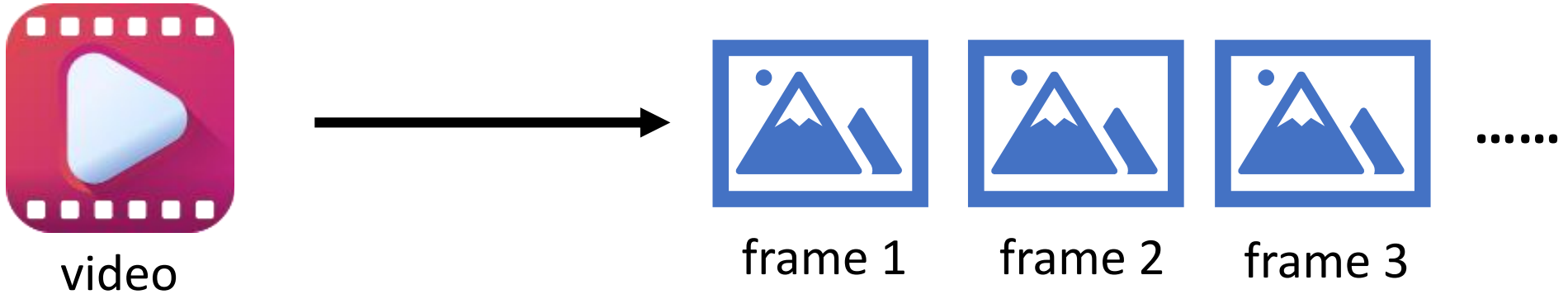
與影像有關的生成式 AI

【生成式AI導論 2024】第15講：為什麼語言模型用文字接龍，圖片生成不用像素接龍呢？– 淺談生成式人工智慧的生成策略

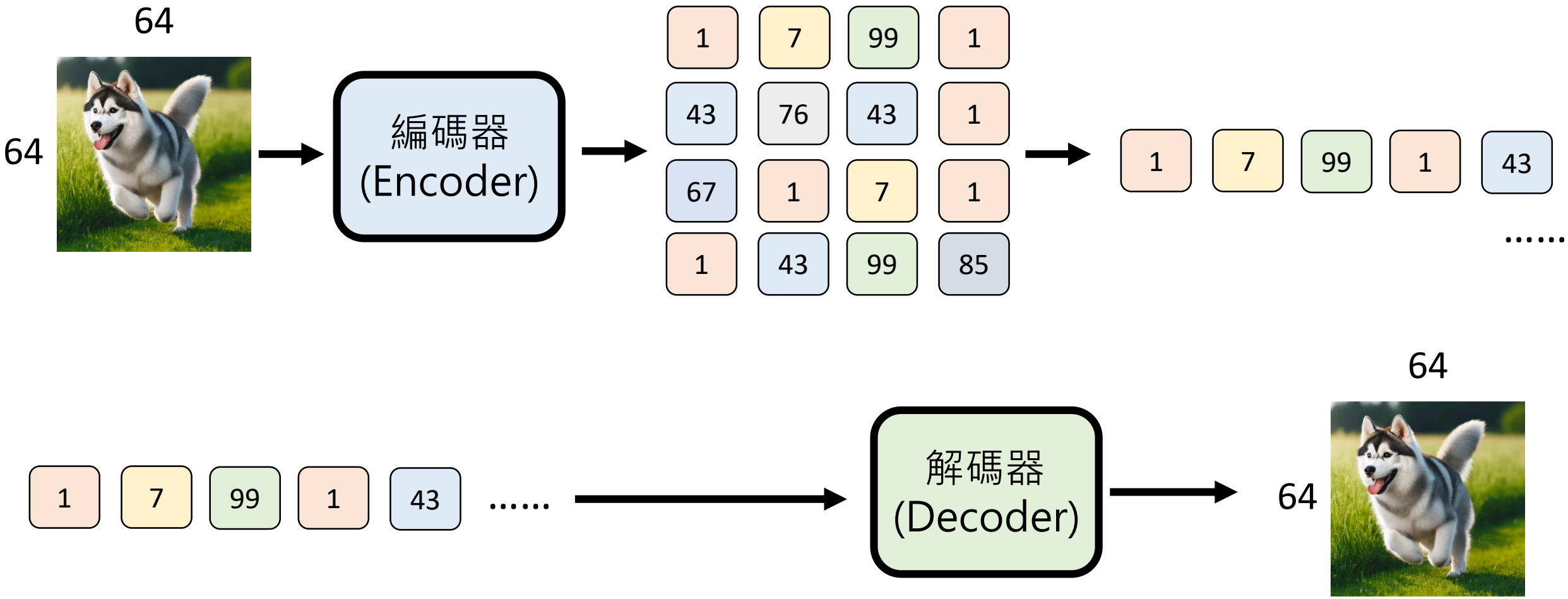
https://youtu.be/QbwQR9sjWbs?si=4svPnXaoD1rpcfgv

