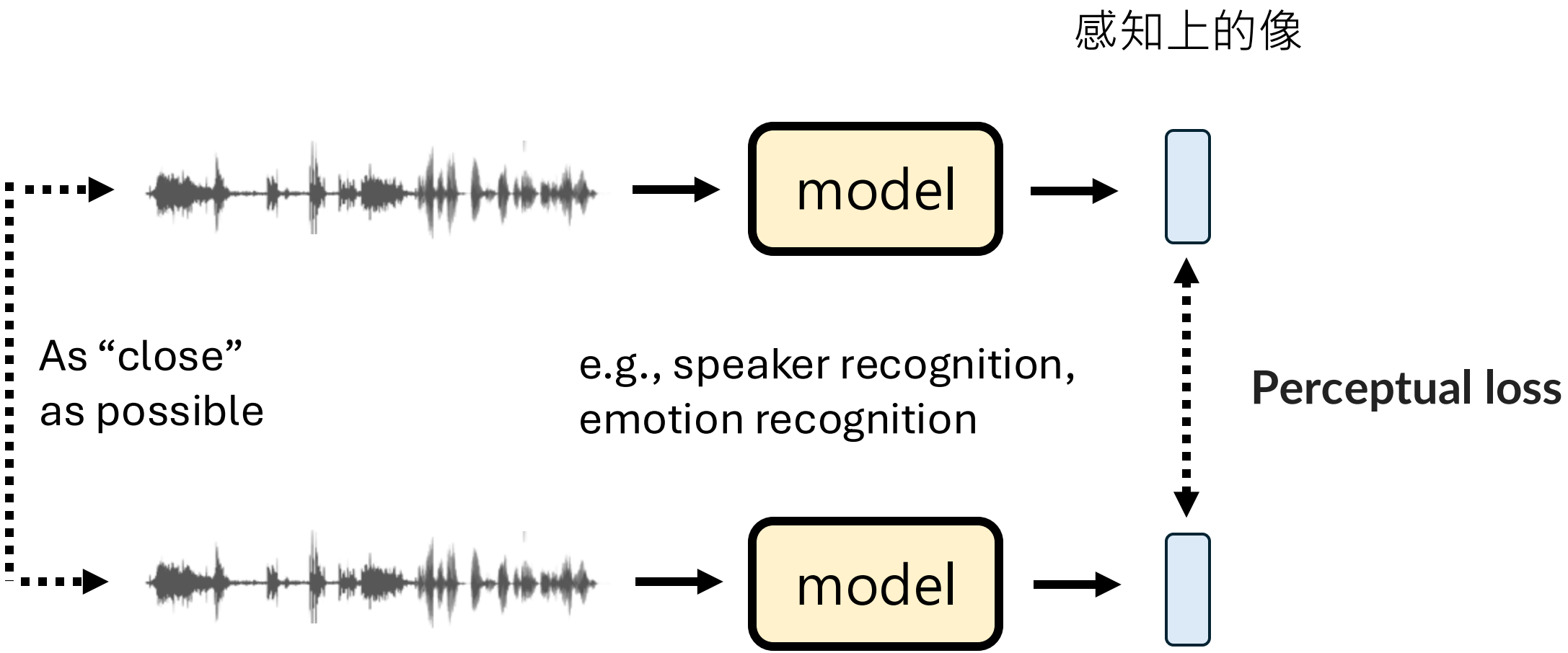
.....

Regression loss
有巨大的差異

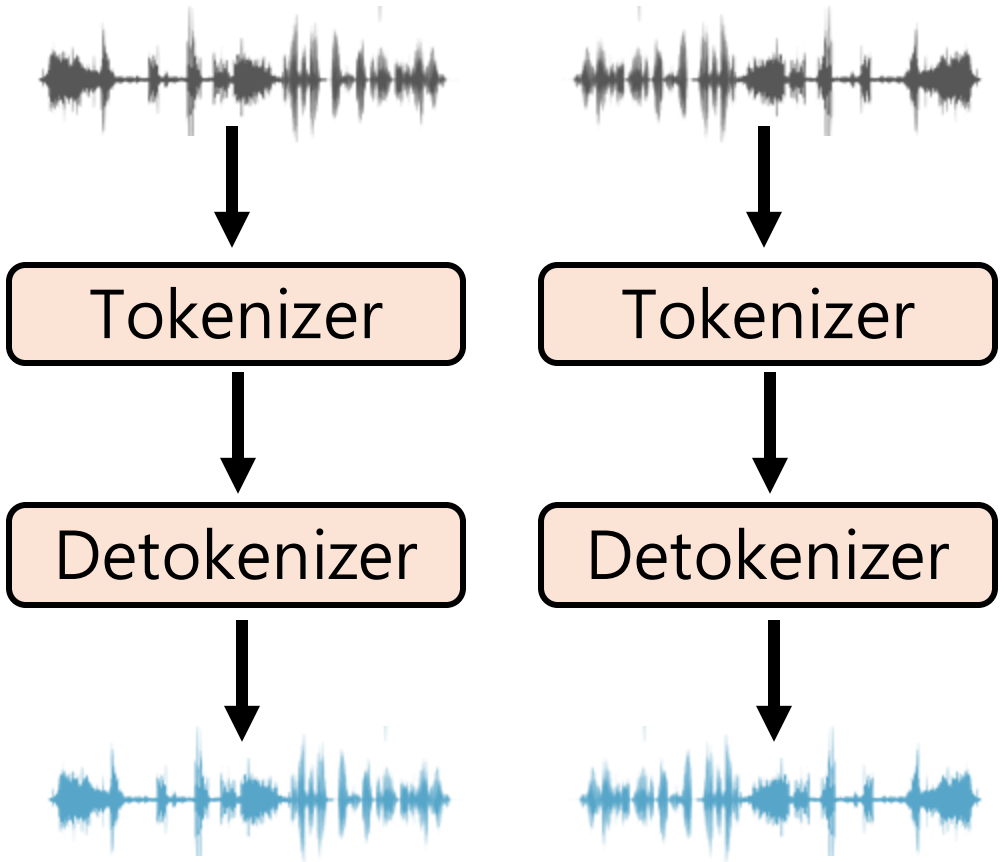
0, 112, 118, 135, 160, 192, 228, 260, 281, -6, -22, -29, ...

人類根本聽不出差異

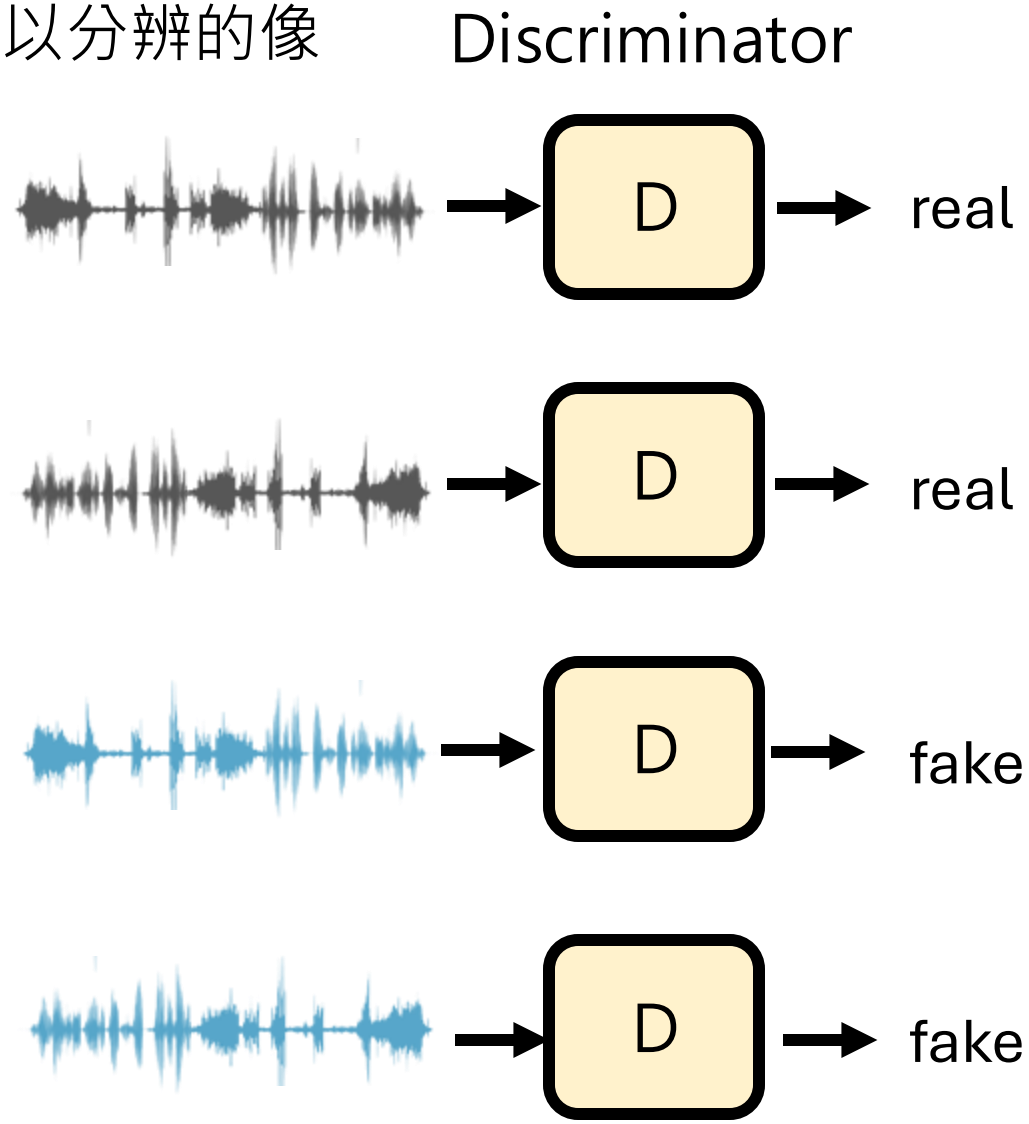
甚麼叫做「像」


感知上的像
model
model
As “close”
as possible
e.g., speaker recognition,
emotion recognition
Perceptual loss

甚麼叫做「像」

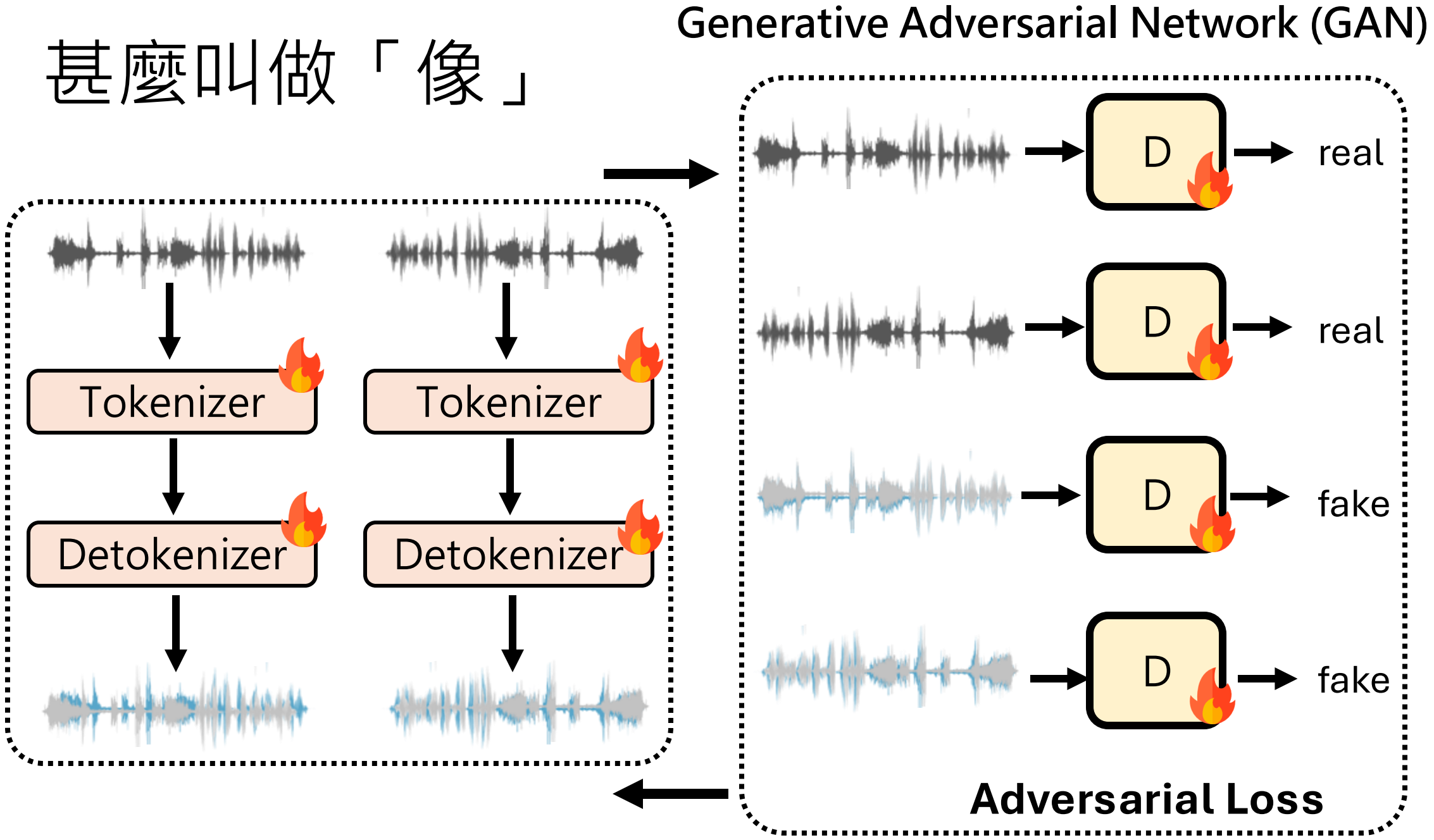


模型難以分辨的像

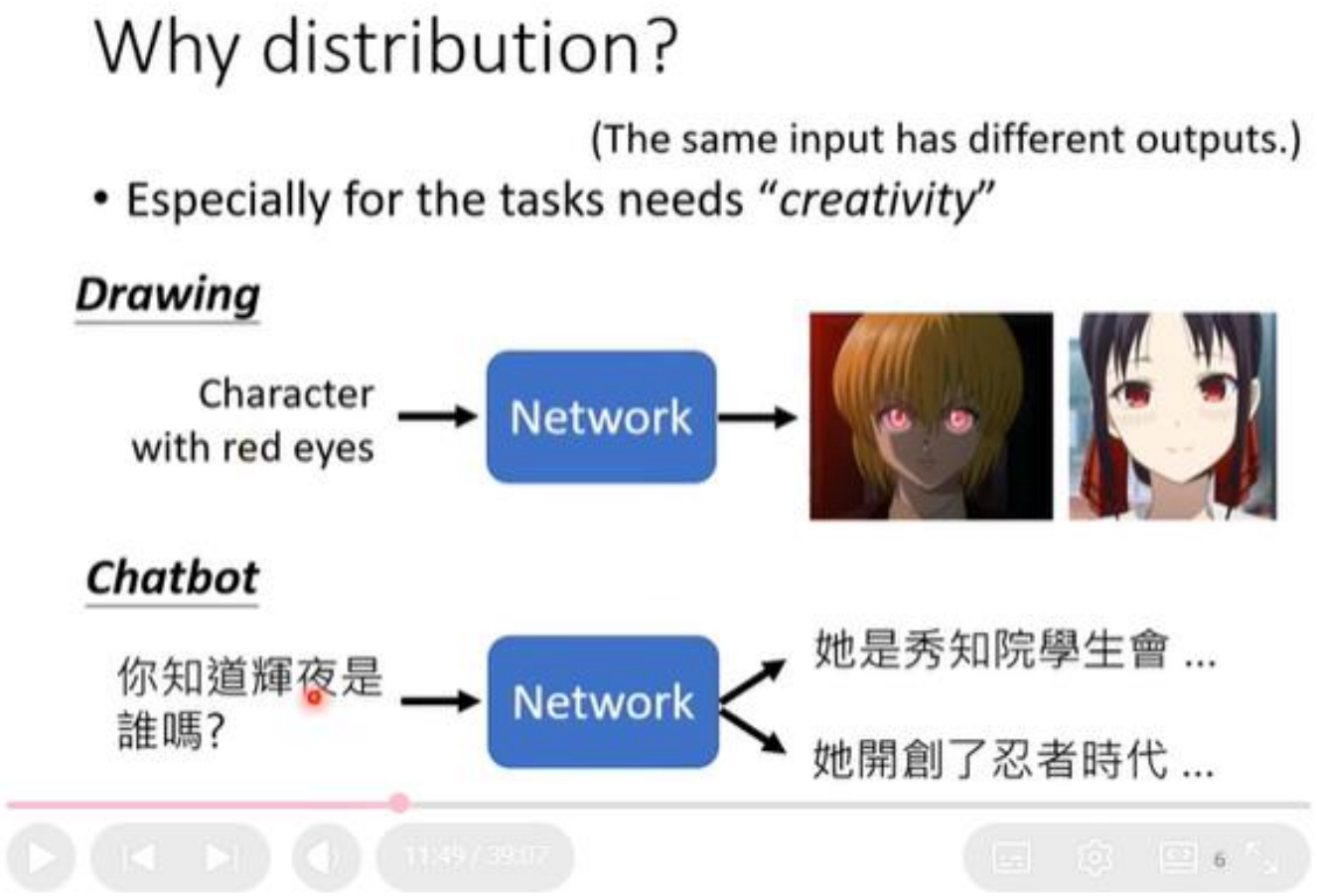


Adversarial Loss

甚麼叫做「像」


Generative Adversarial Network (GAN)
Tokenizer
Tokenizer
Detokenizer
Detokenizer
D
D
D
D
real
real
fake
fake
Adversarial Loss

GAN



https://www.youtube.com/watch?v=4OWp0wDu6Xw&list=PLJV_el3uVTsMhtt7_Y6sgTHGHp1Vb2P2J&index=14

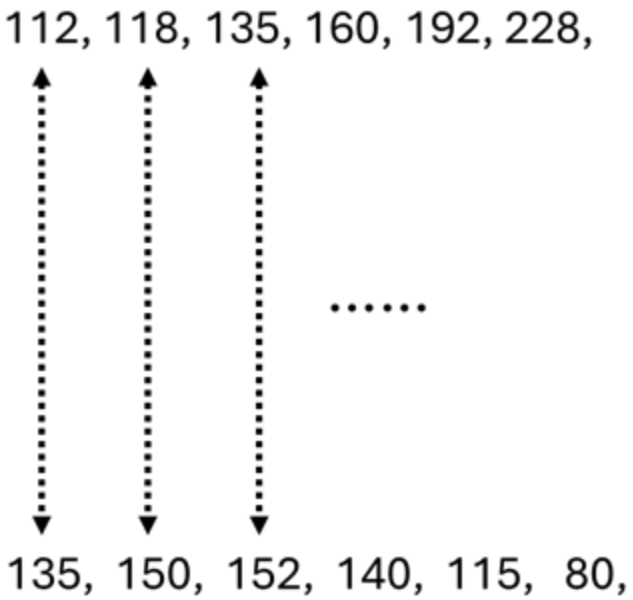
甚麼叫做「像」

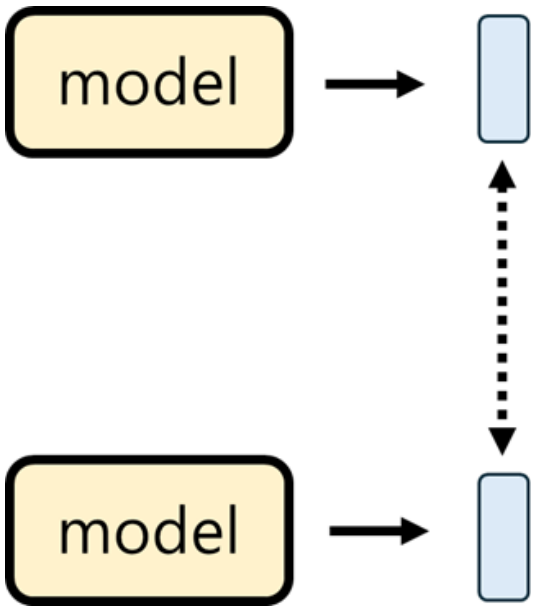
1. 表面上的像

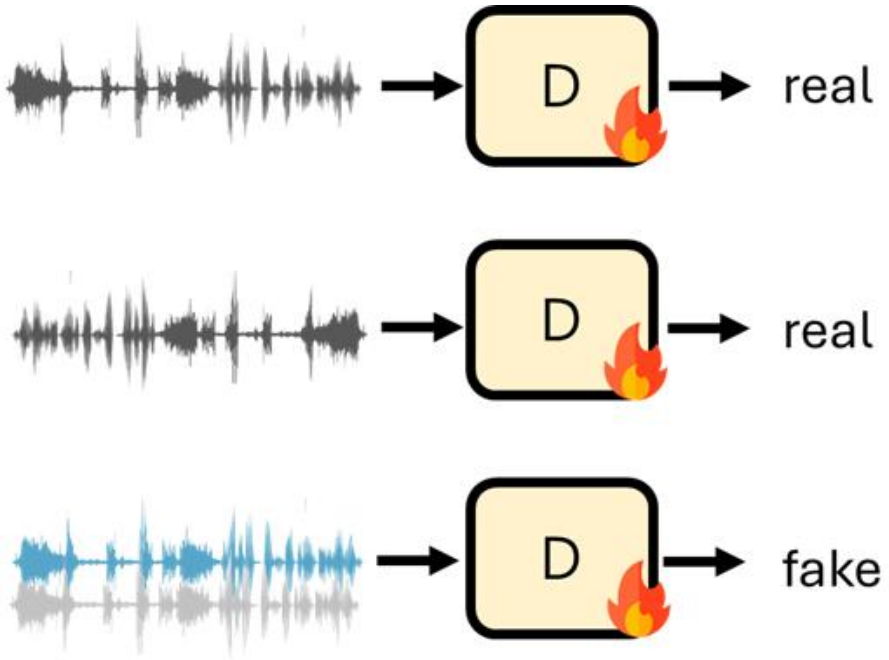
2. 感知上的像

3.模型難以分辨的像

Regression loss + **Perceptual loss** + **Adversarial Loss**

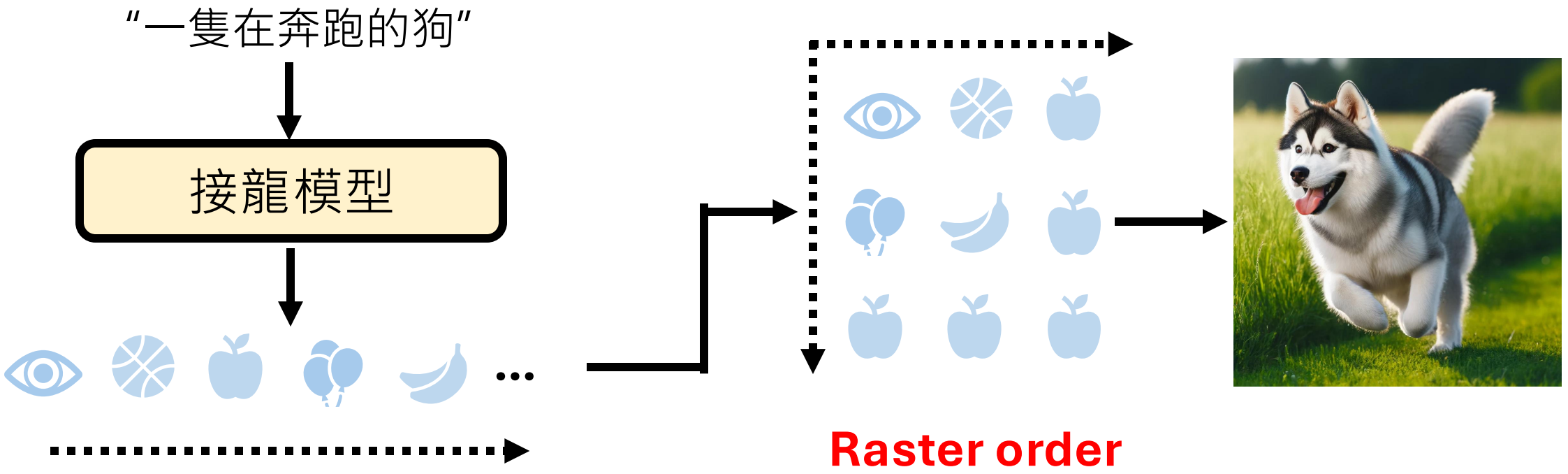






接龍一定要由左向右嗎？

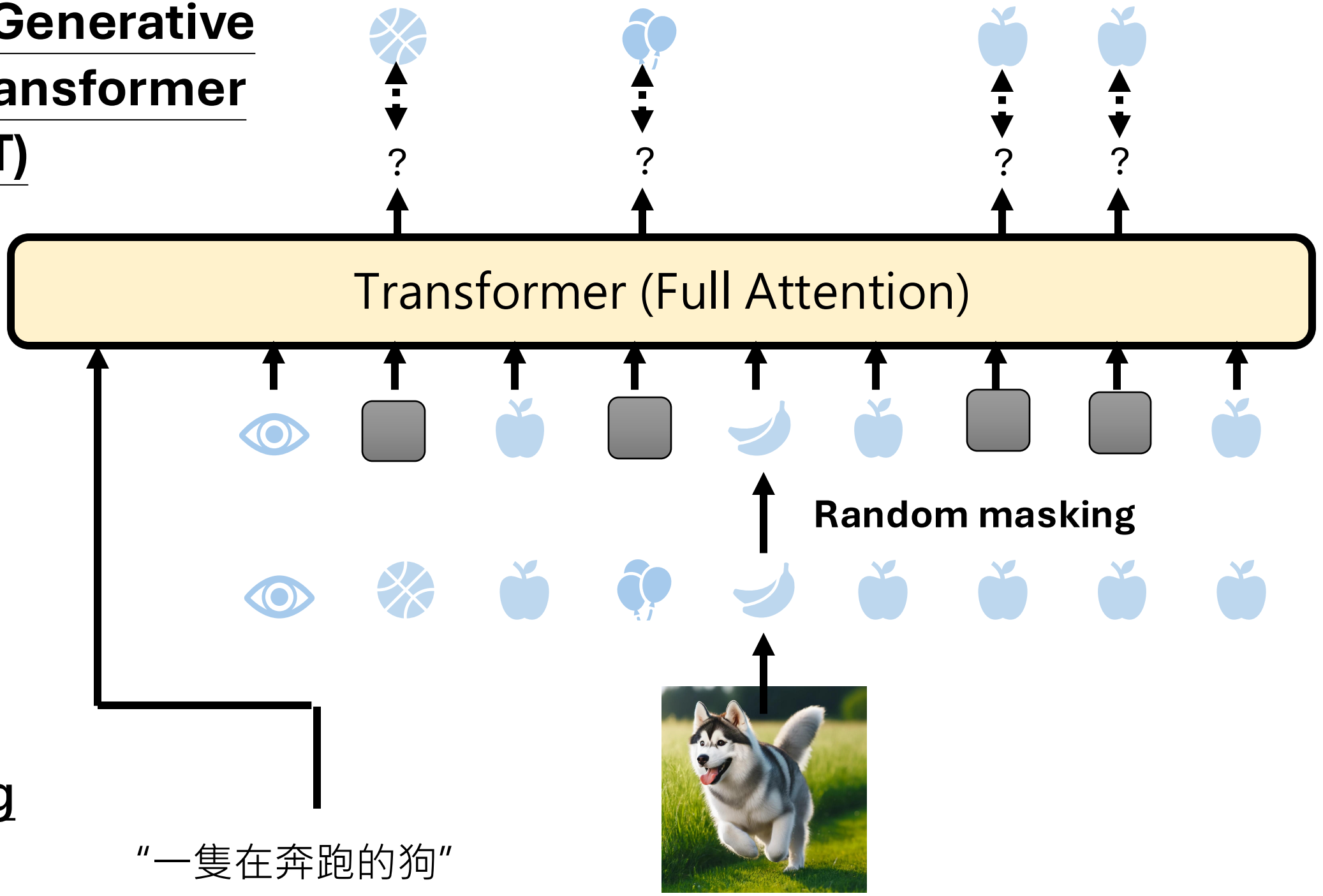
到目前為止接龍都是由左而右 (由上而下)

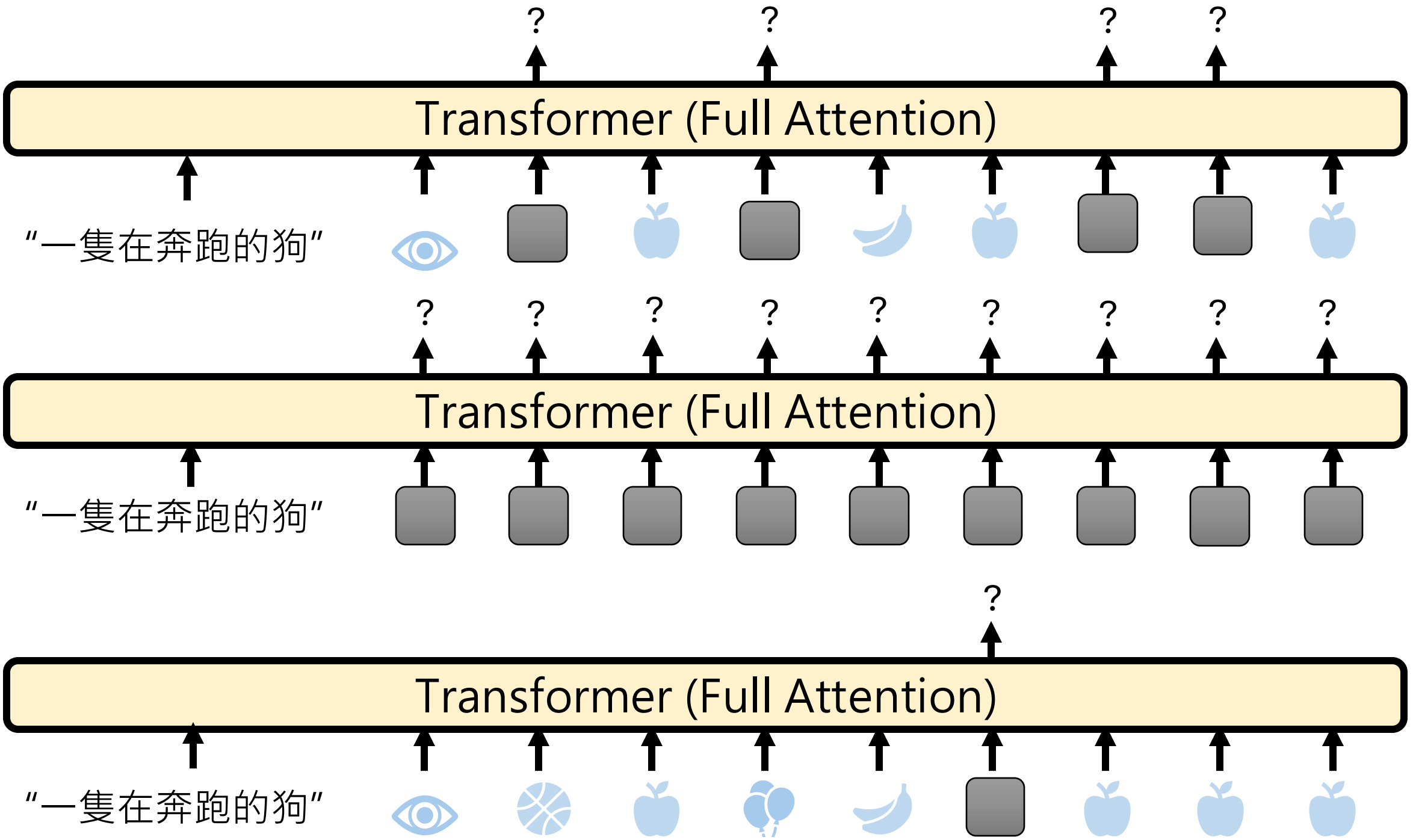


到目前為止接龍都是由左而右 (由上而下)

- **人類畫圖的時候通常不會由左而右、由上而下畫**

Masked Generative Image Transformer (MaskGIT)




Transformer (Full Attention)
"一隻在奔跑的狗"
Transformer (Full Attention)
"一隻在奔跑的狗"
Transformer (Full Attention)
"一隻在奔跑的狗"

Inference

Source of image:
https://arxiv.org/abs/2202.04200

Detokenizer

在左上角畫個頭

我要在中間畫個頭

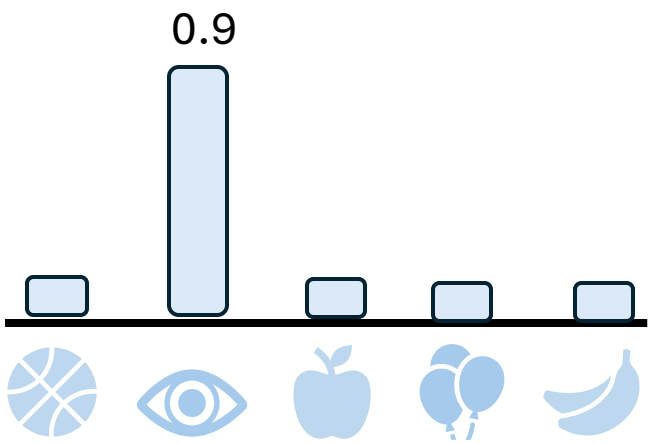
這些輸出是同時發生的

Transformer (Full Attention)

"一隻紅鶴"