圖片是由像素所構成

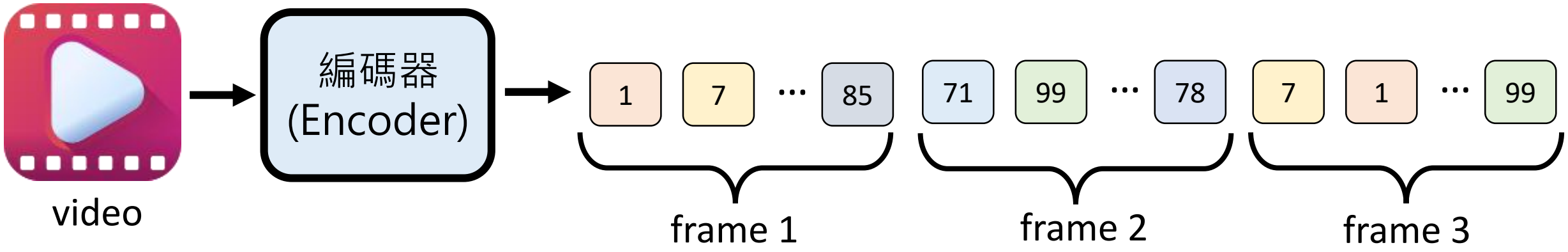
- 影片是由一張一張圖片所構成

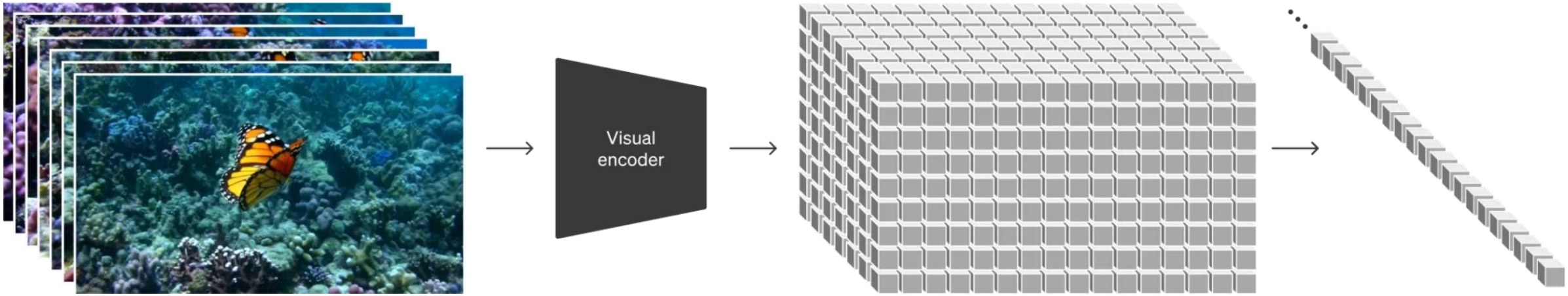


今日的人工智慧如何看影像 – 圖片

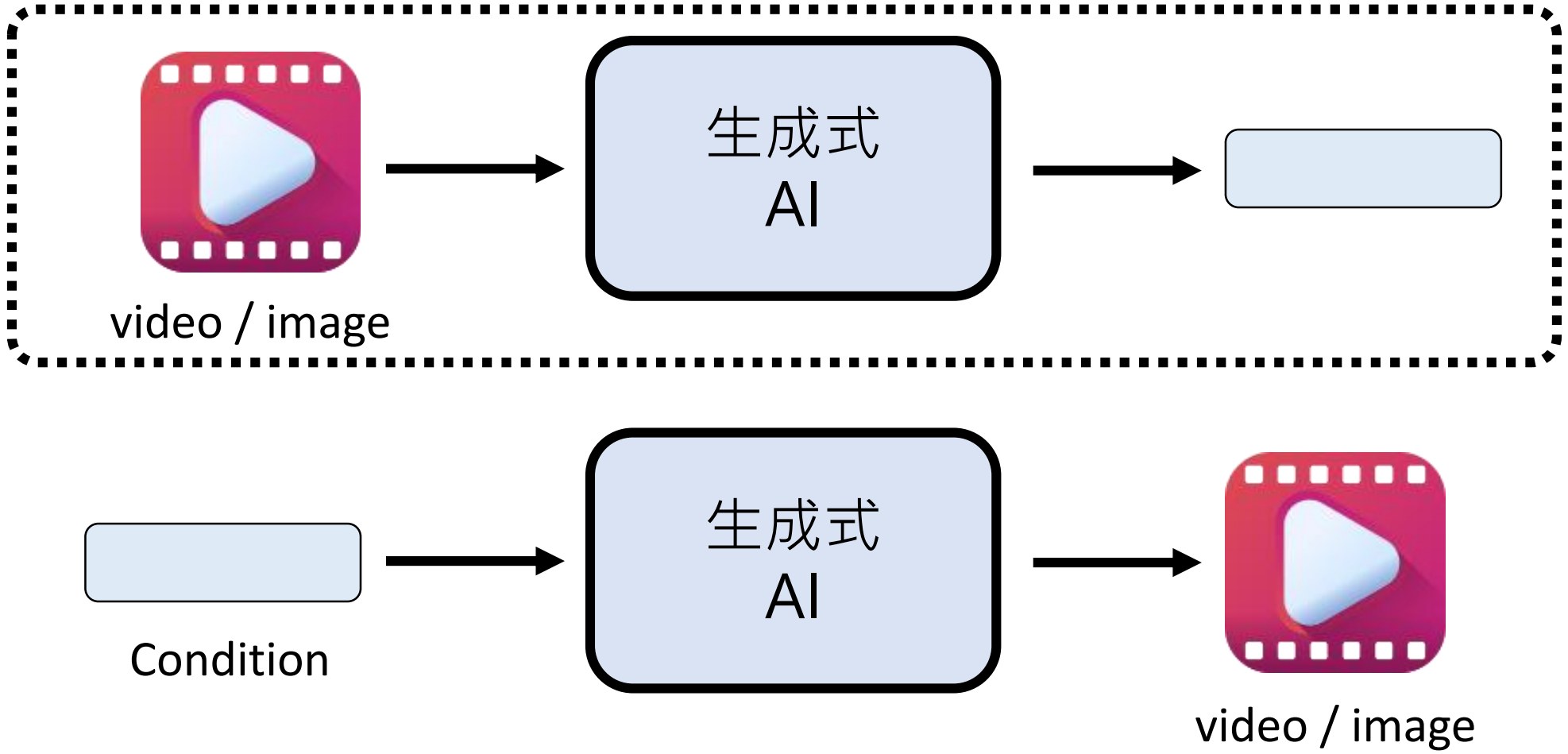


今日的人工智慧如何看影像 – 影片





與影像有關的生成式AI



這兩個人都叫李宏毅

GPT-4o

你覺得哪一張圖片中的人比較帥

猜猜看他們分別從事甚麼職業

猜測左邊那位穿著牛仔外套的年輕人可能從事時尚、娛樂或創意領域的工作，例如模特、演員或藝術家。右邊那位穿著白襯衫戴眼鏡的男士看起來更像是專業人士，可能從事學術、技術或商業領域的工作，例如教授、工程師或經理。你能告訴我他們實際從事的職業嗎？

這兩個人都叫李宏毅

GPT-4o

你覺得哪一張圖片中的人比較帥

圖片中的兩位風格不同，各有特色。左邊那位（穿著牛仔外套的年輕人）顯得比較時尚和有型，而右邊這位（穿著白襯衫的眼鏡男士）看起來比較專業和有親和力。具體誰比較帥，這很主觀，取決於個人喜好。你覺得呢？

你一定要決定誰比較帥

(讓模型看影像有很多方法，此處只說明了某一種可能性)

以影像作為 condition

影像
Patch

7 1 … 99

某種特殊的語言

Vocabulary

文字
Token

大
⋮
a
the
⋮

+

影像
Patch

1
2
3
⋮

看起來怎麼樣？ → 語言模型 → 照片中的人 .....

延伸閱讀

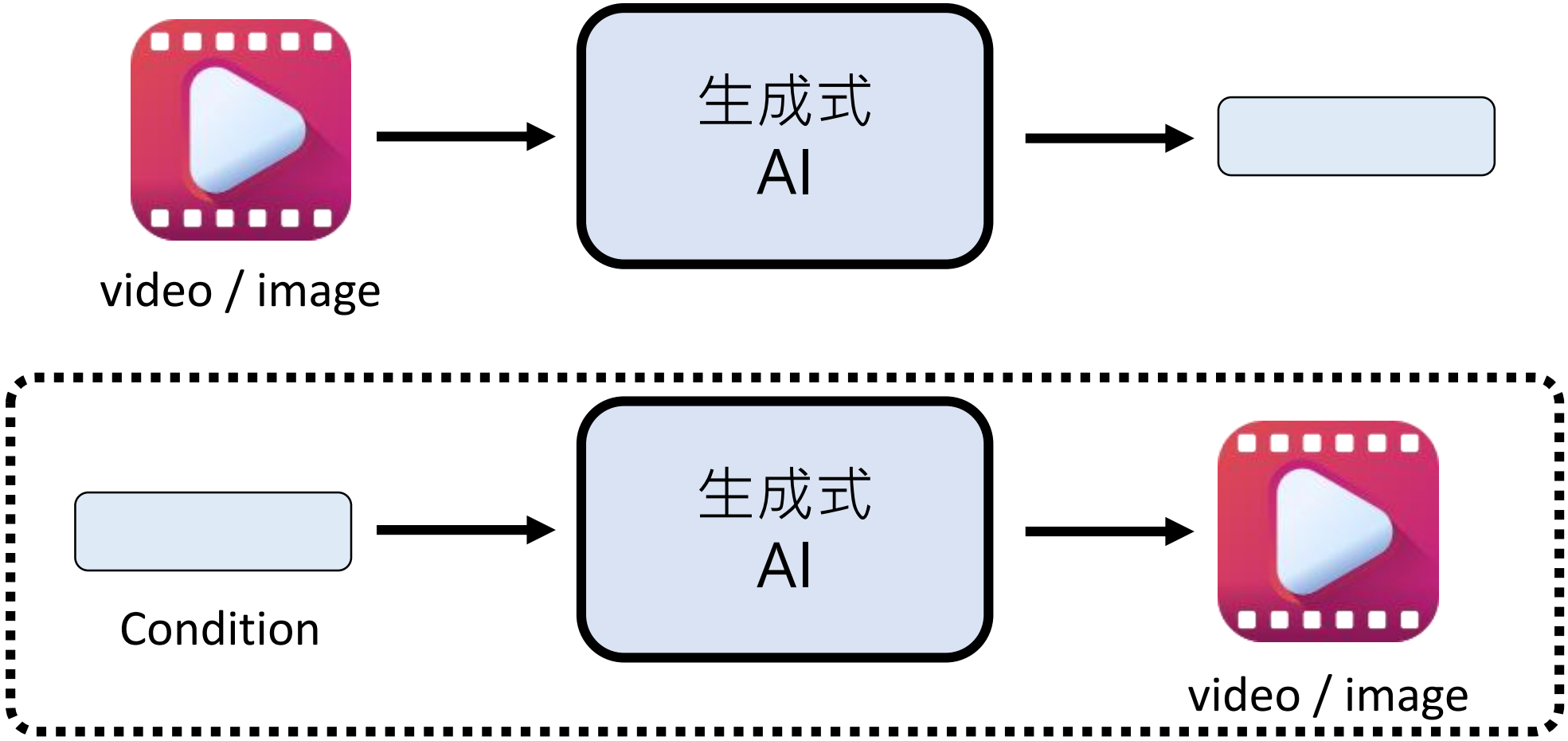
An Introduction to Vision-Language Modeling

Florian Bordes*, Richard Yuanzhe Pang*^, Anurag Ajay**♣, Alexander C. Li**♣, Adrien Bardes*, Suzanne Petryk△, Oscar Mañas*‡, Zhiqiu Lin♣, Anas Mahmoud†, Bargav Jayaraman*, Mark Ibrahim*, Melissa Hall*, Yunyang Xiong*, Jonathan Lebensold*♡, Candace Ross*, Srihari Jayakumar*, Chuan Guo*, Diane Bouchacourt*, Haider Al-Tahan*, Karthik Padthe*, Vasu Sharma*, Hu Xu*, Xiaoqing Ellen Tan*, Megan Richards*, Samuel Lavoie*‡, Pietro Astolfi*, Reyhane Askari Hemmat*, Jun Chen**◇, Kushal Tirumala*, Rim Assouel*‡, Mazda Moayeri▽, Arjang Talattof*, Kamalika Chaudhuri*, Zechun Liu*, Xilun Chen*, Quentin Garrido*, Karen Ullrich*, Aishwarya Agrawal†•, Kate Saenko*, Asli Celikyilmaz* and Vikas Chandra*

*Meta
**Work done while at Meta
‡Université de Montréal, Mila
♡McGill University, Mila
†University of Toronto
♣Carnegie Mellon University
♣Massachusetts Institute of Technology
^New York University
△University of California, Berkeley
▽University of Maryland
◇King Abdullah University of Science and Technology
•Canada CIFAR AI Chair
Core contributors, random ordering
Additional contributors, random ordering
Senior contributors, random ordering

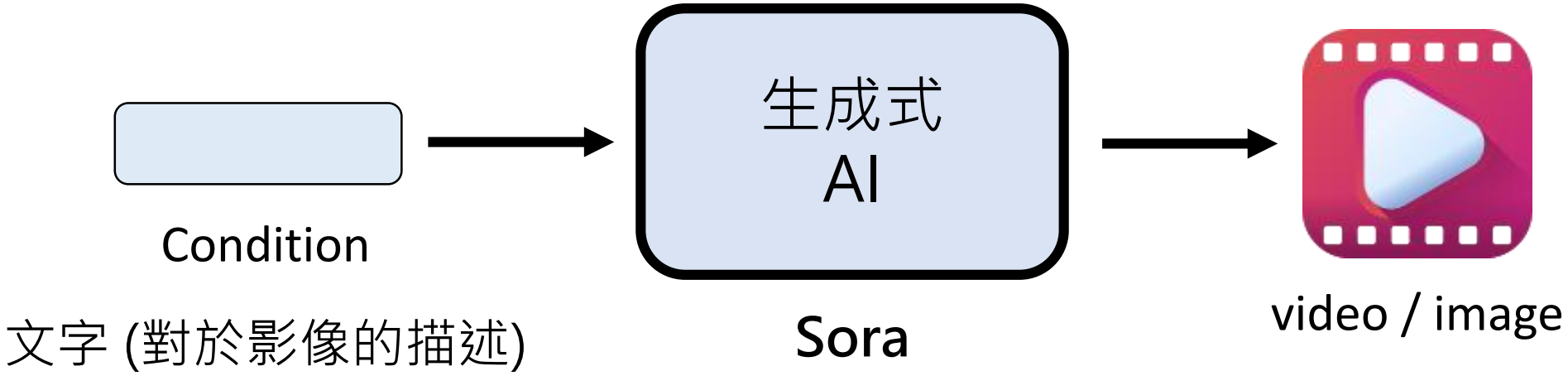
https://arxiv.org/abs/2405.17247

與影像有關的生成式AI



生成影像的生成式 AI — 文字生影像

- 以文字作為 condition



Sora 並沒有真的開放使用

https://openai.com/sora

https://openai.com/research/video-generation-models-as-world-simulators

生成影像的生成式 AI — 文字生影像 **Sora**

Animated scene features a close-up of a short fluffy monster kneeling beside a melting red candle. The art style is 3D and realistic, with a focus on lighting and texture. …..