已經知道圖片大小

Inference



關鍵是：只能留一小部分

0.9 0.6 0.5 0.6 0.8 0.5 0.7 0.5 0.6

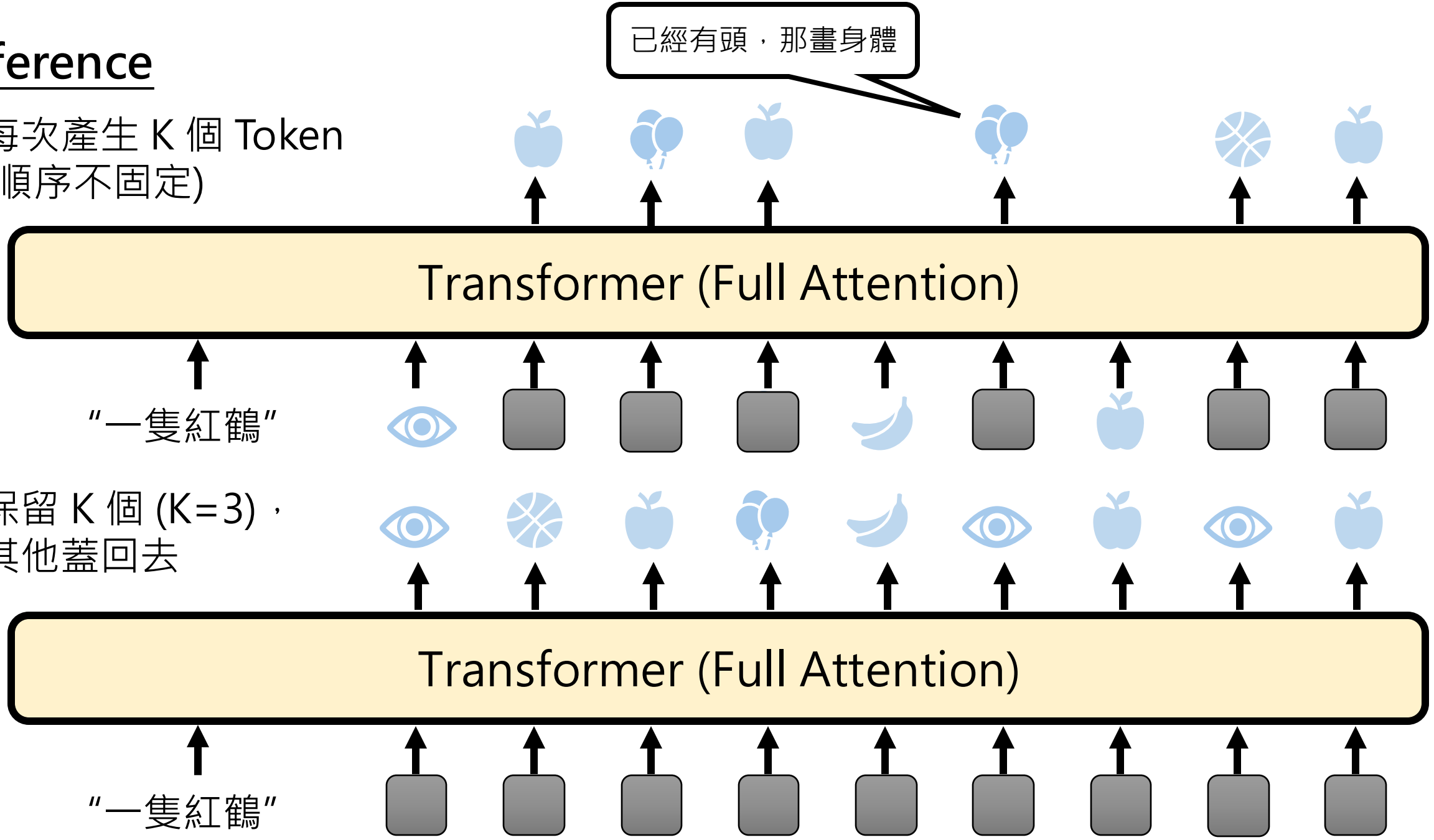
保留 K 個 (K=3)，
其他蓋回去

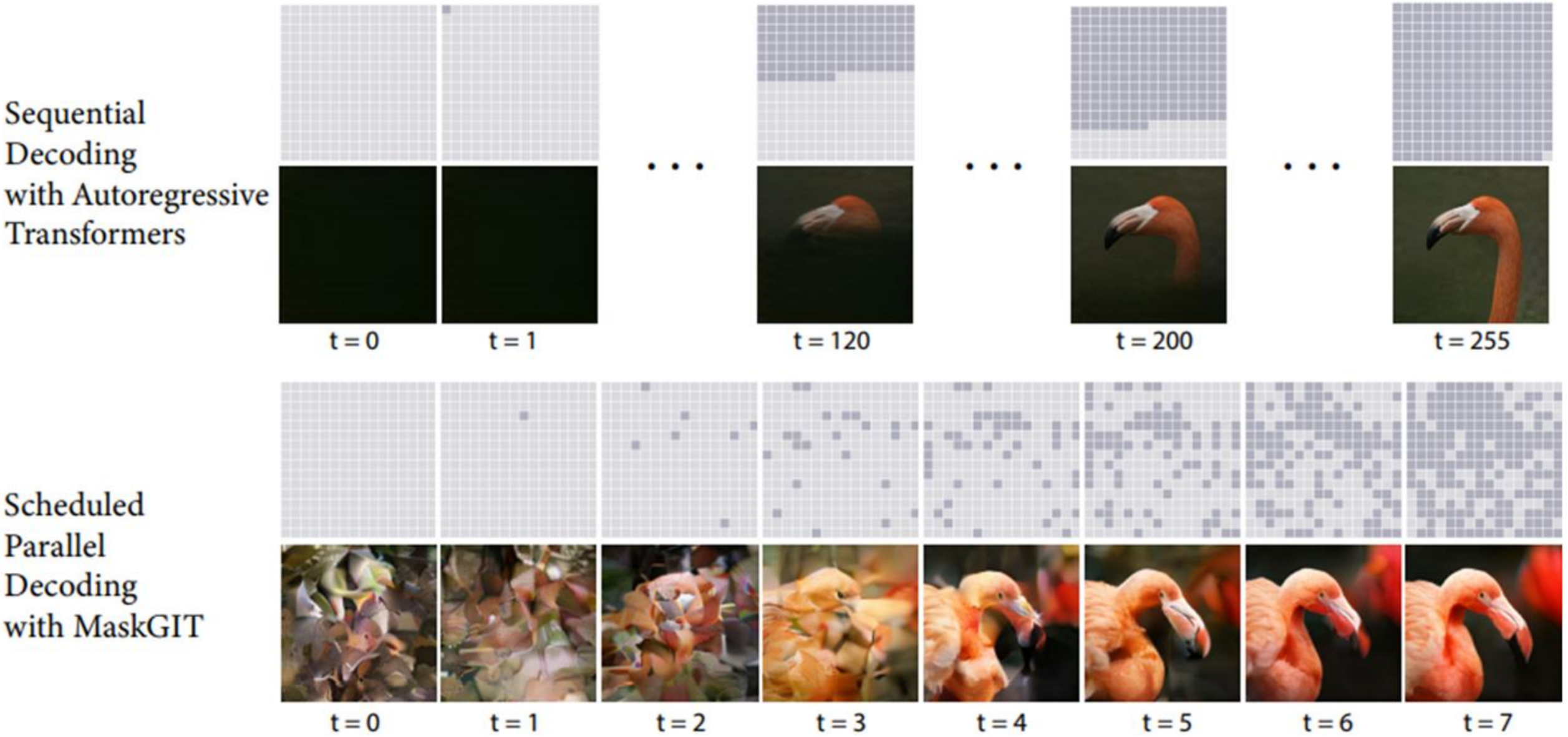
Transformer (Full Attention)

"一隻紅鶴"

Inference

每次產生 K 個 Token
(順序不固定)





https://arxiv.org/abs/2202.04200

另外一層的 Autoregressive Generation

- 從草圖開始再加上細節

"一隻在奔跑的狗"

接龍模型

64

64

Next Scale Prediction

接龍模型

256

256

接龍模型

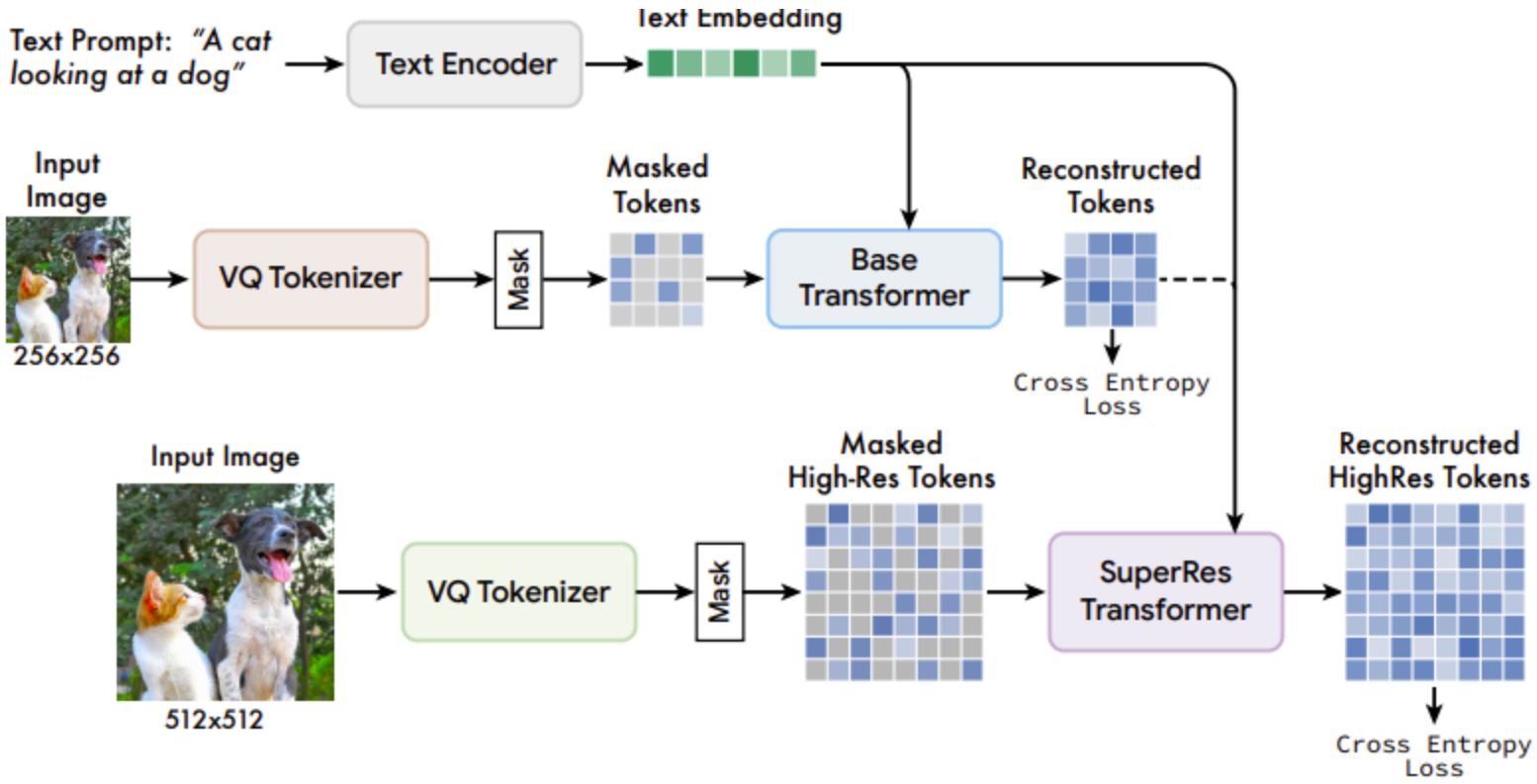
1024

1024

Next Token Prediction

MUSE

https://arxiv.org/pdf/2301.00704



Visual Autoregressive Modeling (VAR)

https://arxiv.org/abs/2404.02905

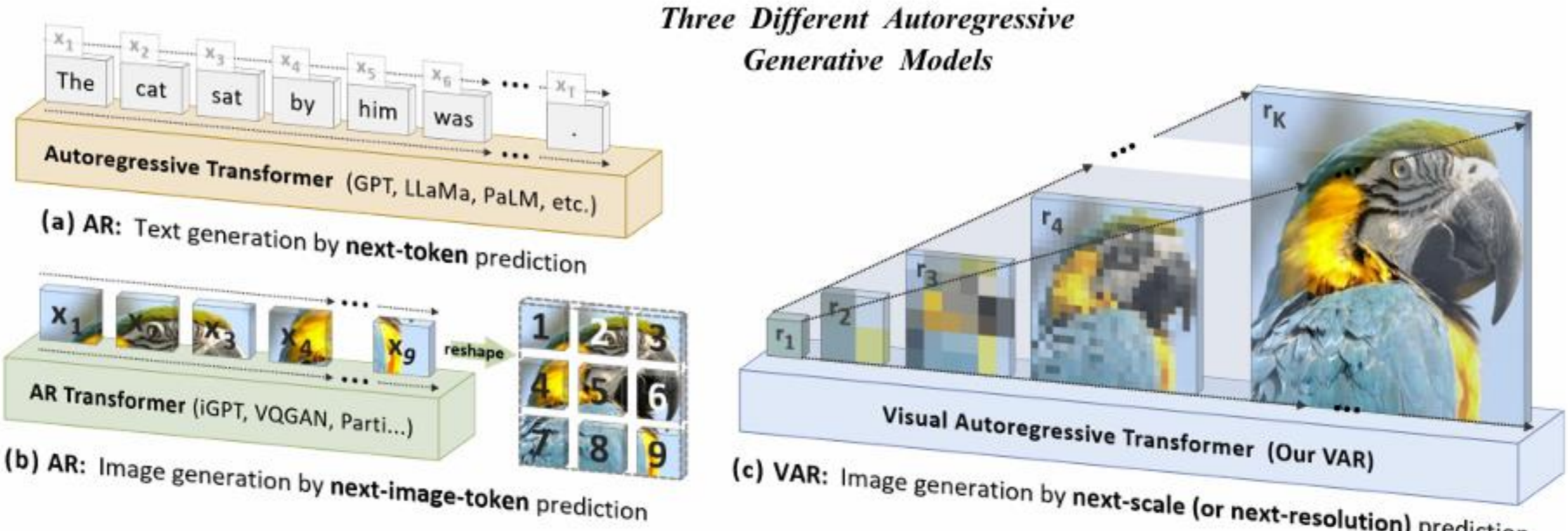
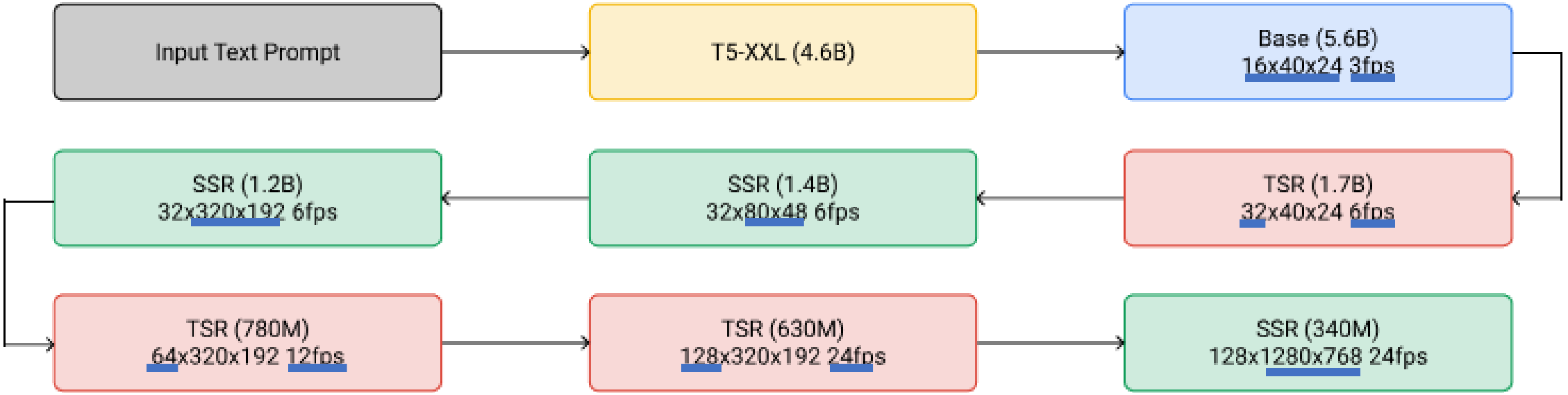