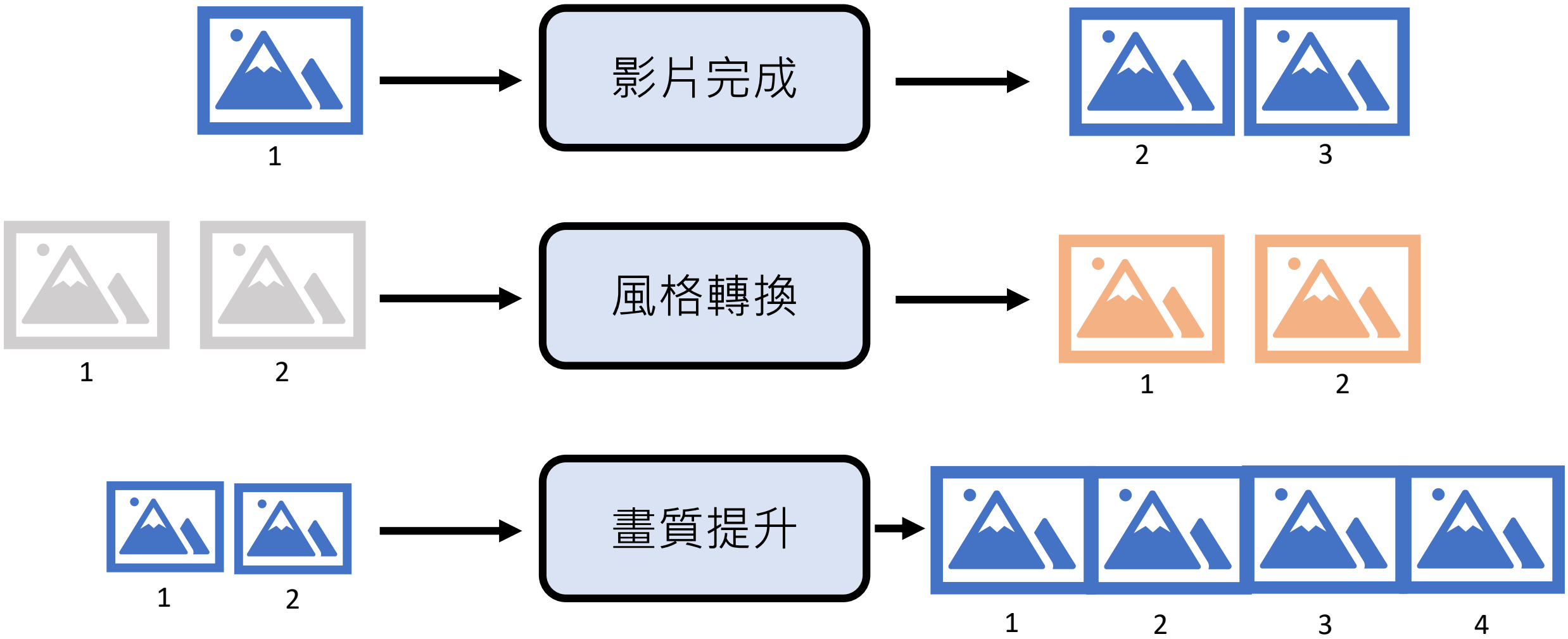
New York City submerged like Atlantis. Fish, whales, sea turtles and sharks swim through the streets of New York.

生成影像的生成式 AI — 文字生影像 **Sora**

Five gray wolf pups frolicking and chasing each other around a remote gravel road, surrounded by grass. The pups run and leap, chasing each other, and nipping at each other, playing.

Archeologists discover a generic plastic chair in the desert, excavating and dusting it with great care.

生成影像的生成式 AI — 影像生影像



生成影像的生成式 AI — 其他輸入生影像

- Talking Head

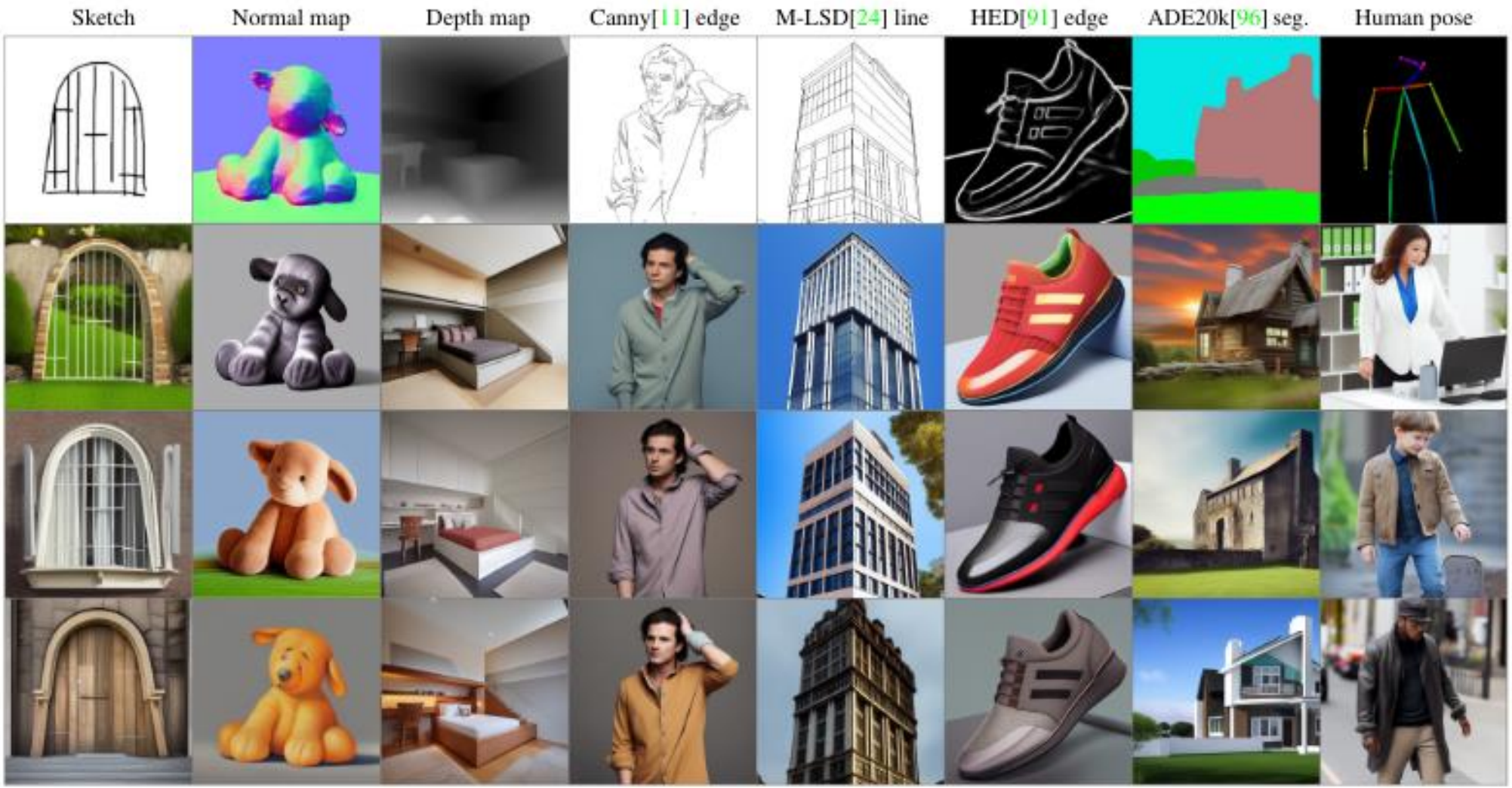
image

生成式
AI

Paper: https://arxiv.org/abs/2403.17694
Demo: https://huggingface.co/spaces/ZJYang/AniPortrait_official

ControlNet: https://arxiv.org/abs/2302.05543

生成影像的生成式 AI — 其他輸入生影像



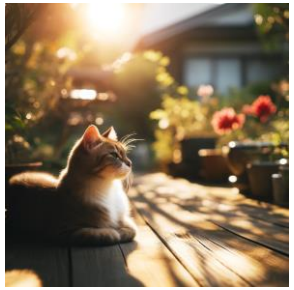
以文字生圖為例

一隻在奔跑的狗 → 圖片生成 → 

訓練資料

一隻在奔跑的狗

雪地裡的貓



陽光下的貓

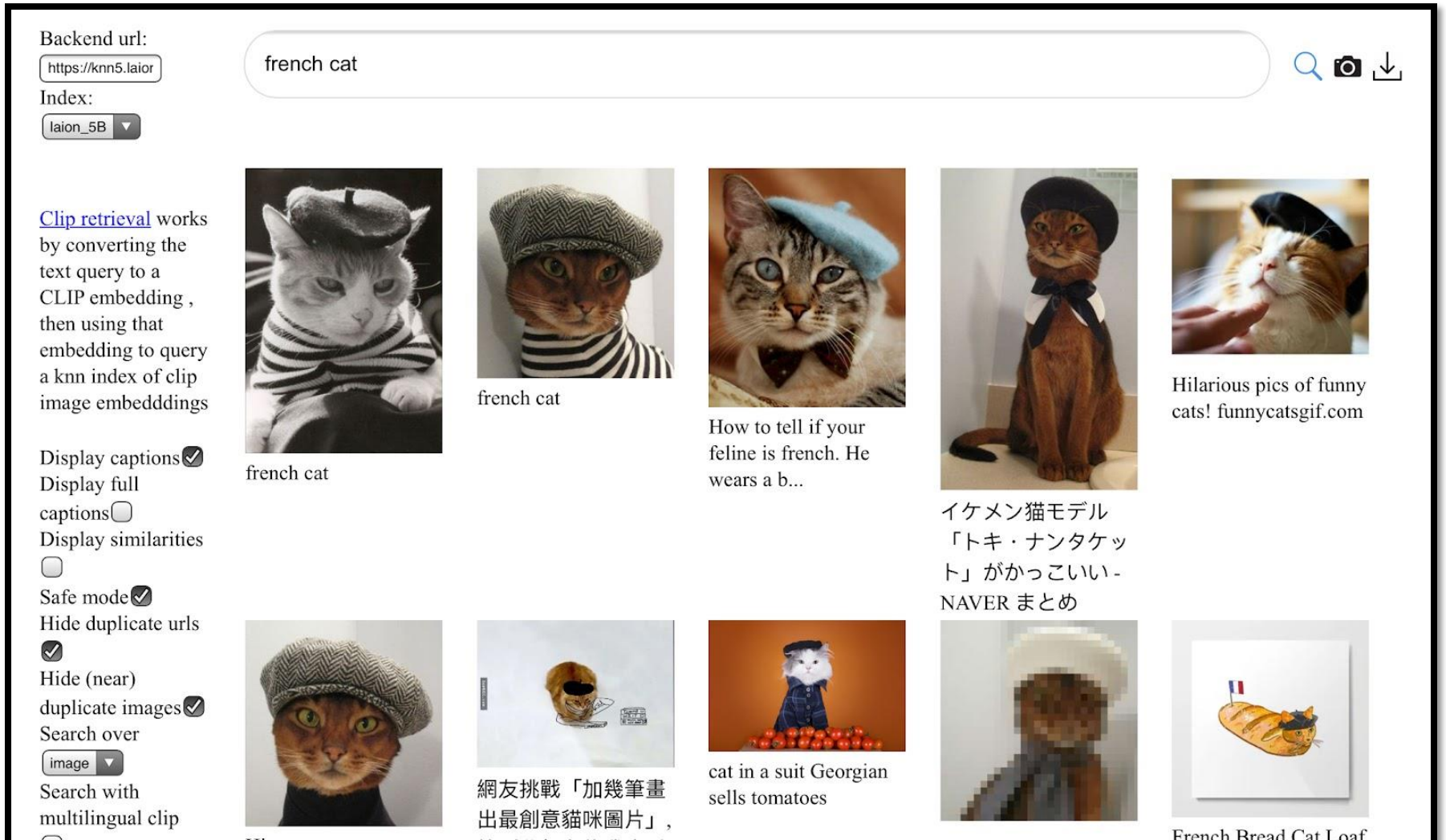
沙灘上的狗

以文字生圖為例

https://laion.ai/blog/laion-5b/

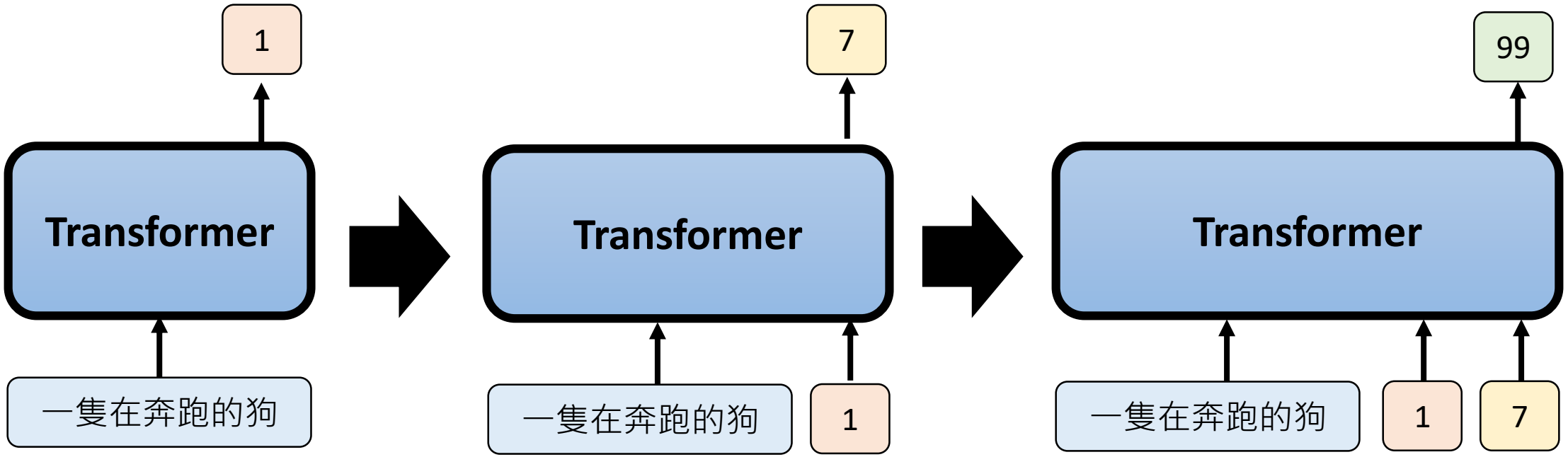
LAION

5.85B



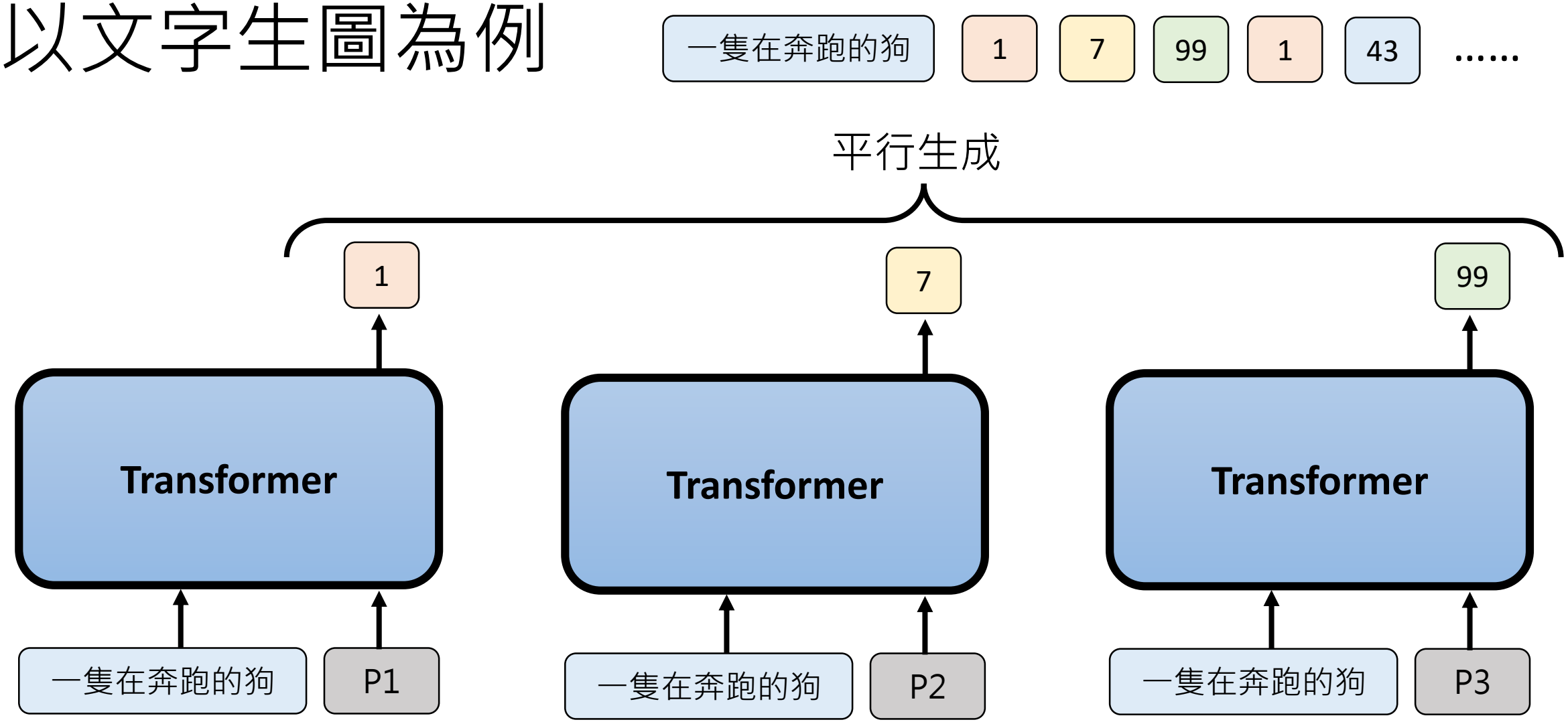
以文字生圖為例

一隻在奔跑的狗 | 1 | 7 | 99 | 1 | 43 | .....