Imagen Video

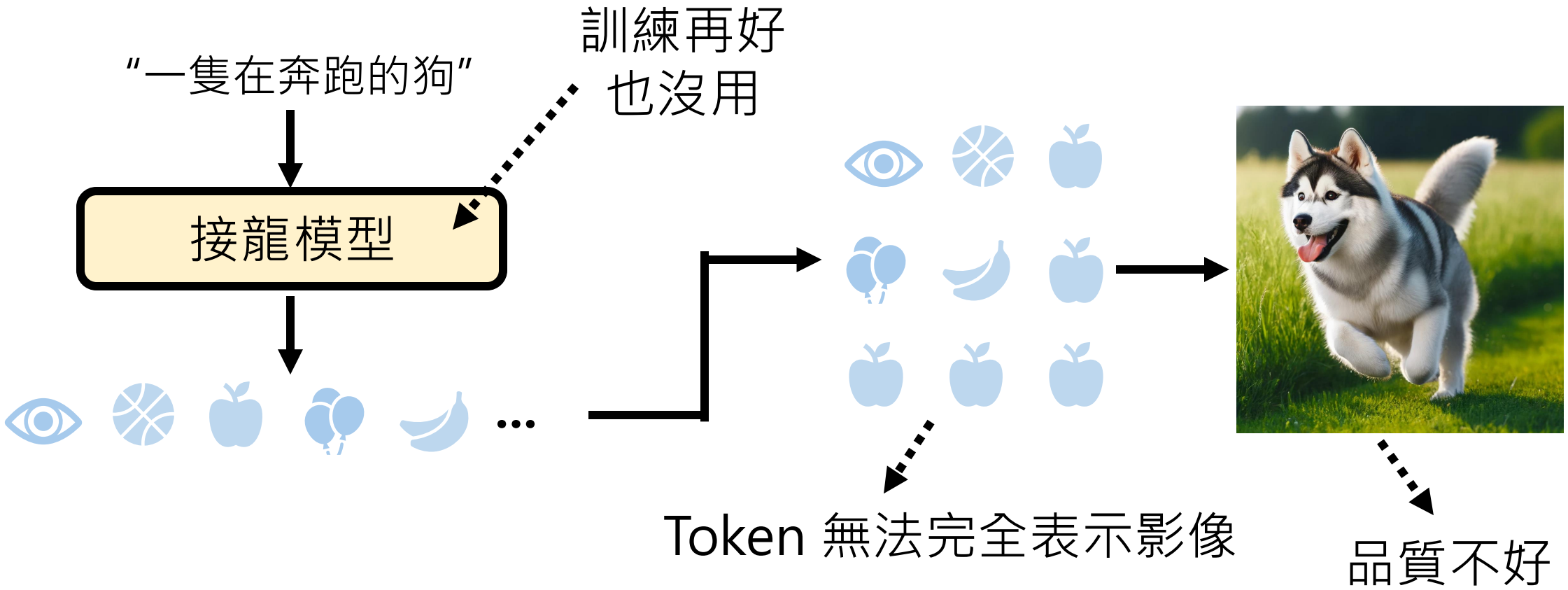
https://arxiv.org/abs/2210.02303

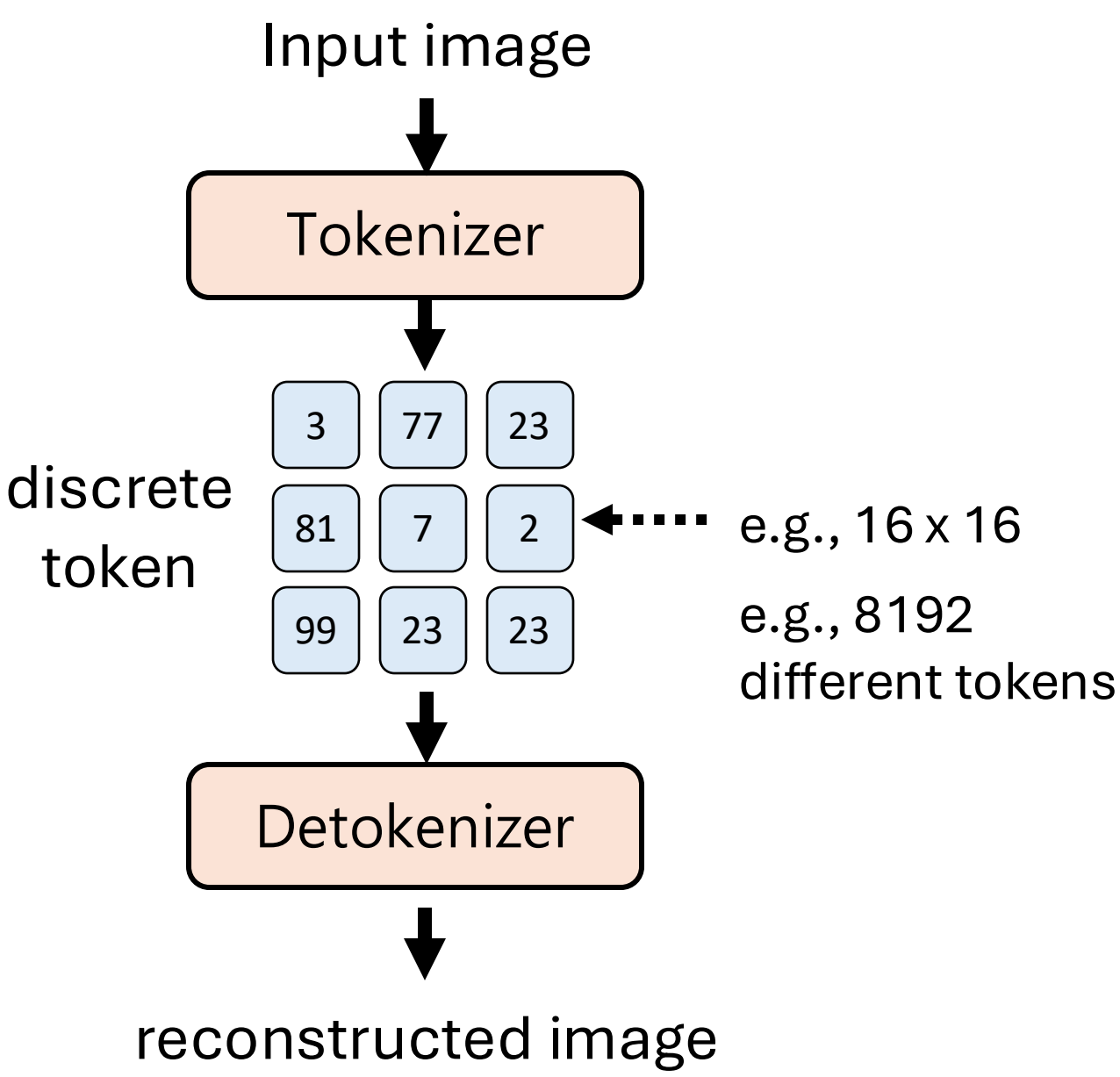


Token 的極限

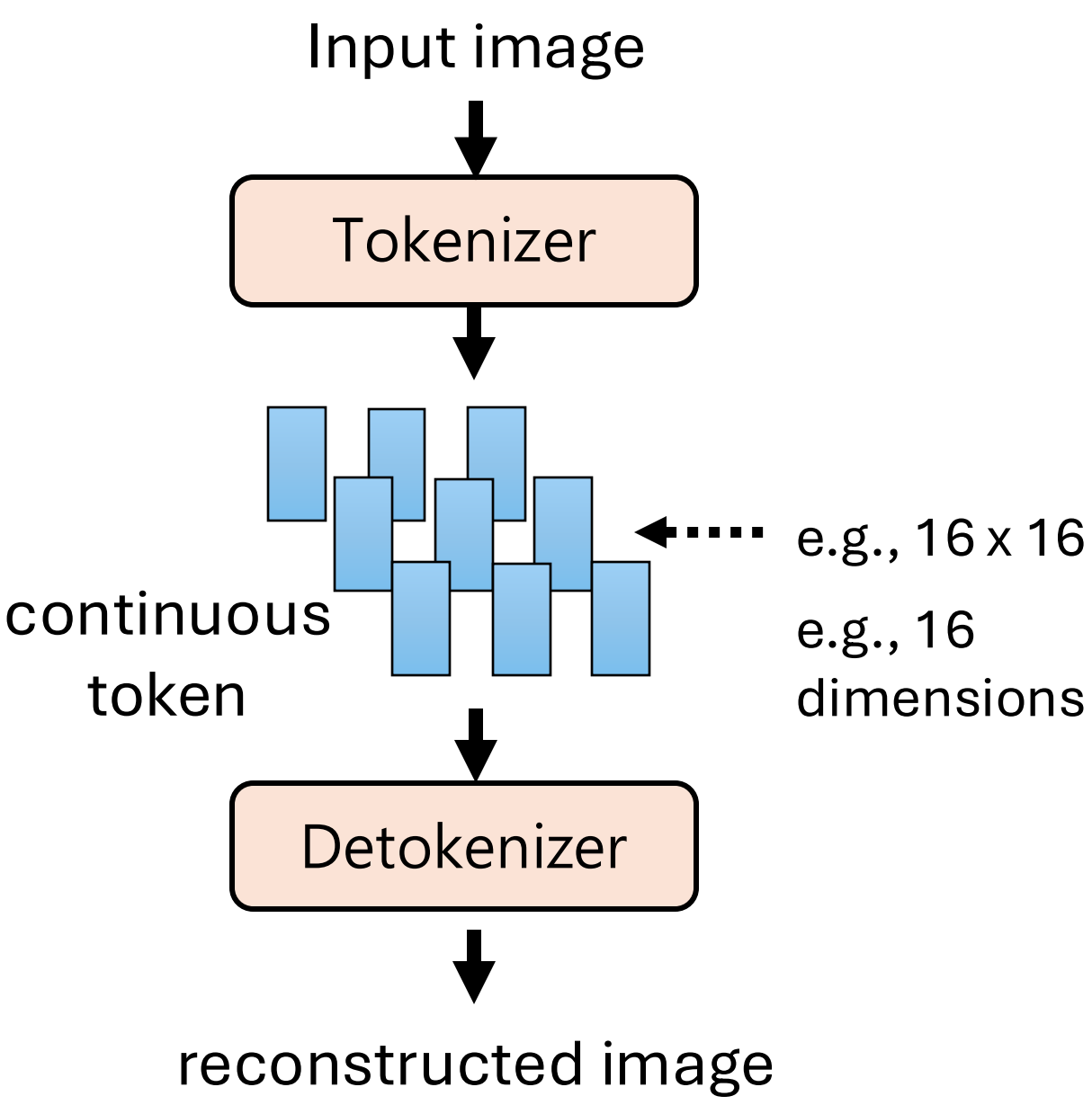
Token 限制了生成的品質

https://arxiv.org/abs/2312.02116
https://arxiv.org/abs/2410.13863



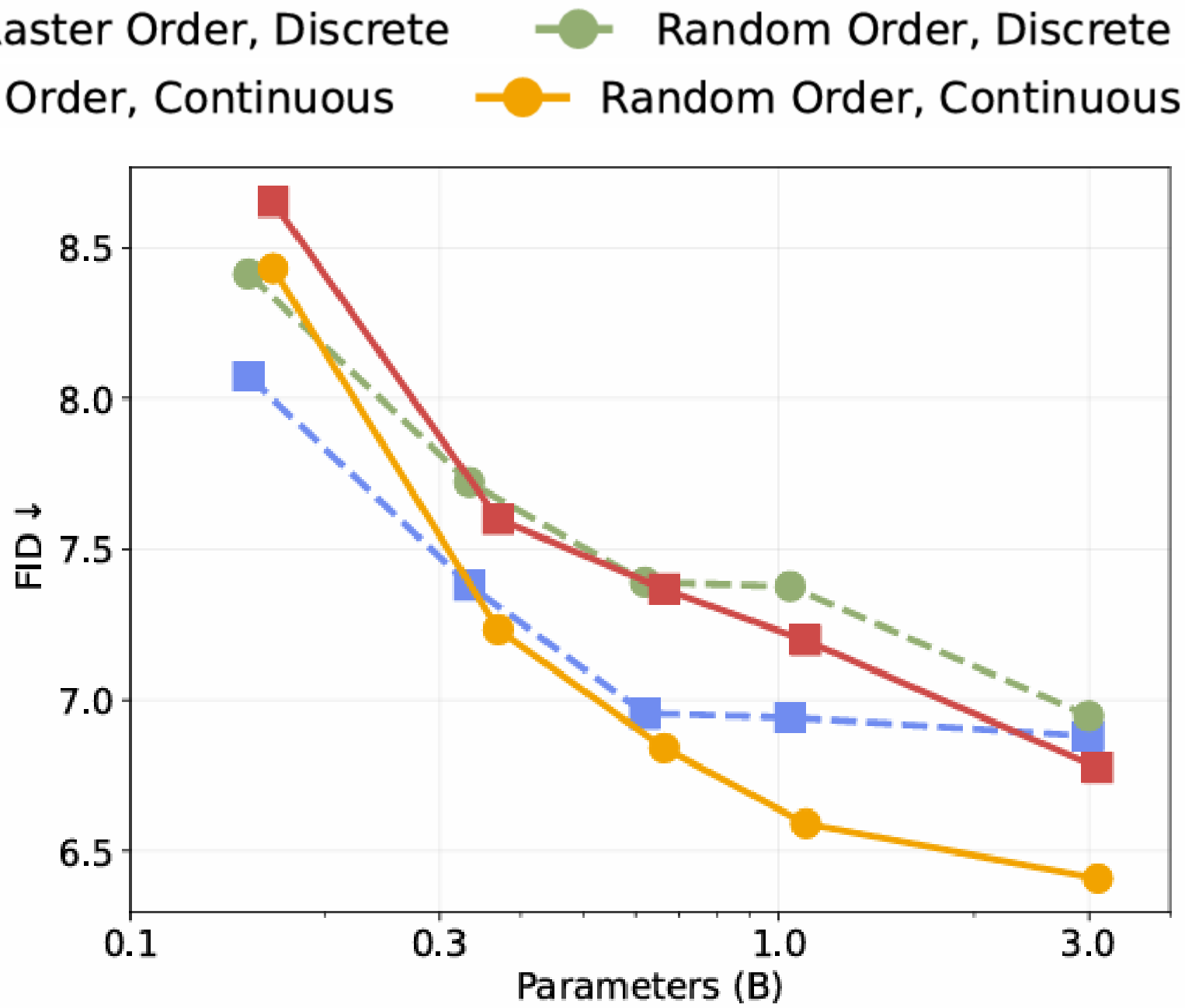


Source of image: https://arxiv.org/abs/2410.13863



Source of image: https://arxiv.org/abs/2410.13863

Perceptual Loss



https://arxiv.org/abs/2410.13863

能夠對 Continuous Token 做接龍嗎？

蒙娜麗莎的微笑

$\hat{z}_1$ $\hat{z}_2$ $\hat{z}_3$ $\hat{z}_4$ $\hat{z}_5$ $\hat{z}_6$ …

MSE?

蒙娜麗莎的微笑 → 接龍模型 → z_1 l_1 $\hat{z}_1$

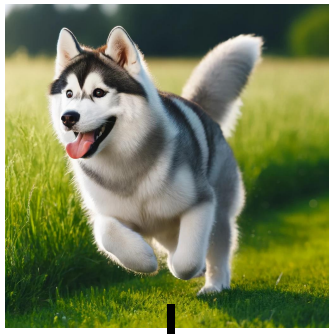
$$l_1 = \|z_1 - \hat{z}_1\|_2$$

蒙娜麗莎的微笑 $\hat{z}_1$ → 接龍模型 → z_2 l_2 $\hat{z}_2$

$$l_2 = \|z_2 - \hat{z}_2\|_2$$

$$L = l_1 + l_2 + \cdots$$

一隻在奔跑的狗



…

一隻在奔跑的狗

…

一隻在奔跑的狗 → 接龍模型 → 一隻在奔跑的狗

一隻在奔跑的狗 → 接龍模型 → 一隻在奔跑的狗

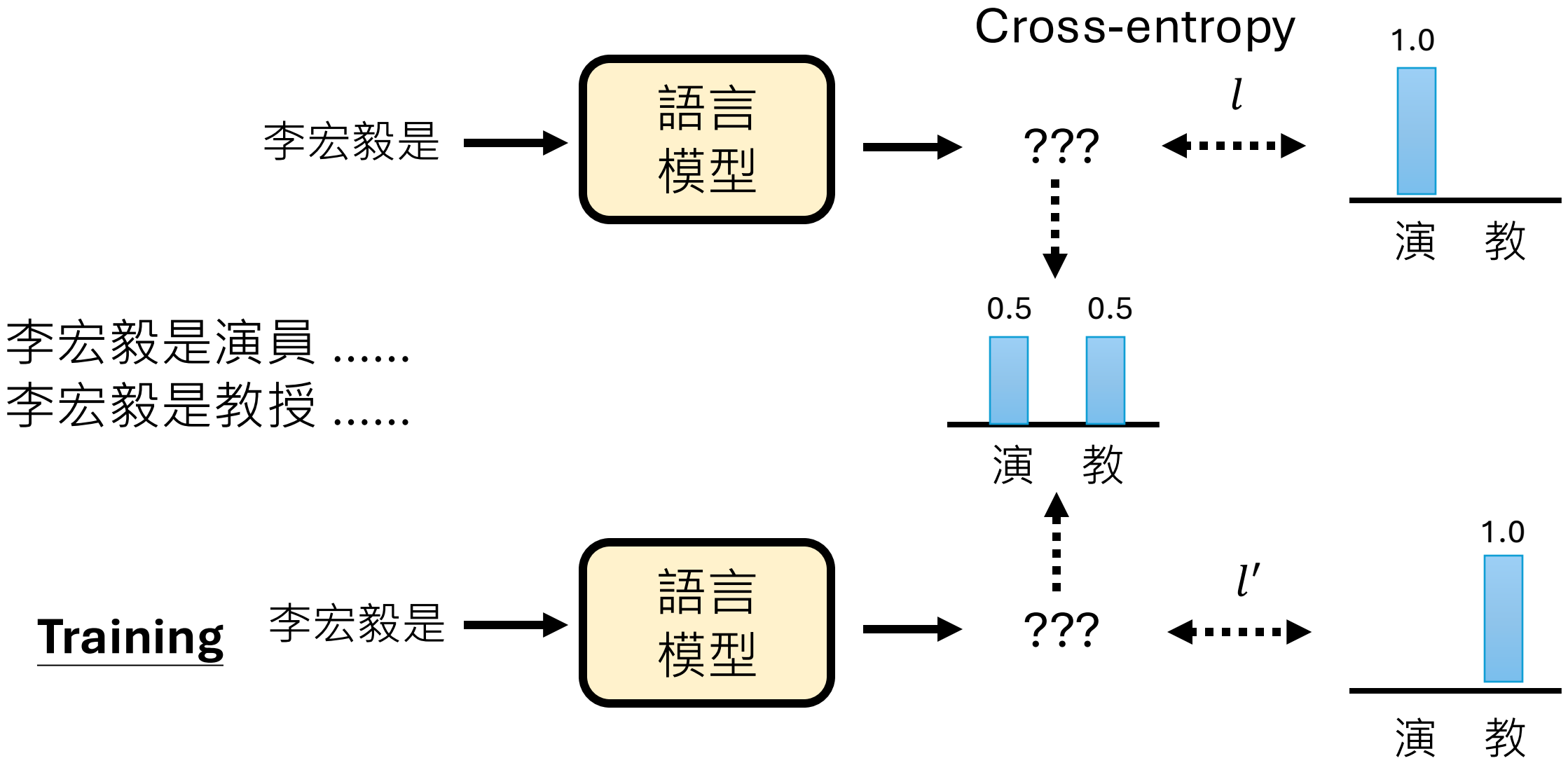
MSE

l

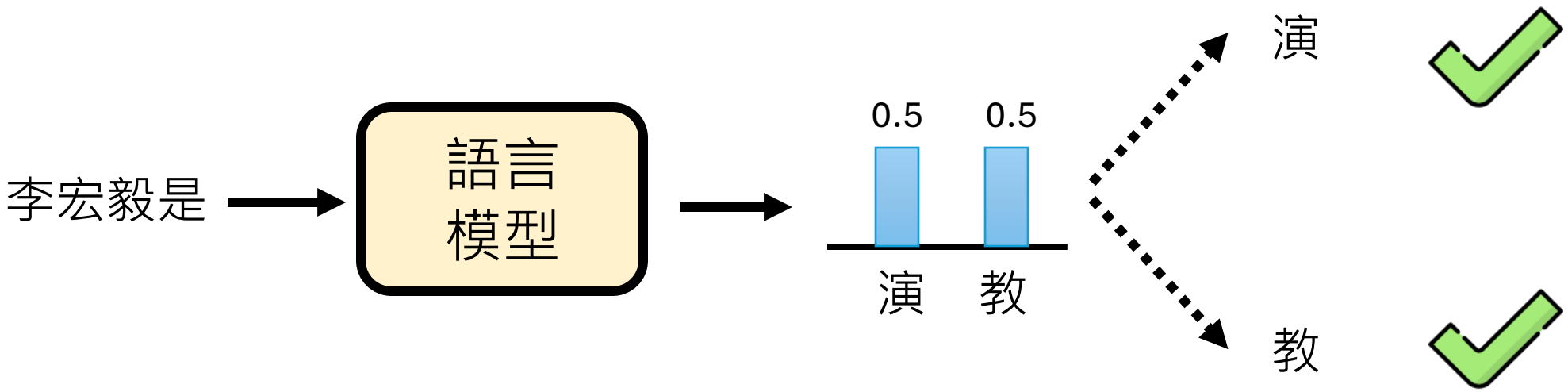
l'

$L = l + l' + \cdots$

之前訓練文字模型的時候沒有這種問題嗎？

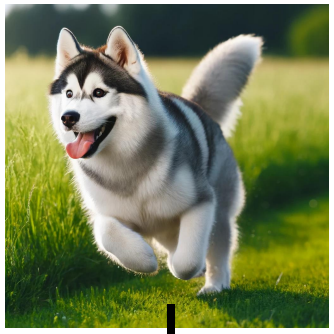


之前訓練文字模型的時候沒有這種問題嗎？



Inference

一隻在奔跑的狗



...