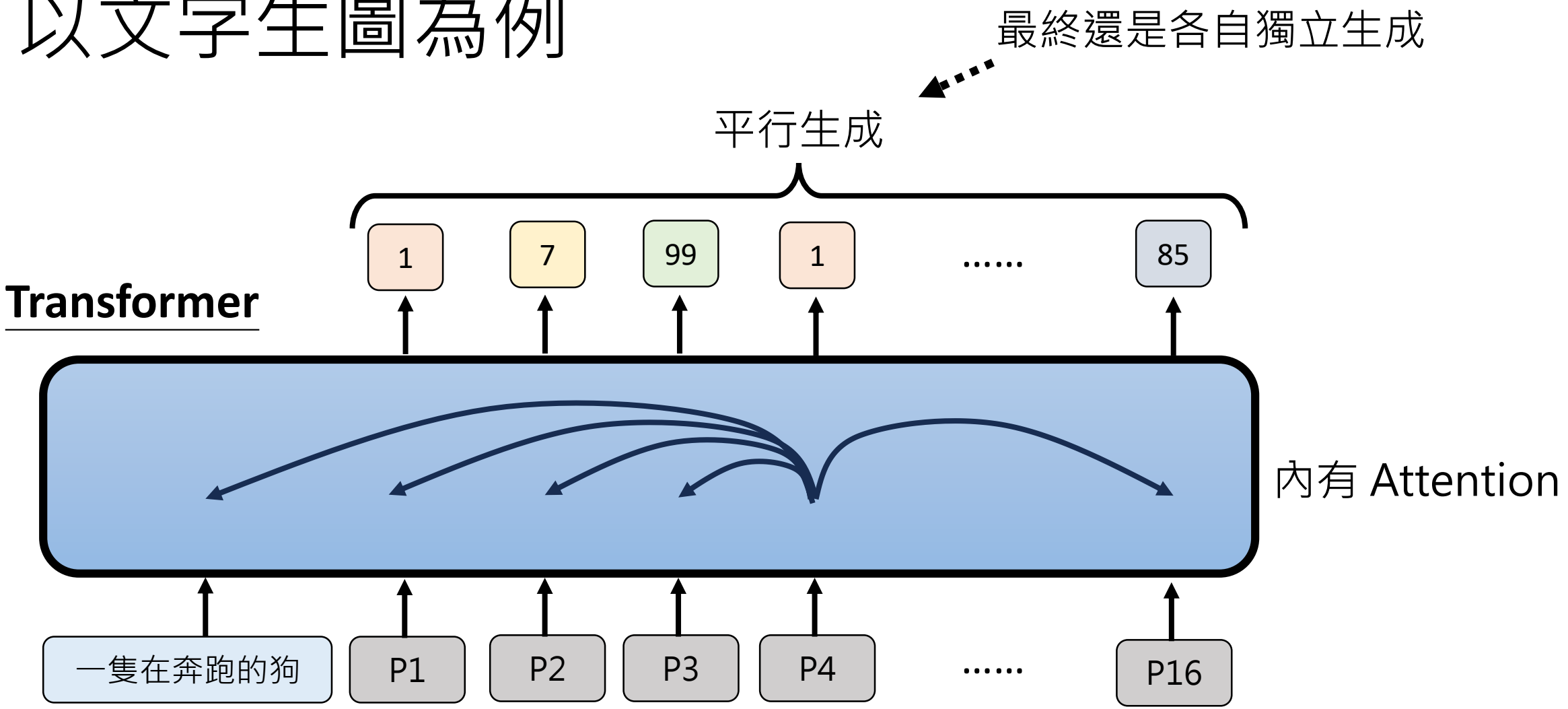
(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

以文字生圖為例



(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

以文字生圖為例



(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

如何評量影像生成的好壞

https://arxiv.org/abs/2103.00020

怎麼知道生成結果好不好？

如何評量影像生成的好壞

https://arxiv.org/abs/2103.00020

沙灘上的狗

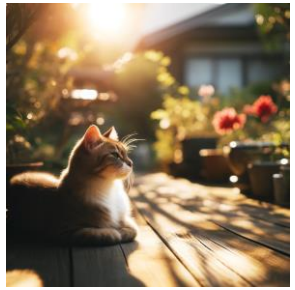
CLIP

高分

陽光下的貓

CLIP

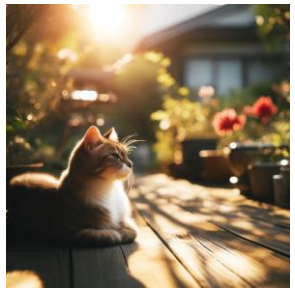
低分



陽光下的貓

CLIP

高分



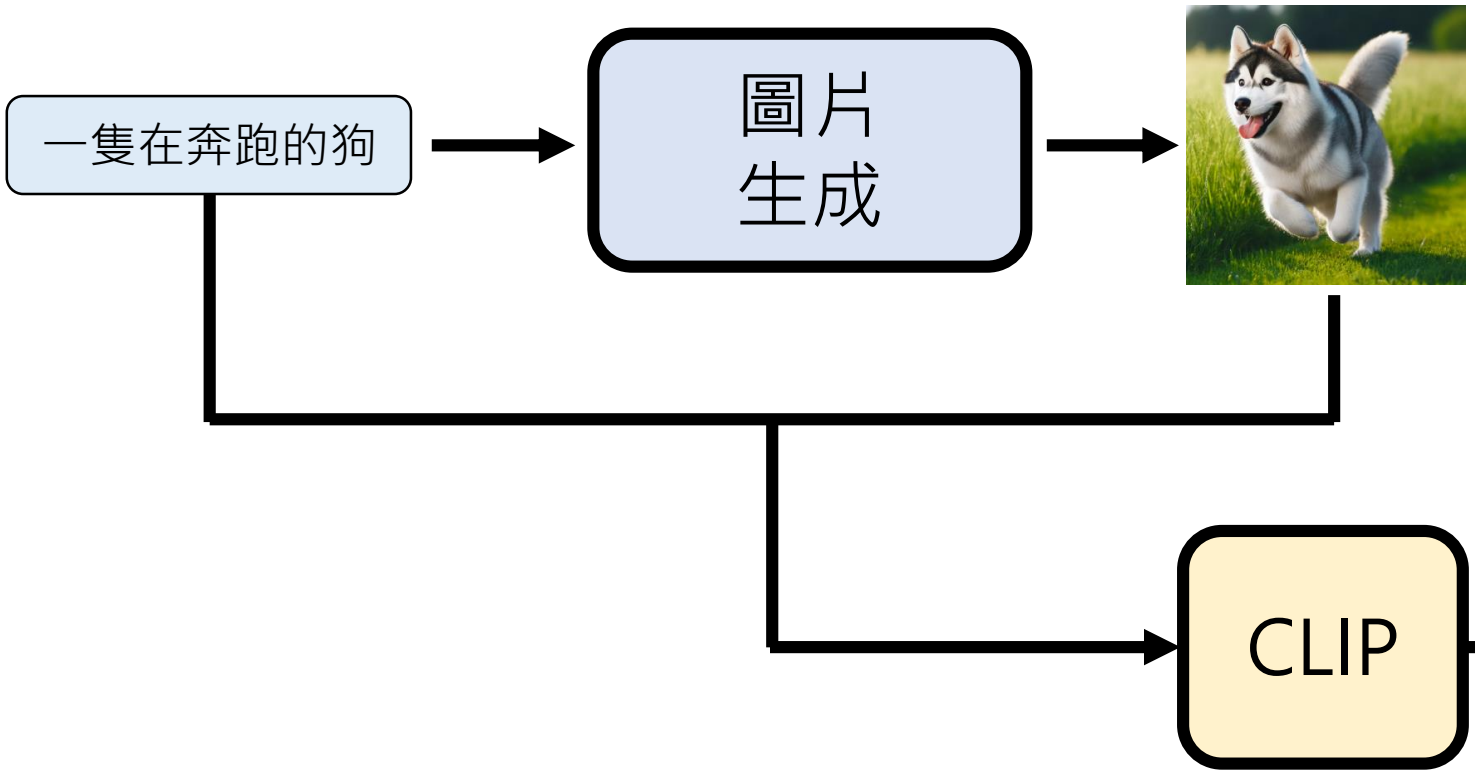
沙灘上的狗

CLIP

低分

如何評量影像生成的好壞

https://arxiv.org/abs/2103.00020



怎麼知道生成結果好不好？


PATTIHE

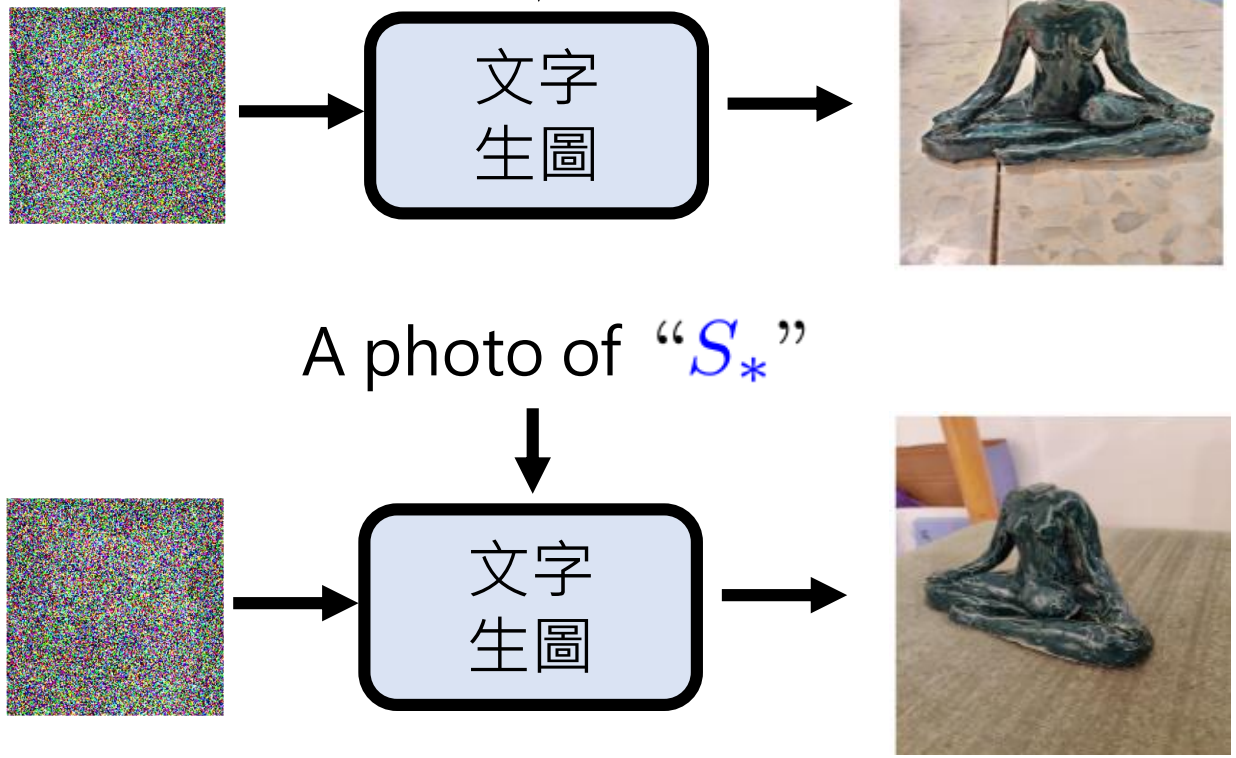

patya

個人化的圖像生成

https://arxiv.org/abs/2208.01618
https://arxiv.org/abs/2208.12242

“S_*”

(平常沒有在用的符號)



Source of image: https://arxiv.org/abs/2208.01618

Input samples $\xrightarrow{invert}$ "S_*"

"An oil painting of S_*"

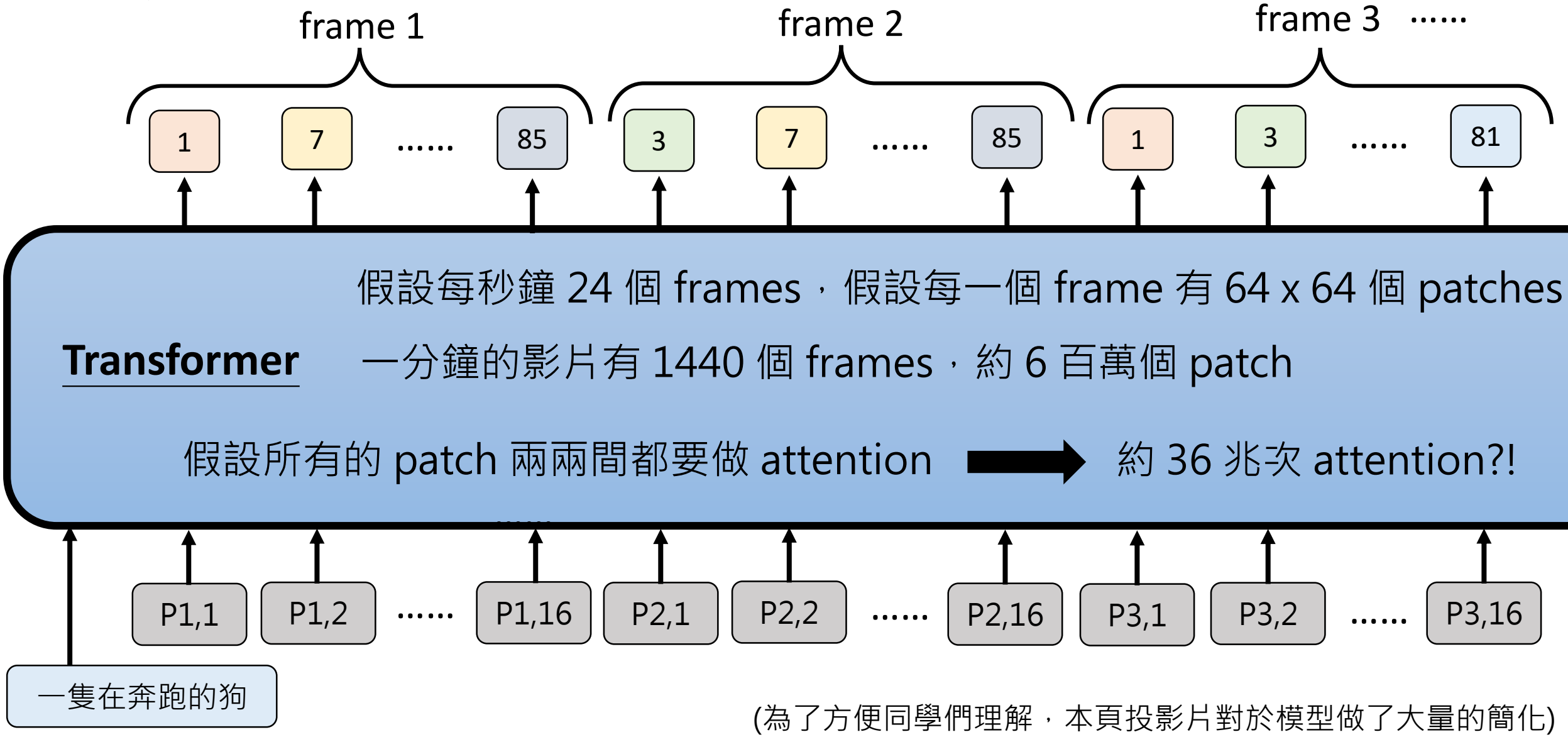
"App icon of S_*"

"Elmo sitting in the same pose as S_*"

"Crochet S_*"

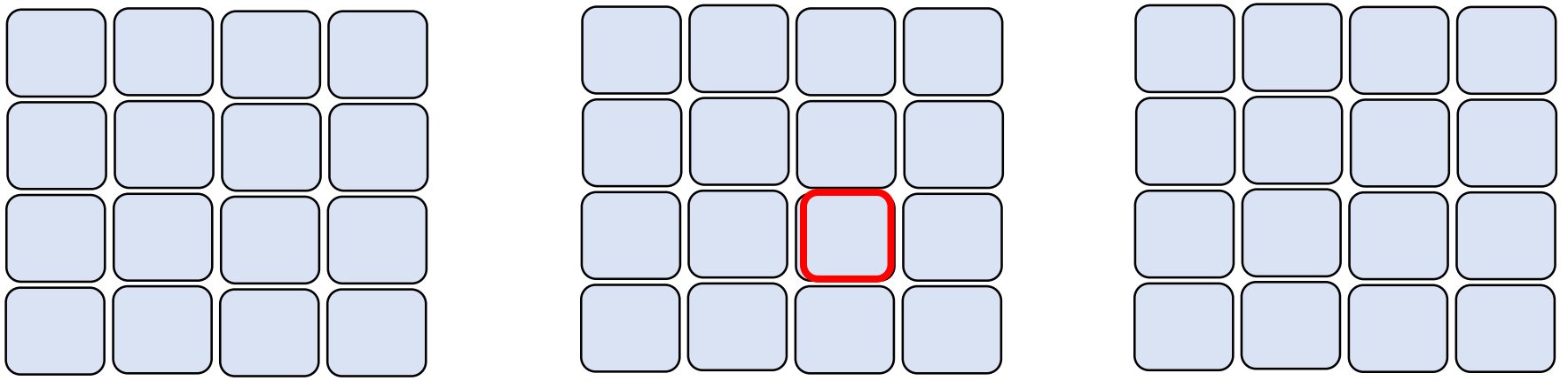
Source of image: https://arxiv.org/abs/2208.01618

文字生影片的挑戰

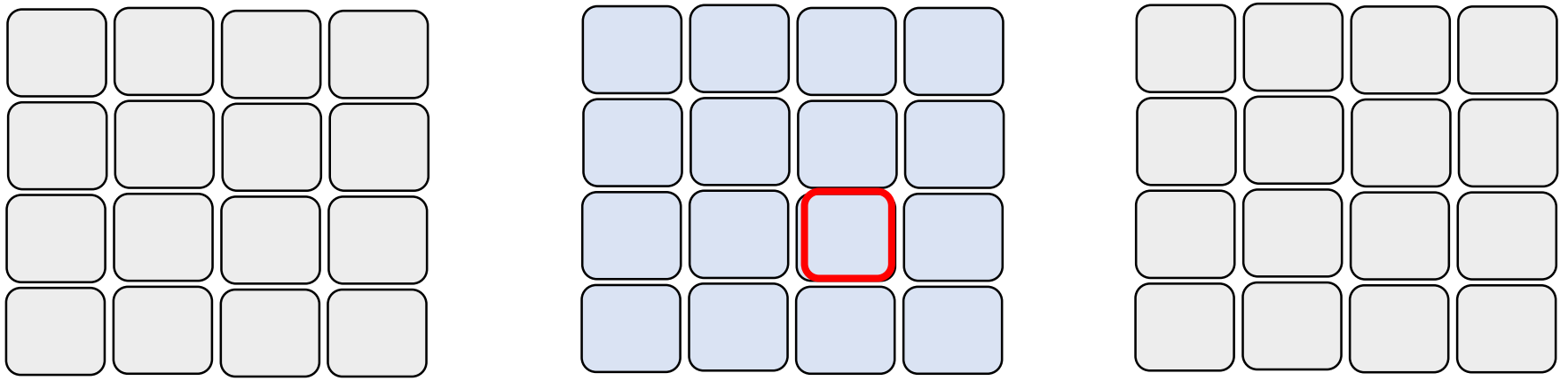


(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

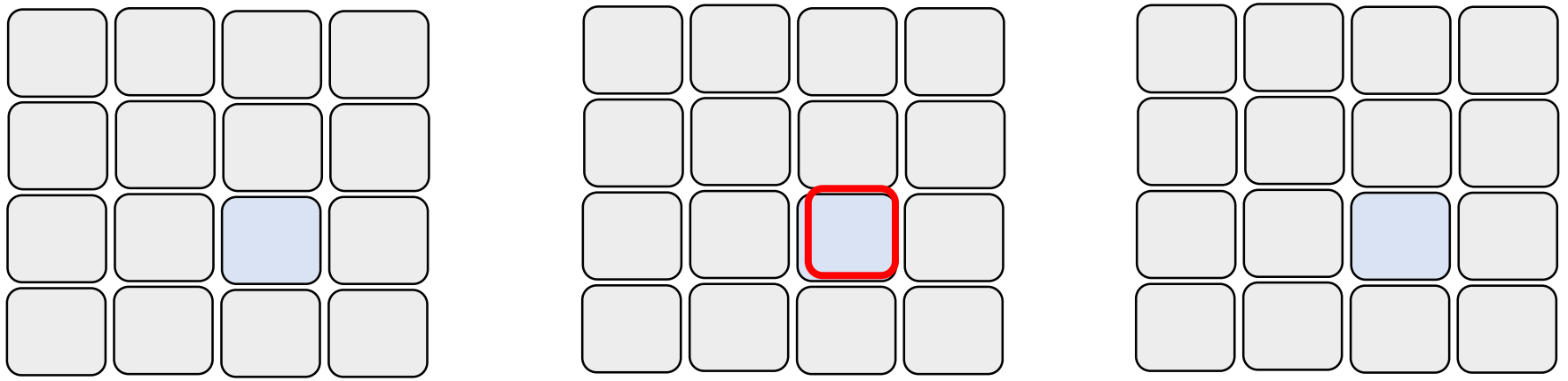
Spatio-temporal Attention (3D)