一隻在奔跑的狗

...

一隻在奔跑的狗 → 接龍模型 → 一隻在奔跑的狗

一隻在奔跑的狗 → 接龍模型 → 一隻在奔跑的狗

MSE

l

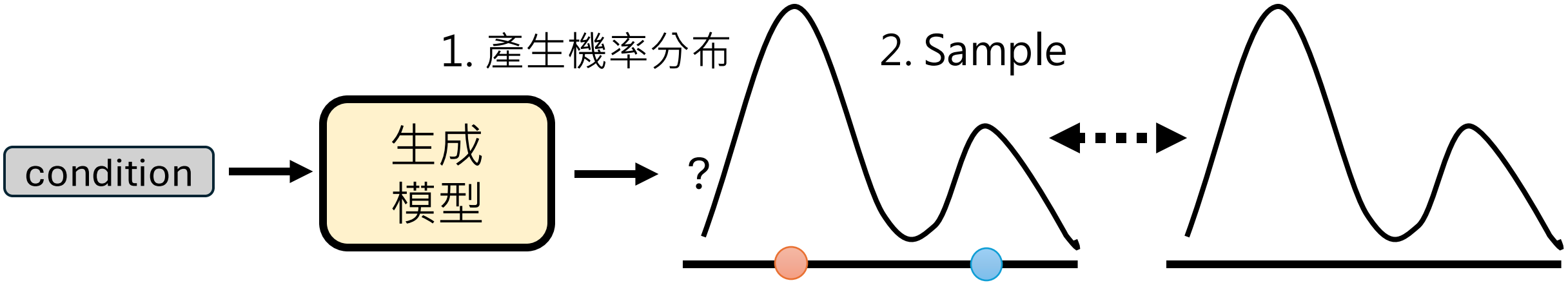
l'

使用 continuous token 時，不能用 MSE，需要用其他方法來定義 Loss

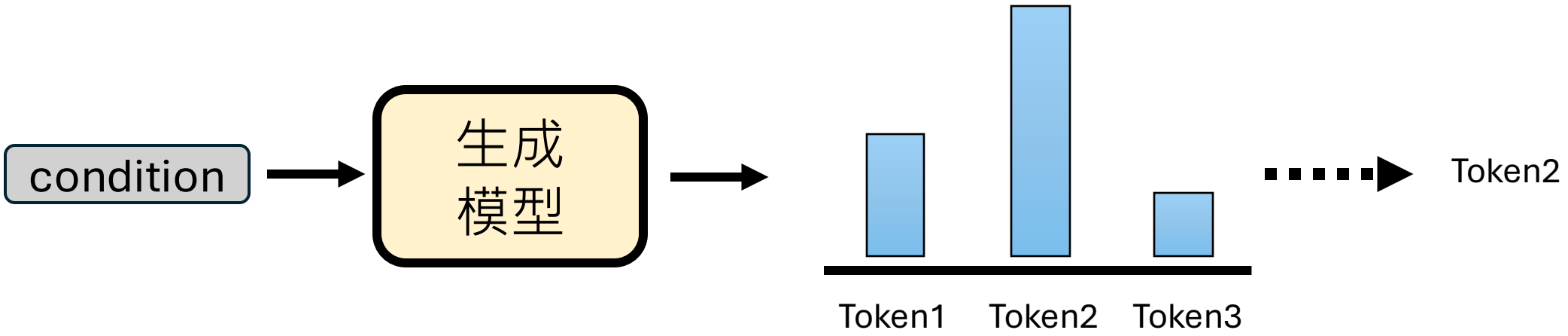
另外一系列生成的方法

VAE, GAN, Diffusion, Flow, etc.

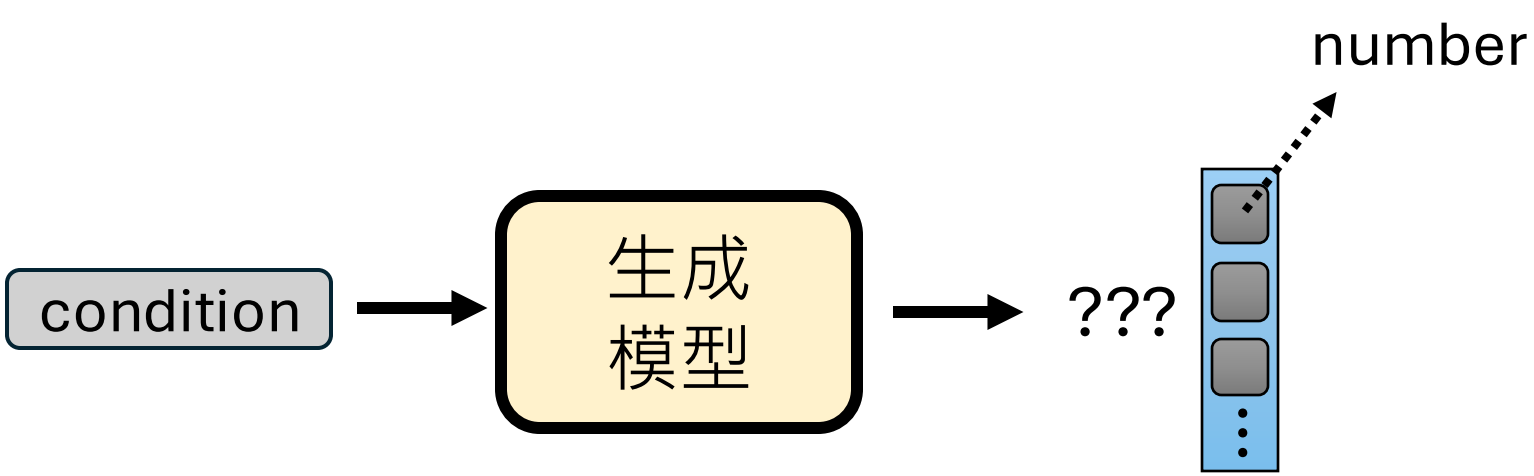
這個世界沒有標準答案 ...

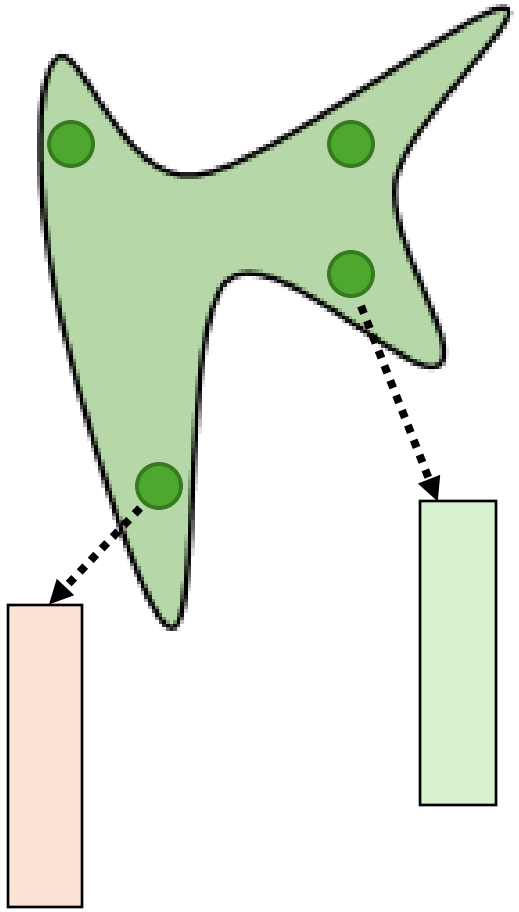


如果生成的目標是 Discrete Token，這個問題自然就被解決了!



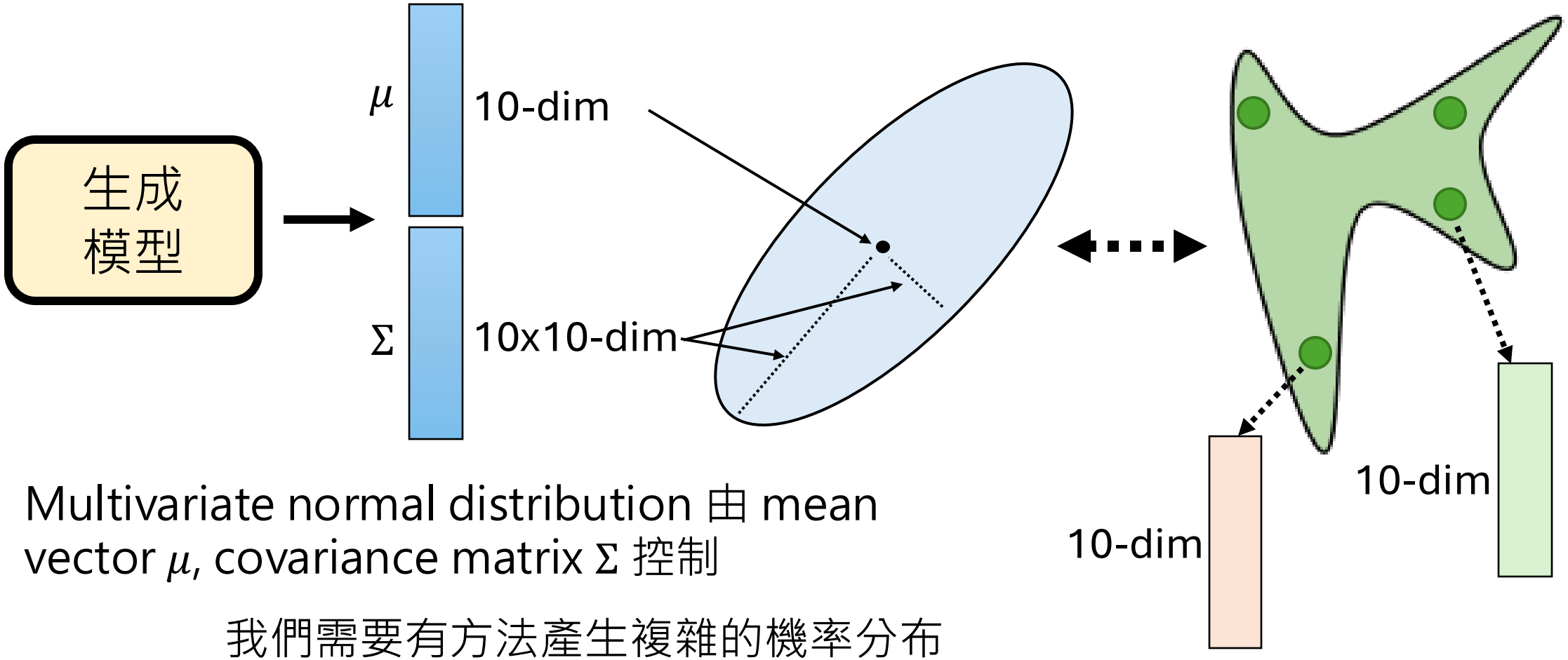
如果生成的目標是連續的 (Continuous)





如果生成的目標是連續的 (Continuous)

把生成模型的輸出當作是控制機率分布的參數



Multivariate normal distribution 由 mean vector μ, covariance matrix Σ 控制

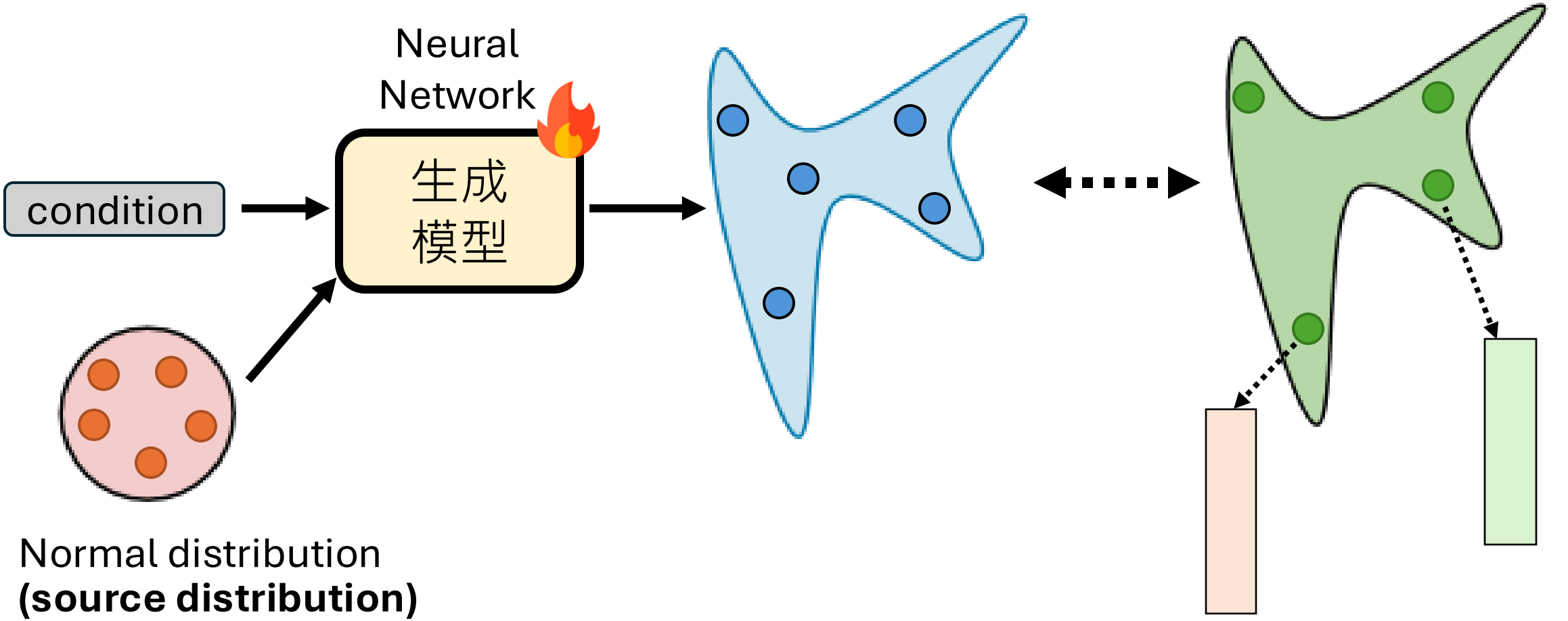
我們需要有方法產生複雜的機率分布

Generative Model

VAE, GAN, Diffusion, Flow, etc.