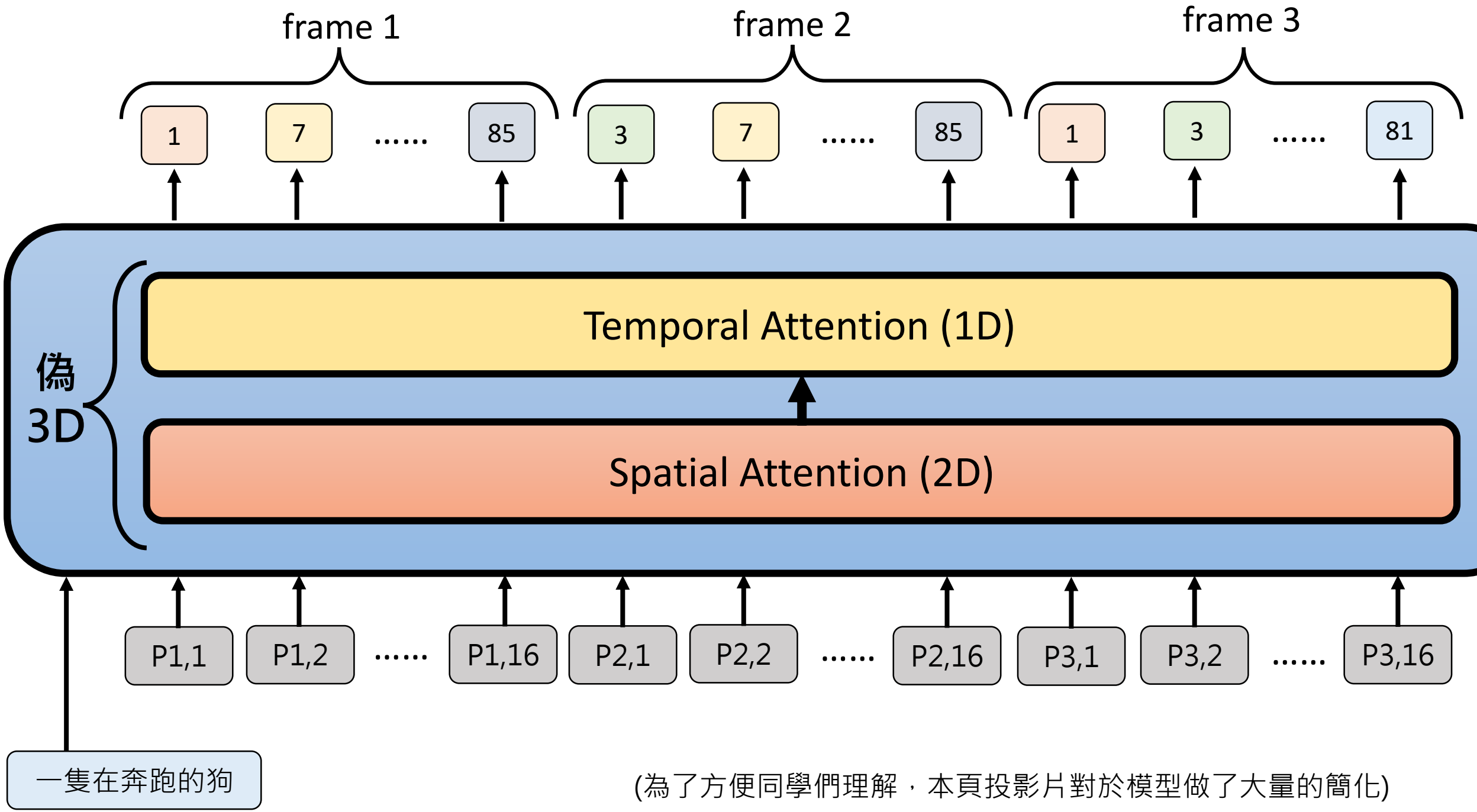


Spatial Attention (2D)



Temporal Attention (1D)





(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

假設每秒鐘 24 個 frames，假設每一個 frame 有 64 x 64 個 patches

一分鐘的影片有 1440 個 frames，約 6 百萬個 patch

Spatio-temporal Attention (3D)

假設所有的 patch 兩兩間都要做 attention

→ 約 36 兆次 attention?!

相差千倍

Spatial Attention (2D)

同一個 frame 中的 patch 才做 attention

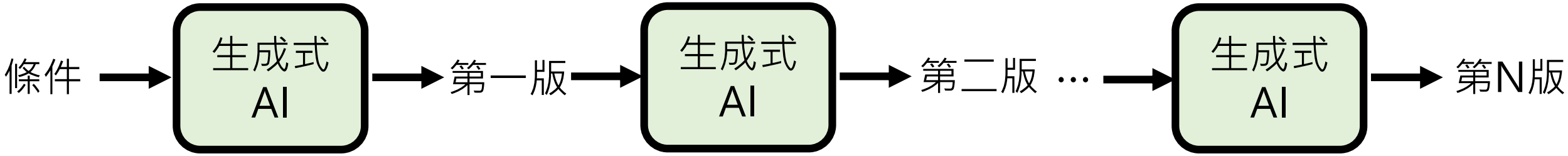
$(64 \times 64)^2 \times 1440$ → 約 240 億次 attention

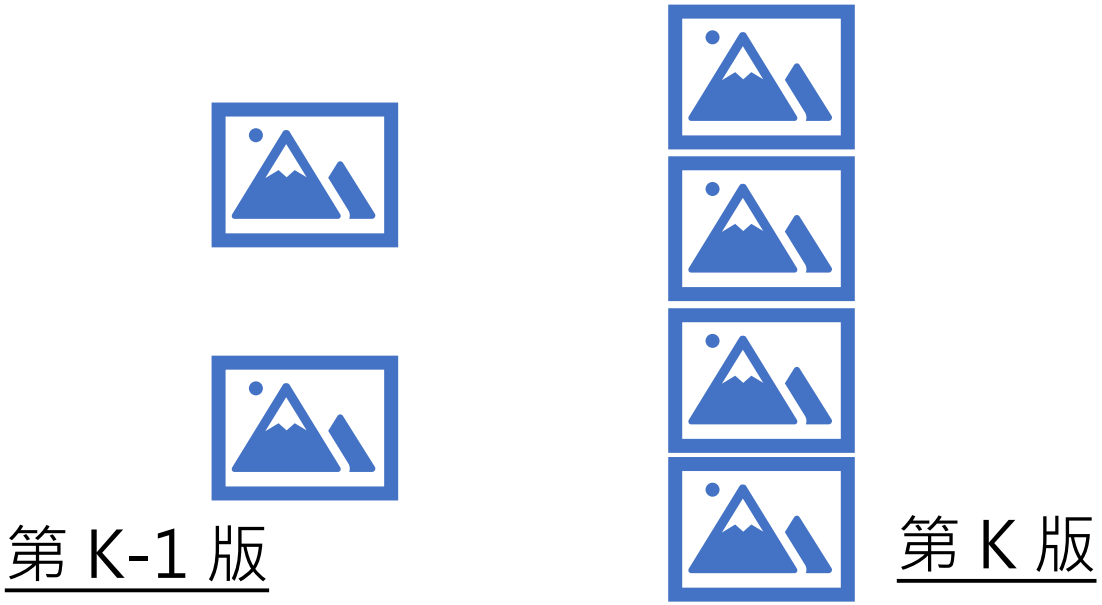
Temporal Attention (1D)

不同 frame 中同樣位置的 patch 才做 attention

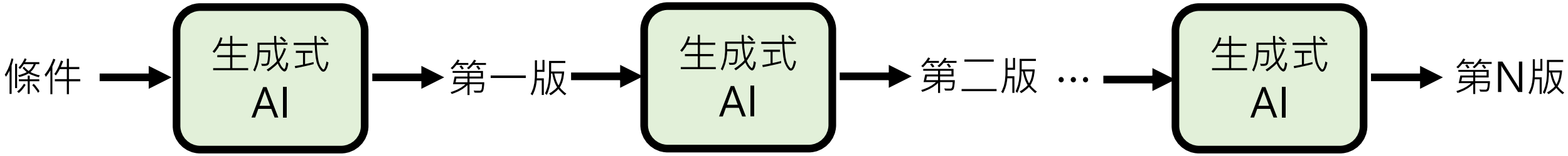
$1440^2 \times (64 \times 64)$ → 約 85 億次 attention