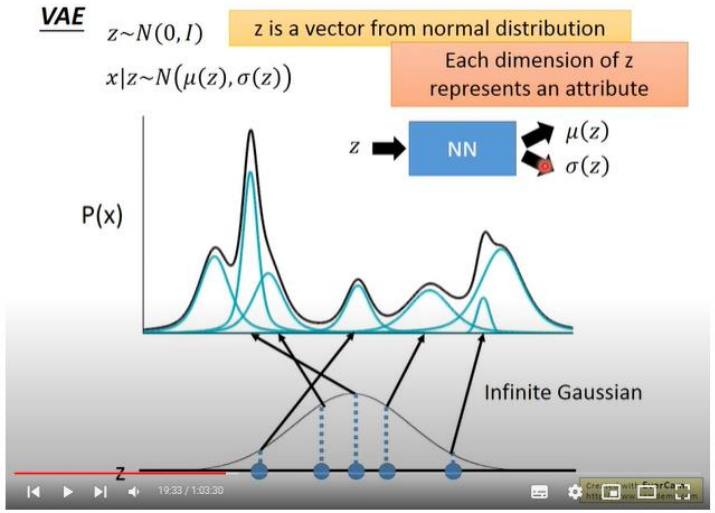
As close as possible

(target distribution)



Normal distribution
(source distribution)

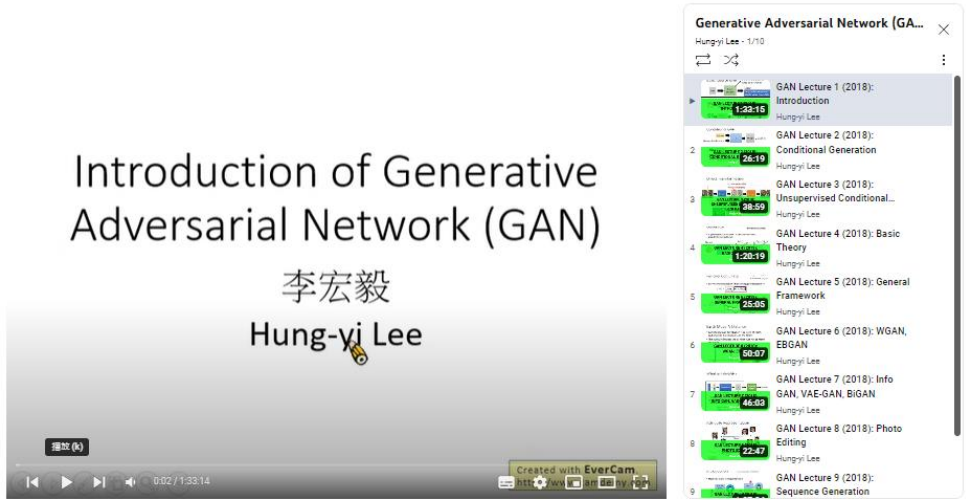
VAE



ML Lecture 18: Unsupervised Learning - Deep Generative Model (Part II)

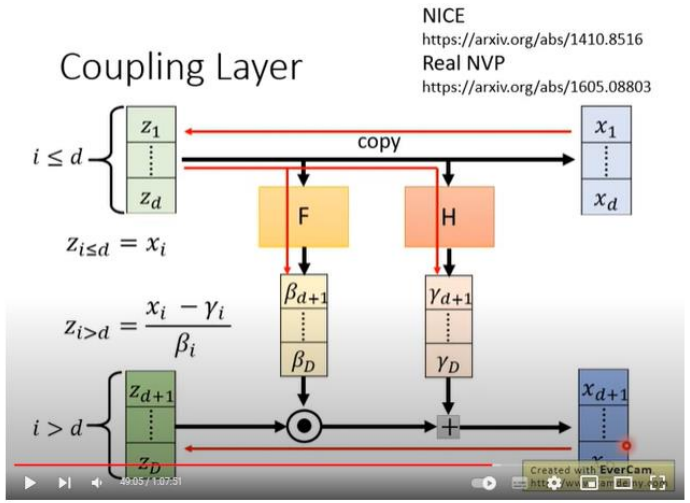
https://youtu.be/8zomhgKrsmQ
(2016 機器學習)

GAN



https://www.youtube.com/watch?v=DQNNMiAP5lw&list=PLJV_el3uVTsMq6JEFPW35BCiOQTsoqwNw (2018 機器學習及其深層與結構化)

Normalizing Flow



Flow-based Generative Model

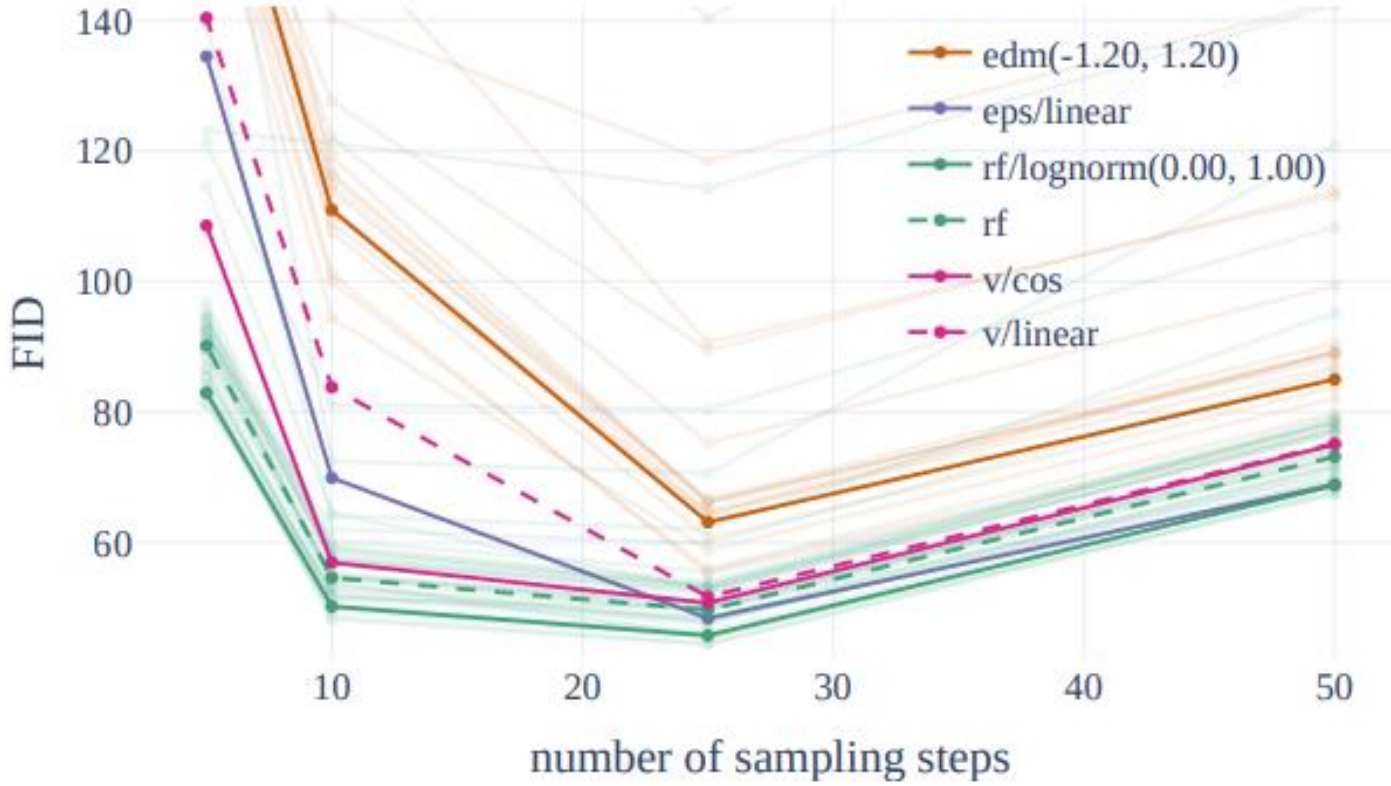
https://youtu.be/uXY18nzdSsM
(2019 機器學習)

Diffusion

https://www.youtube.com/watch?v=azBugJzmz-o&list=PLJV_el3uVTsNi7PgekEUFsyVllAJXRsP- (2023 機器學習)

Flow Matching

- Stable Diffusion 3
  - https://arxiv.org/abs/2403.03206
- Meta Movie Gen
  - https://arxiv.org/abs/2403.03206
- Flux.2
  - https://bfl.ai/blog/flux-2 (NOVEMBER 25, 2025)



Flow Matching

https://diffusion.csail.mit.edu/2025/index.html

Introduction to Flow Matching and Diffusion Models

MIT Computer Science Class 6.S184: Generative AI with Stochastic Differential Equations

Diffusion and flow-based models have become the state of the art for generative AI across a wide range of data modalities, including images, videos, shapes, molecules, music, and more! This course aims to build up the mathematical framework underlying these models from first principles. At the end of the class, students will have built a toy image diffusion model from scratch, and along the way, will have gained hands-on experience with the mathematical toolbox of stochastic differential equations that is useful in many other fields. This course is ideal for students who want to develop a principled understanding of the theory and practice of generative AI.

Course Notes

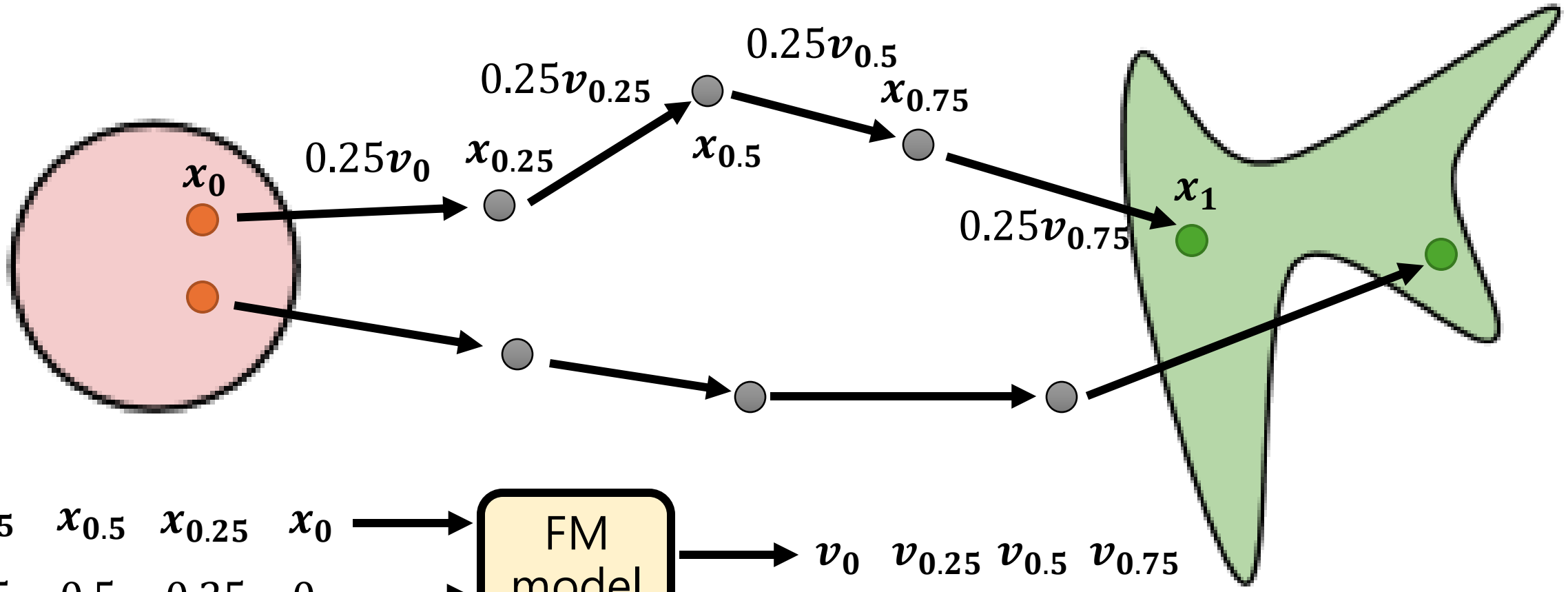
The course notes serve as the backbone of the course and provide a self-contained explanation of all material in the class. In contrast, lectures slides will generally not be self-contained and are intended to provide accompanying visualizations during the lecture. You may view the notes by clicking on the colored link below.

View the course notes here!

(Ignore the condition in the slides.)

Flow Matching

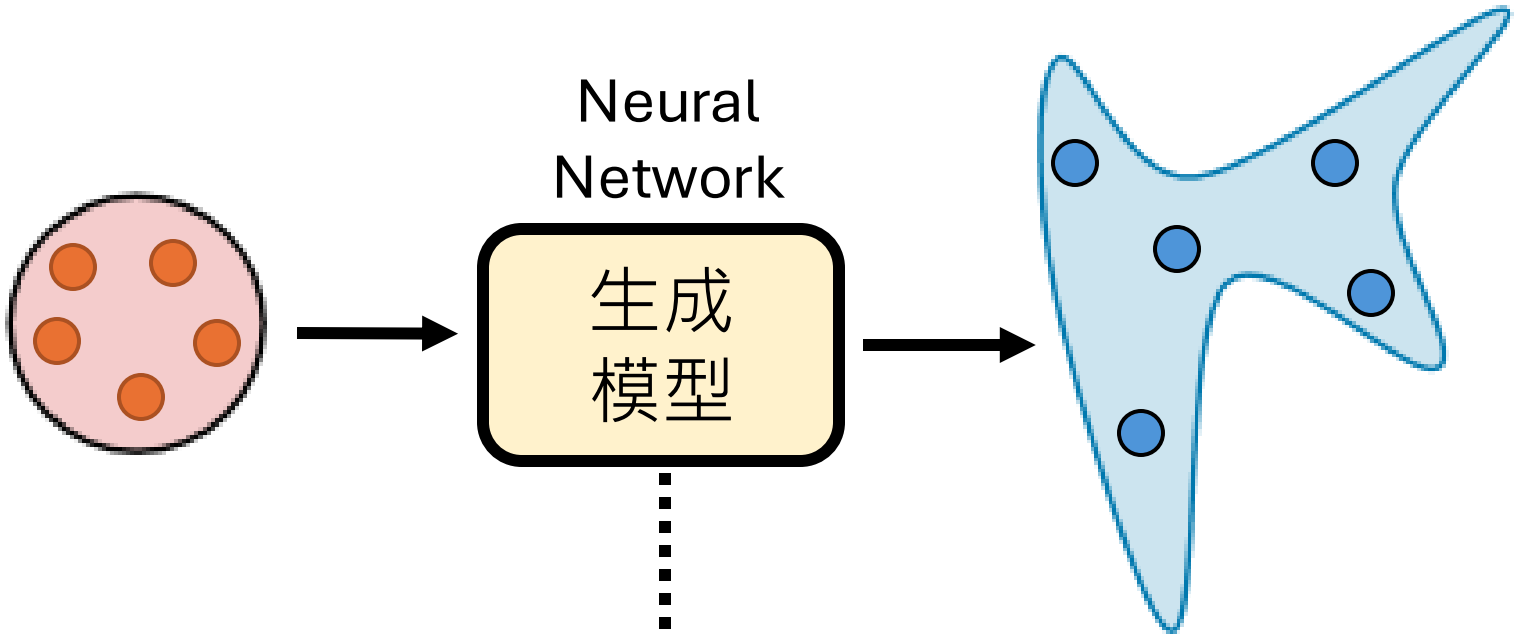
先決定從 source distribution 變到 target distribution 需要幾步 (例如：4步)



Vector Field (給一個位置、一個時間，指示接下來往哪裡走)

(Ignore the condition in the slides.)

Flow Matching



步數可以看做是「層數」

x_0 → FM model → $v_0 \times 0.25 + x_0 = x_{0.25}$

0 → FM model

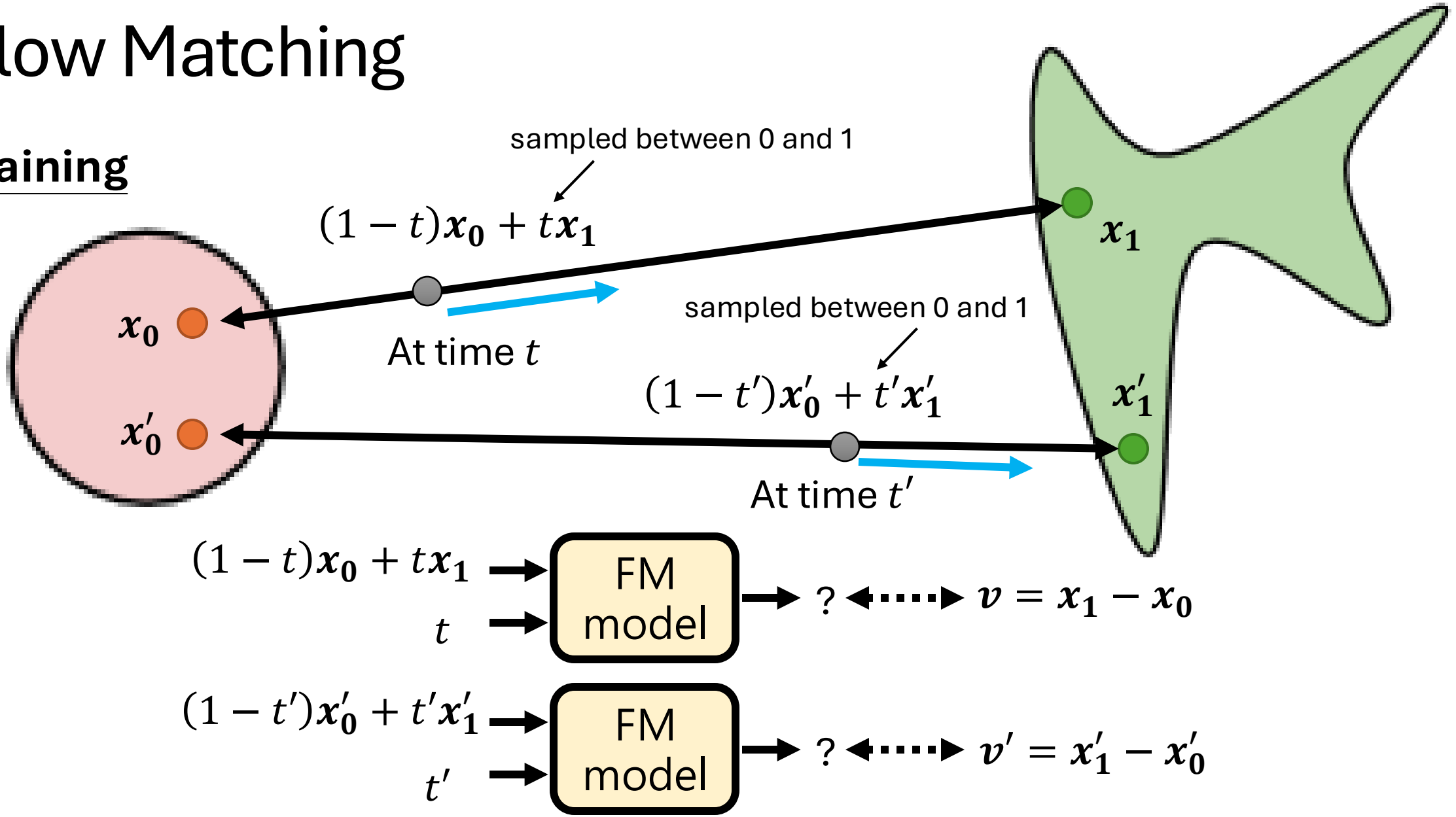
$x_{0.25}$ → FM model → $v_{0.25} \times 0.25 + x_{0.25}$ ……

0.25 → FM model

(Ignore the condition in the slides.)

Flow Matching

Training

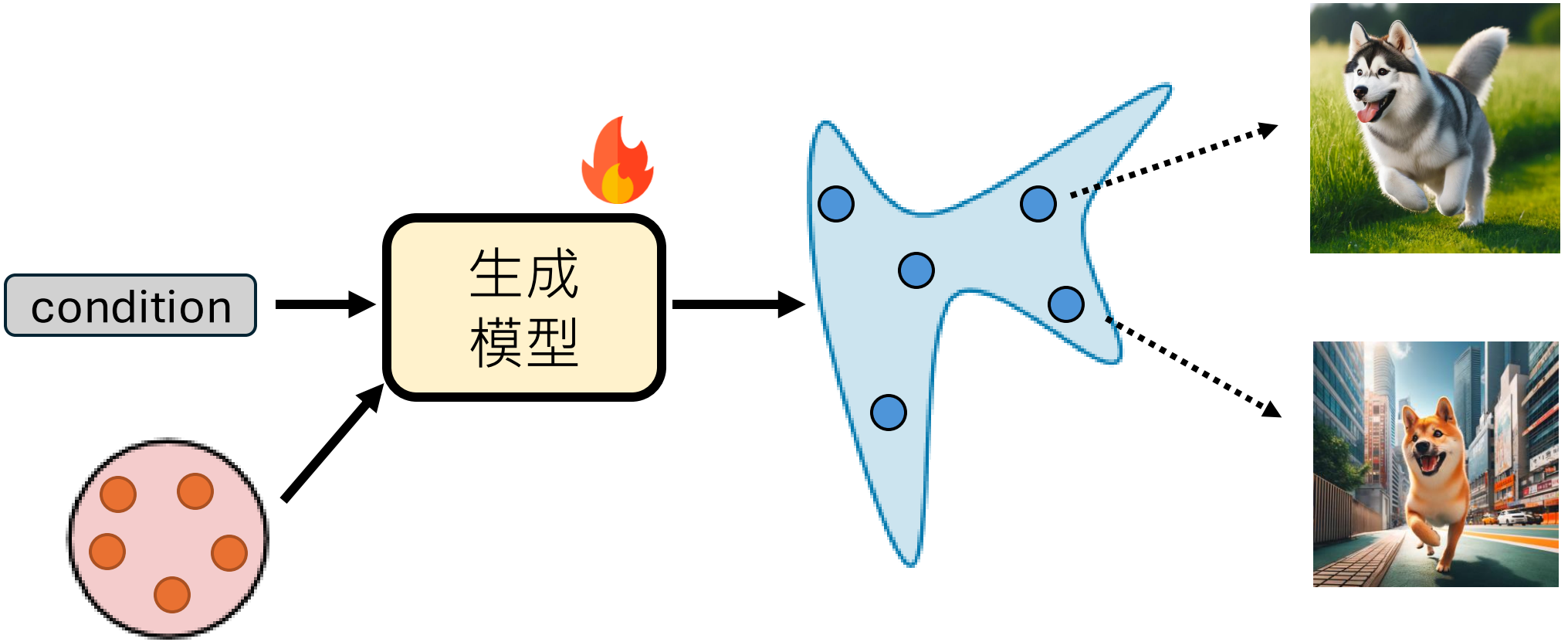


範例程式

- https://colab.research.google.com/drive/162Z6c_3d3GCRBeRwCjS7N_mrIonccEgz?usp=sharing

Generative Model

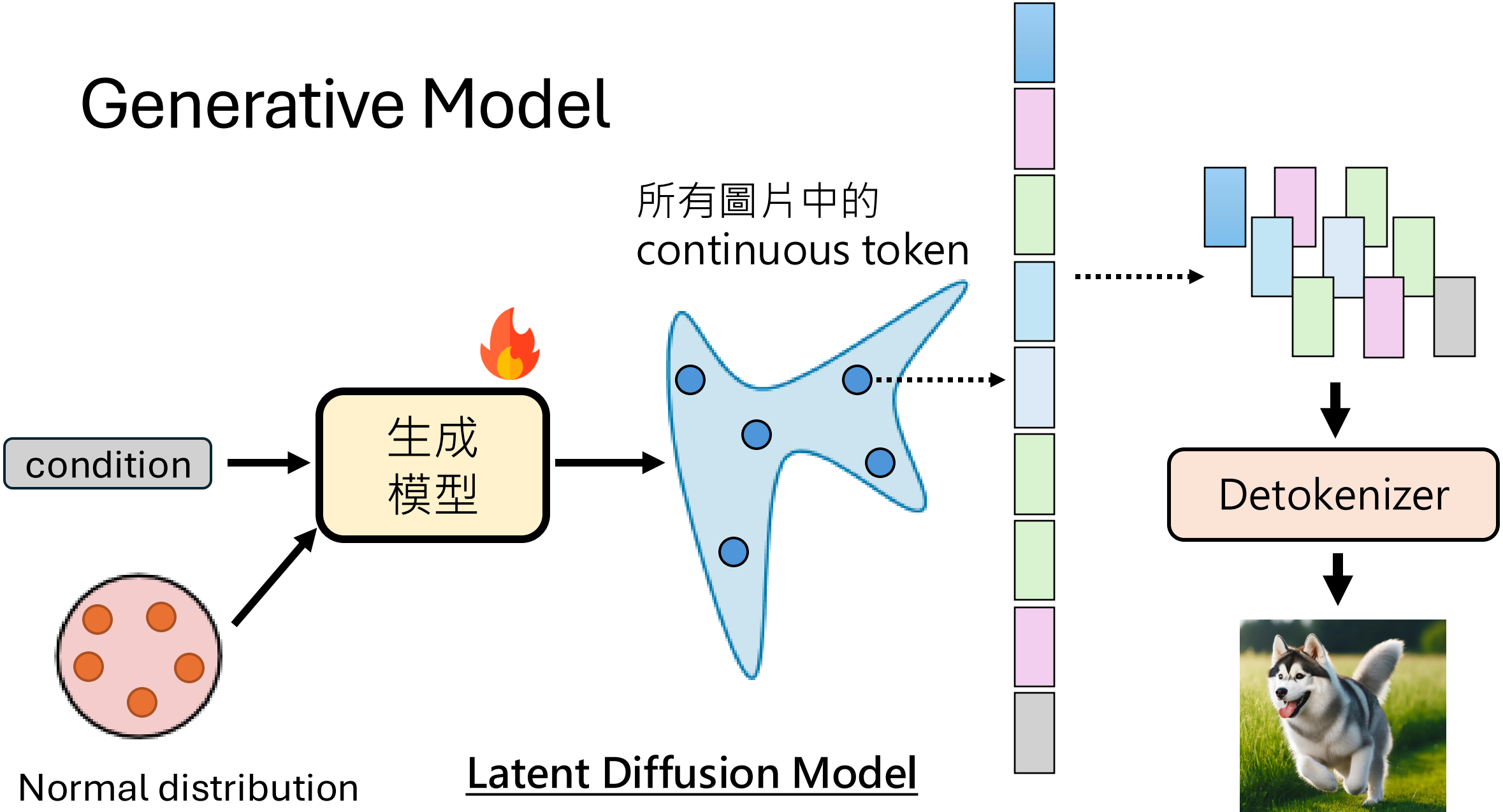
一張圖片可以被視為一個向量



Normal distribution

在 2022 年以前，影像生成就是這麼做的

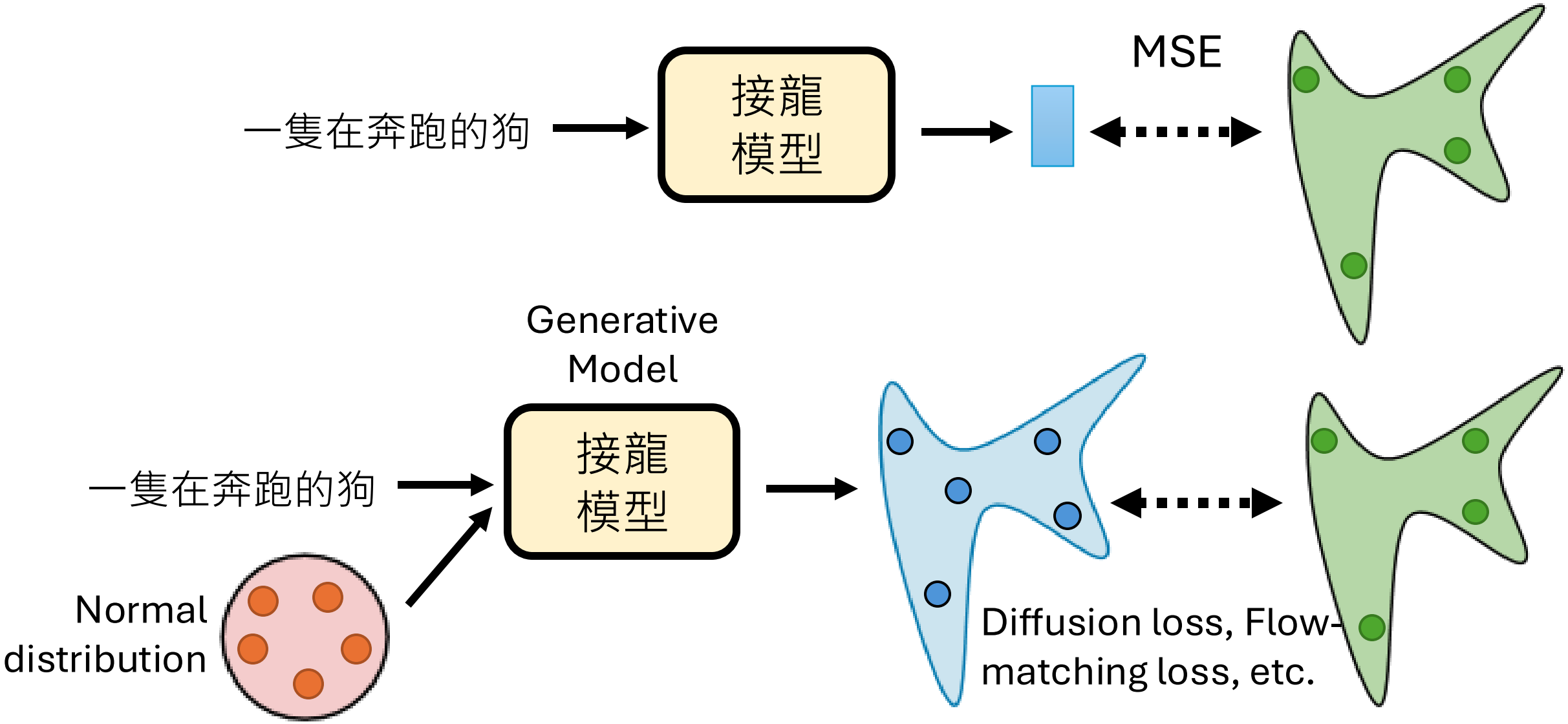
Generative Model



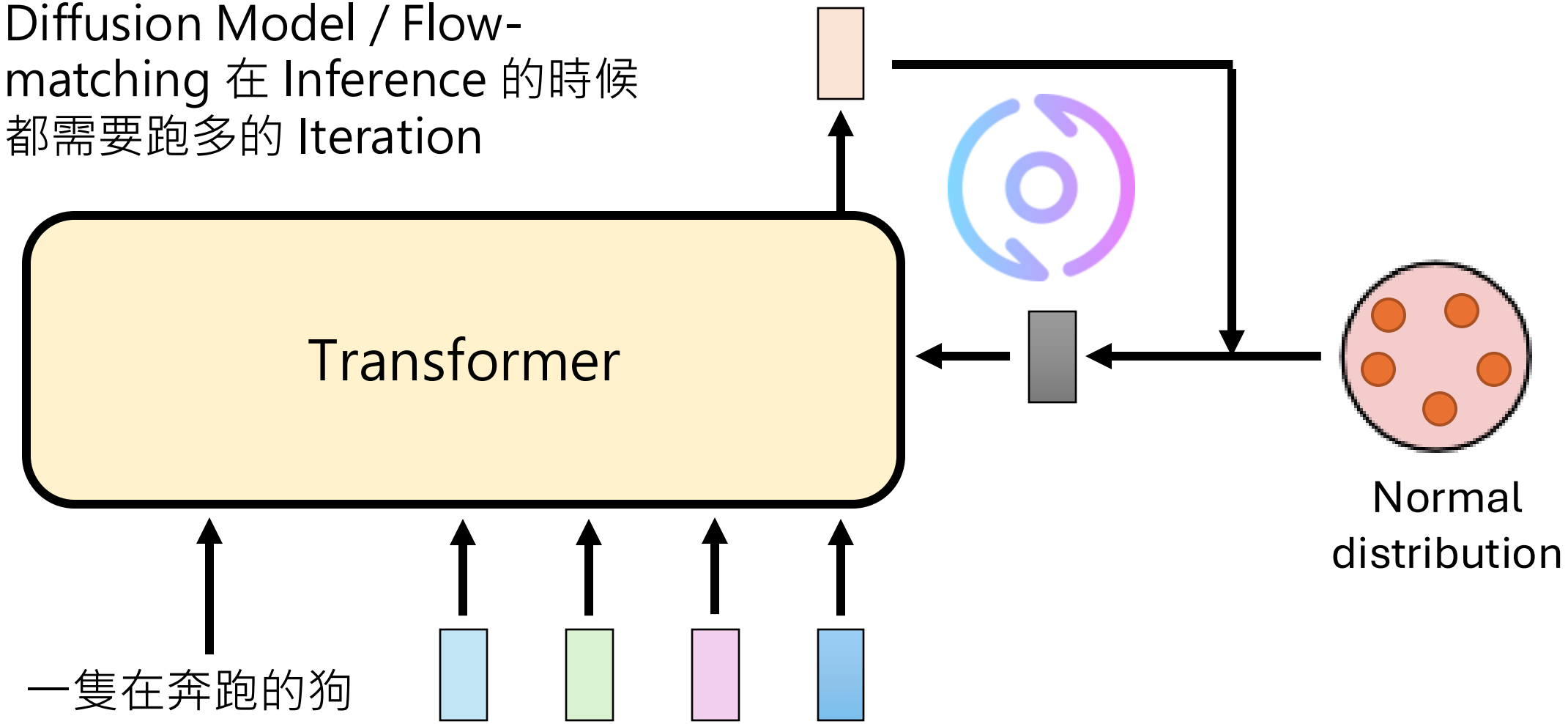
https://youtu.be/JbfcAaBT66U?si=p_clS-G3SJAUJZ1A

兩條世界線的匯聚

輸出 continuous token 的困難可以用 generative model 來解



但是 Inference 的時候遇到問題 ...

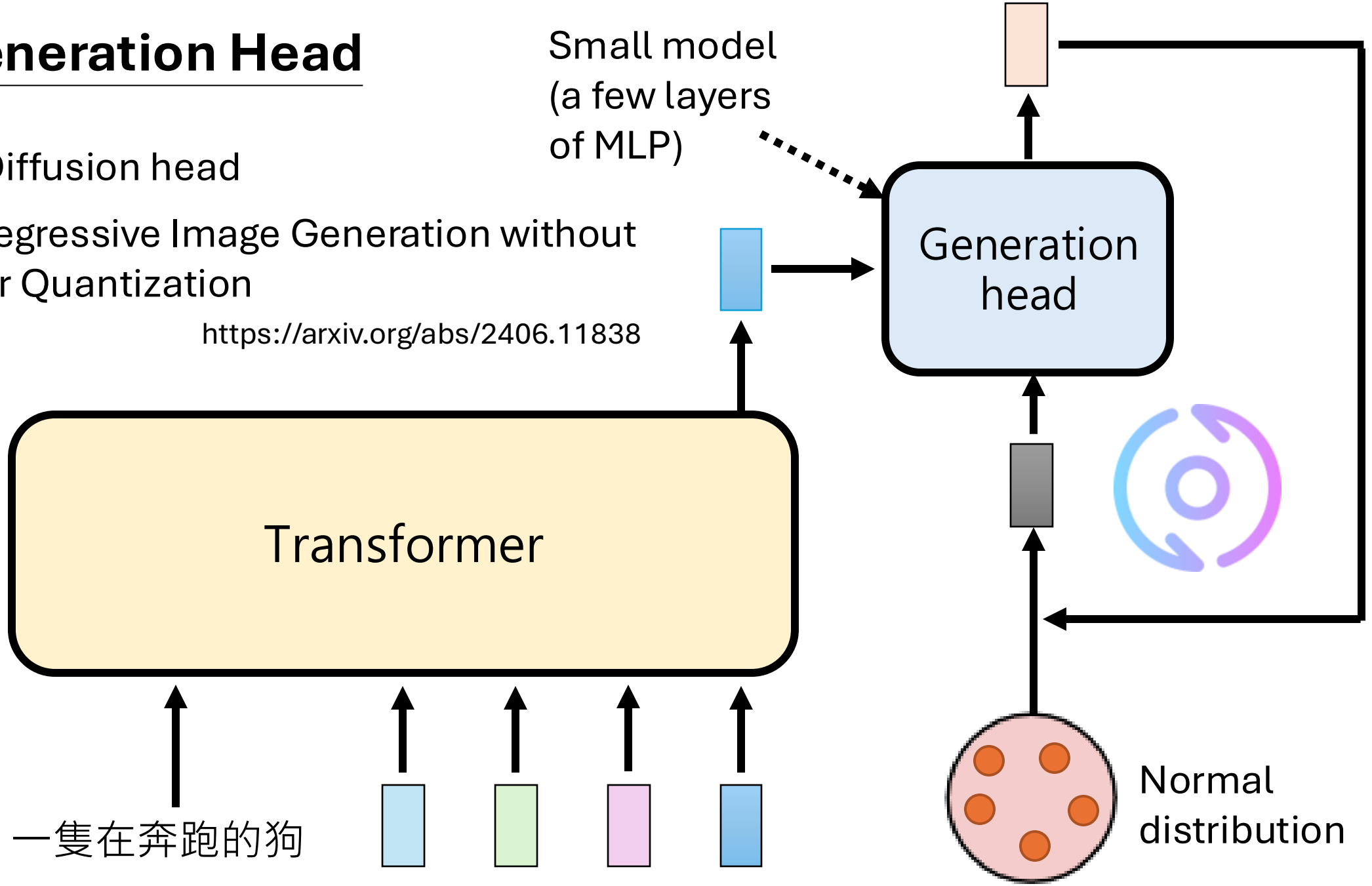


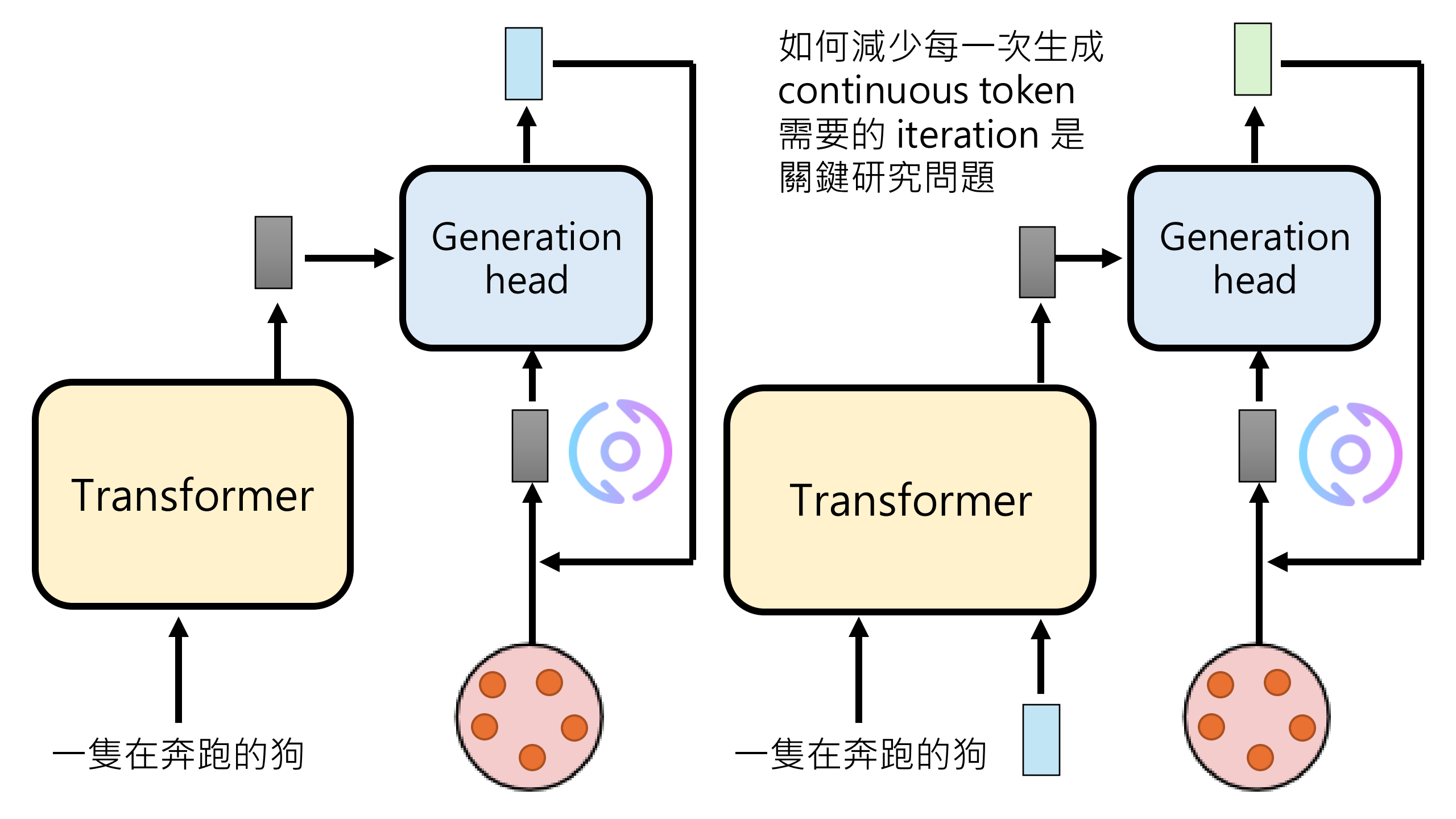
Generation Head

e.g., Diffusion head

Autoregressive Image Generation without Vector Quantization

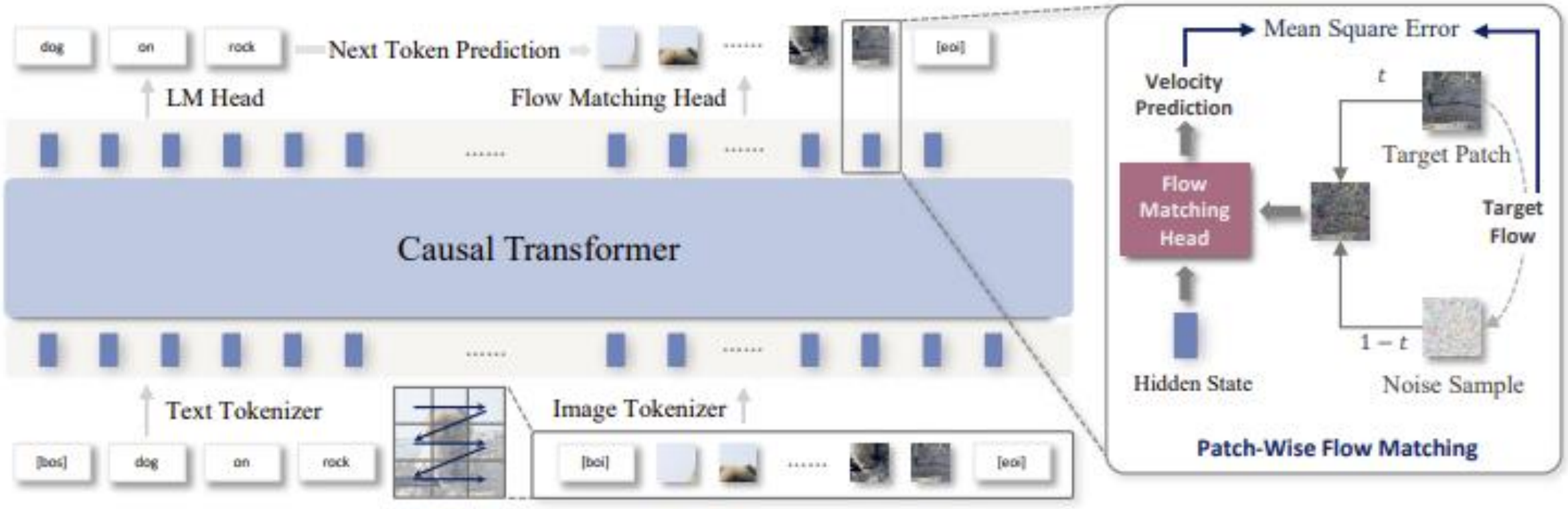
https://arxiv.org/abs/2406.11838




Generation head
Transformer
一隻在奔跑的狗
如何減少每一次生成 continuous token 需要的 iteration 是關鍵研究問題
Generation head
Transformer
一隻在奔跑的狗

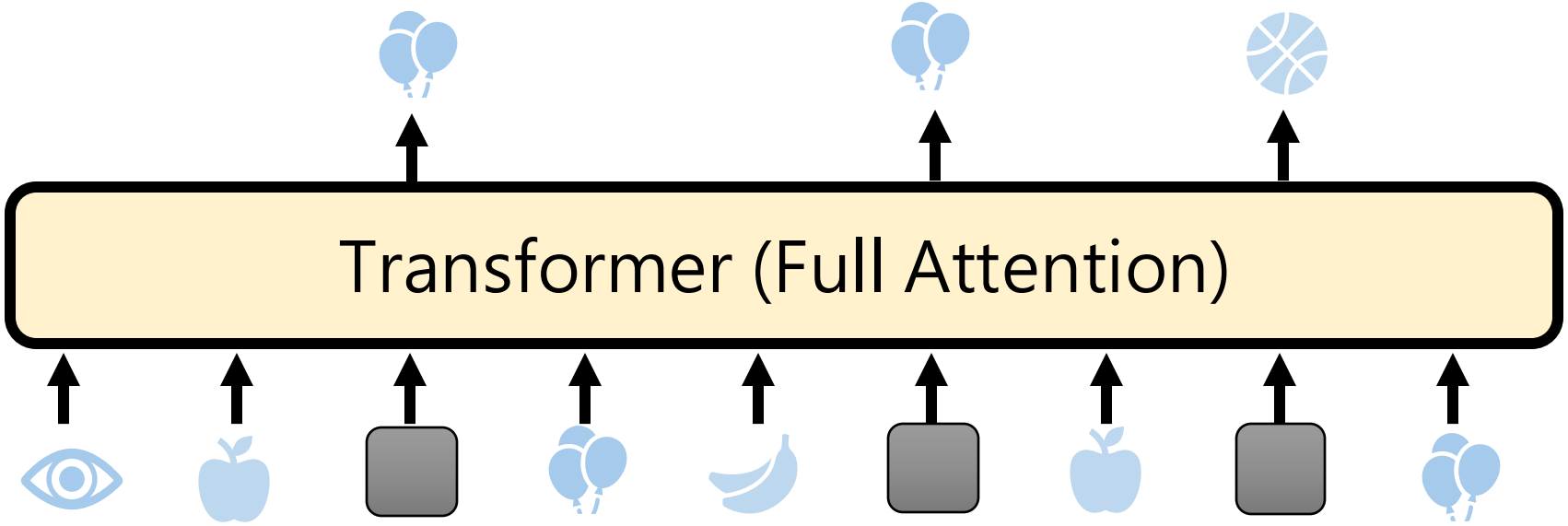
Flow-matching head

- NextStep-1 https://arxiv.org/abs/2508.10711



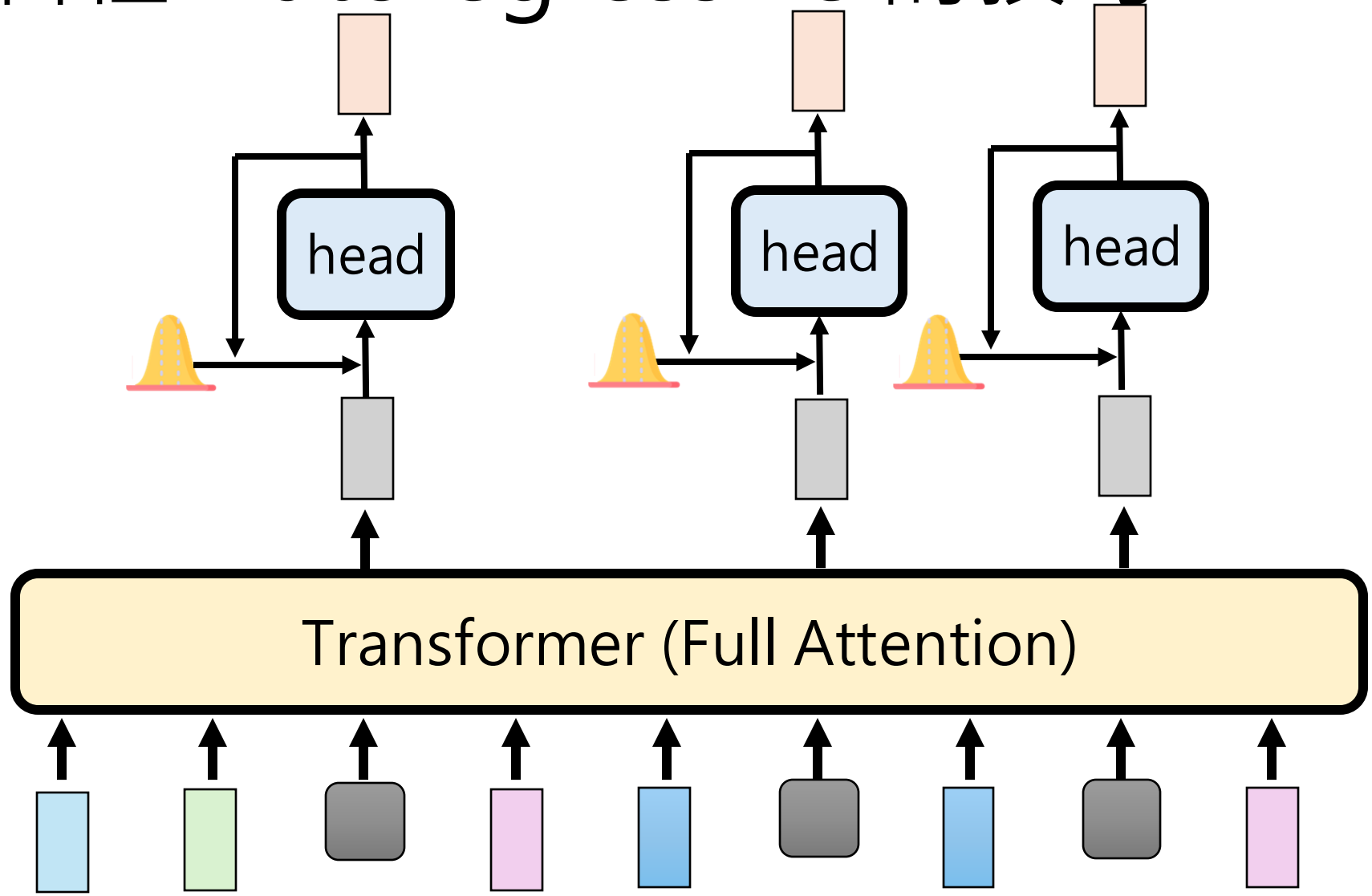
可以搭配各種 Autoregressive 的技巧

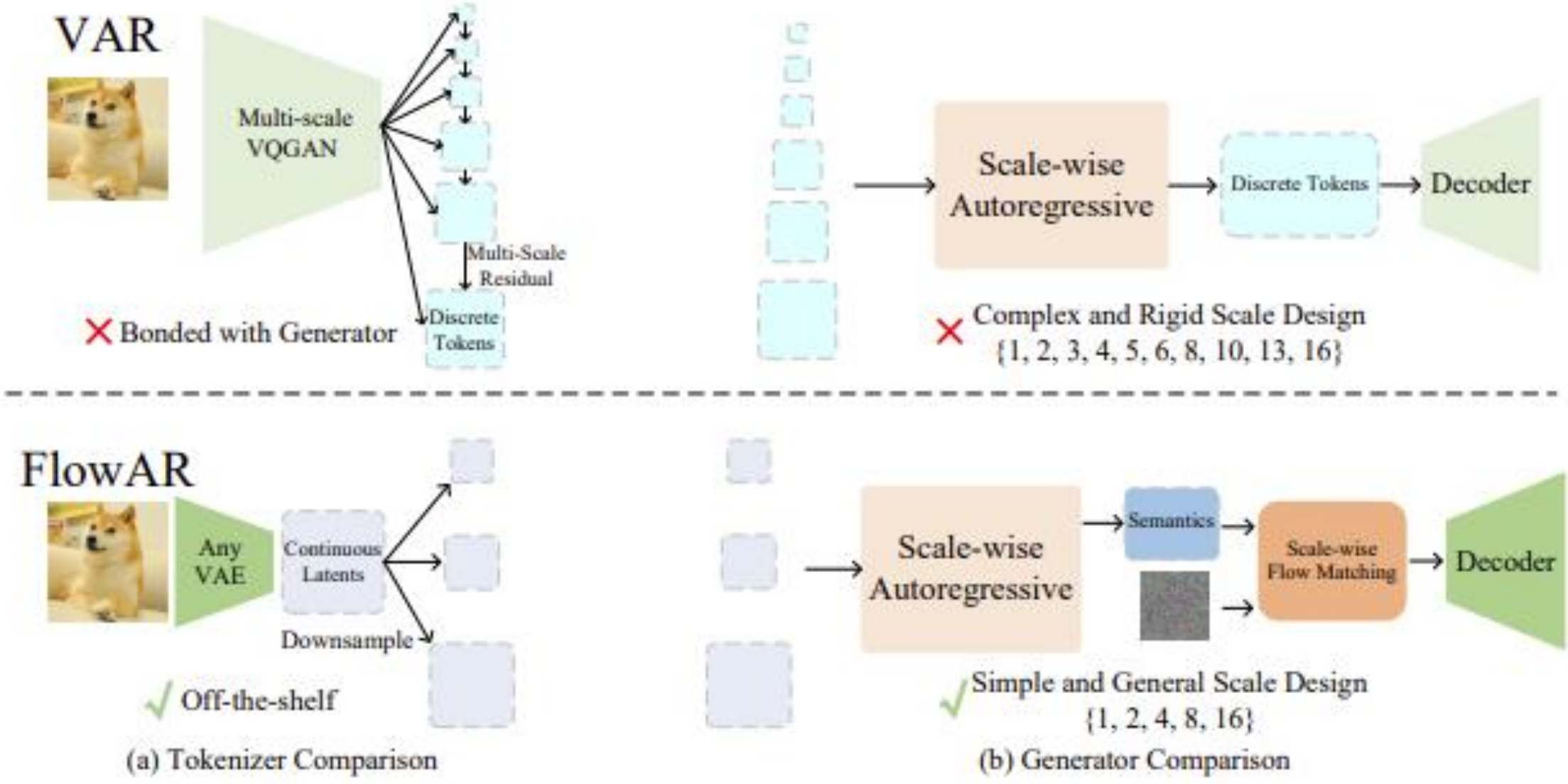
- MaskGIT



可以搭配各種 Autoregressive 的技巧

- MaskGIT





https://arxiv.org/abs/2412.15205

More Application of Generation Head

- Text-to-audio
  - IMPACT: Iterative Mask-based Parallel Decoding for Text-to-Audio Generation with Diffusion Modeling https://arxiv.org/abs/2506.00736
  - Generative Audio Language Modeling with Continuous-valued Tokens and Masked Next-Token Prediction https://arxiv.org/abs/2507.09834
- Text-to-speech
  - Efficient Speech Language Modeling via Energy Distance in Continuous Latent Space https://arxiv.org/abs/2505.13181
- Speech-to-speech LM
  - Flow-SLM: Joint Learning of Linguistic and Acoustic Information for Spoken Language Modeling https://arxiv.org/abs/2508.09350
- Video Generation
  - VideoMAR: Autoregressive Video Generation with Continuous Tokens https://arxiv.org/pdf/2506.14168v1

Concluding Remarks