文字生影片的挑戰





文字生影片的挑戰



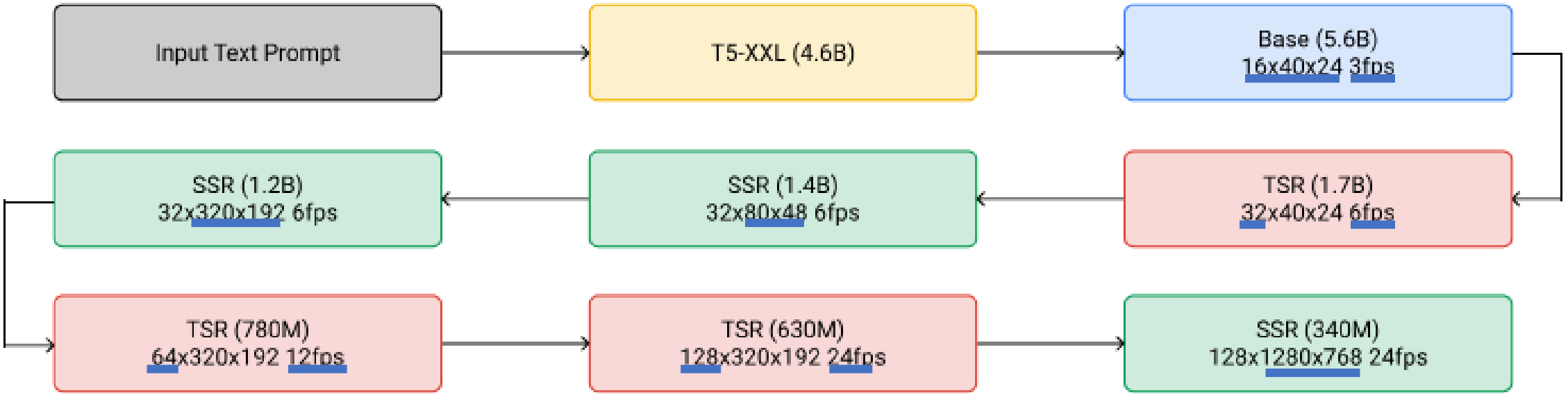


Imagen Video https://arxiv.org/abs/2210.02303

延伸閱讀

https://arxiv.org/abs/2405.03150

Video Diffusion Models: A Survey

Andrew Melnik *andrew.melnik.papers@gmail.com*
Bielefeld University

Michal Ljubljanac *mljubljanac@techfak.uni-bielefeld.de*
Bielefeld University

Cong Lu *conglu@cs.ubc.ca*
University of British Columbia

Qi Yan *qi.yan@ece.ubc.ca*
University of British Columbia

Weiming Ren *w2ren@uwaterloo.ca*
University of Waterloo

Helge Ritter *helge@techfak.uni-bielefeld.de*
Bielefeld University

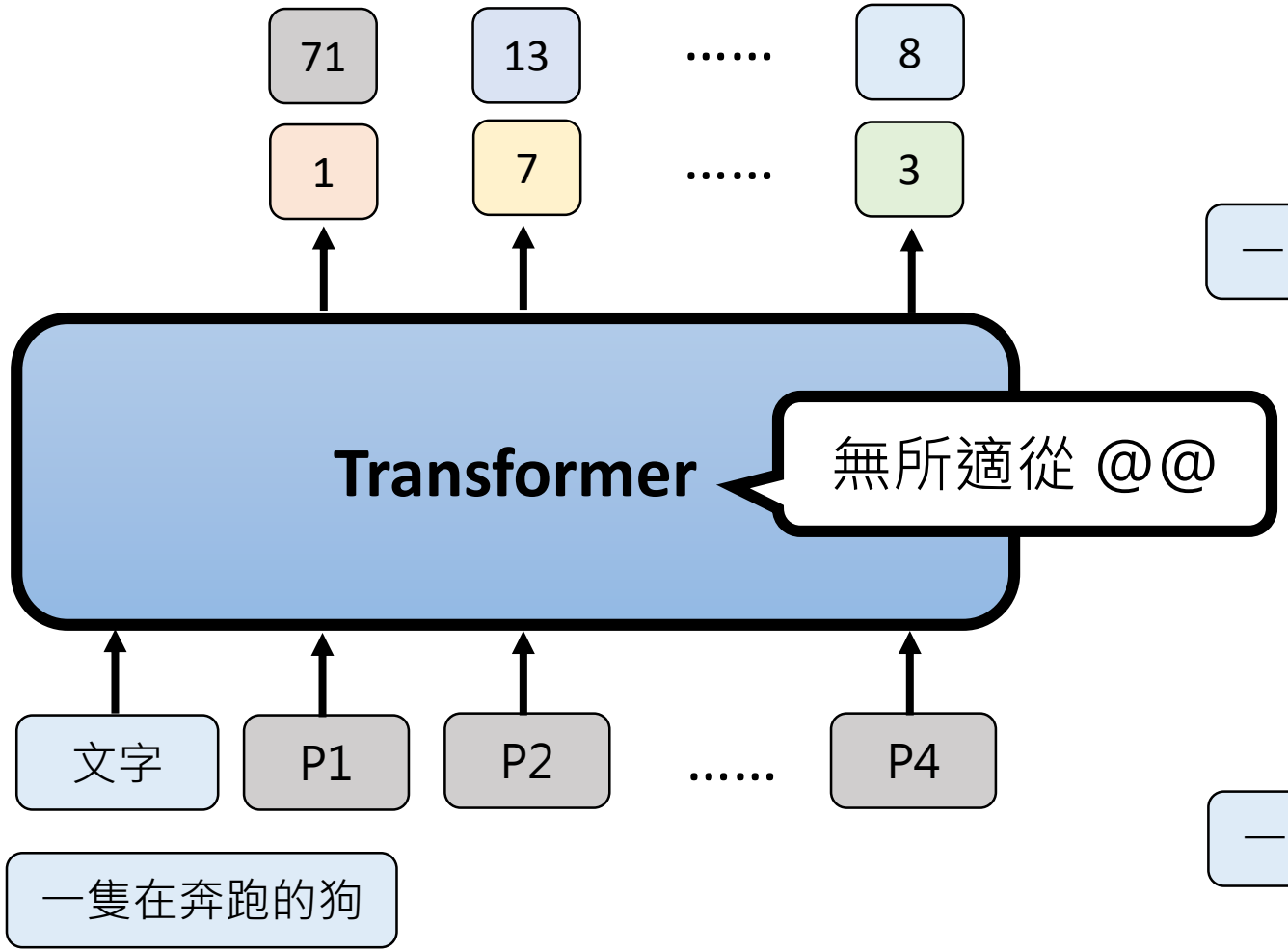
經典影像生成方法介紹

- Variational Auto-encoder (VAE)
- Flow-based Method
- Diffusion Method
- Generative Adversarial Network (GAN)

Sora 使用的是 Diffusion

Source of image: https://openai.com/index/video-generation-models-as-world-simulators/

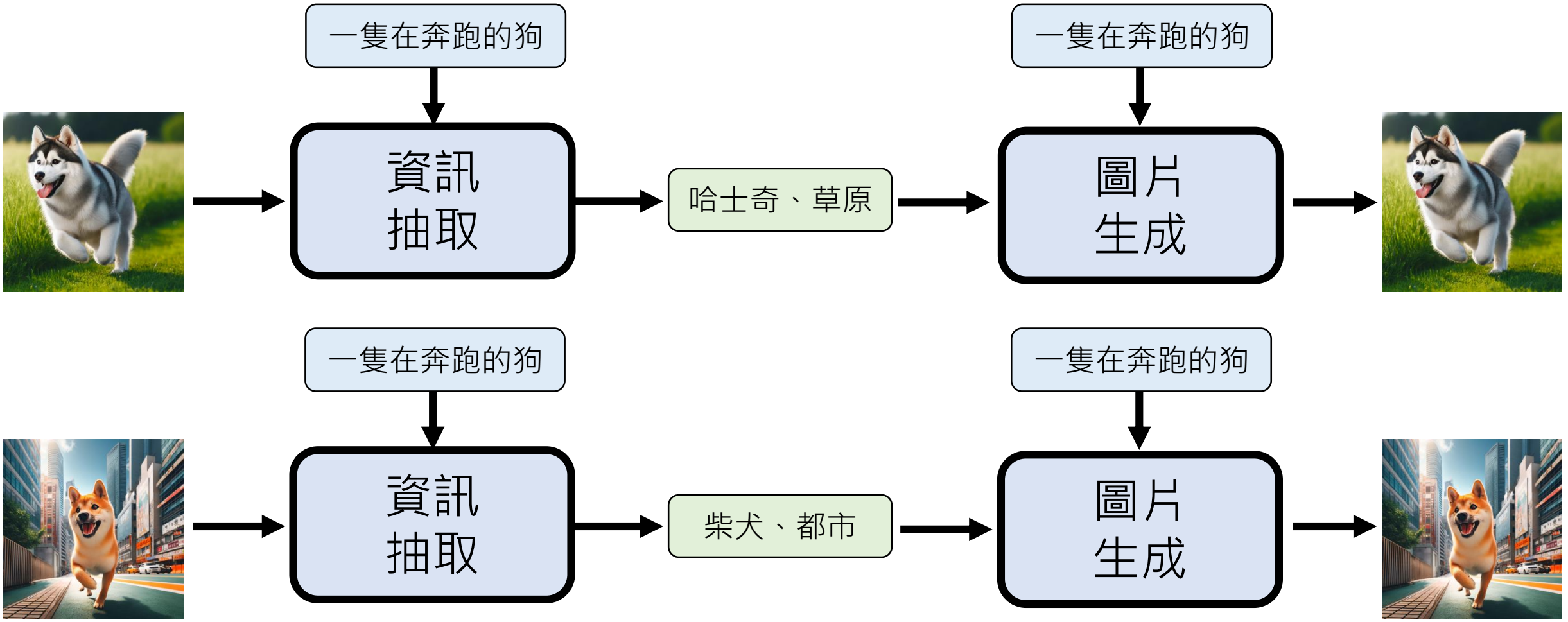
文字生影像的挑戰

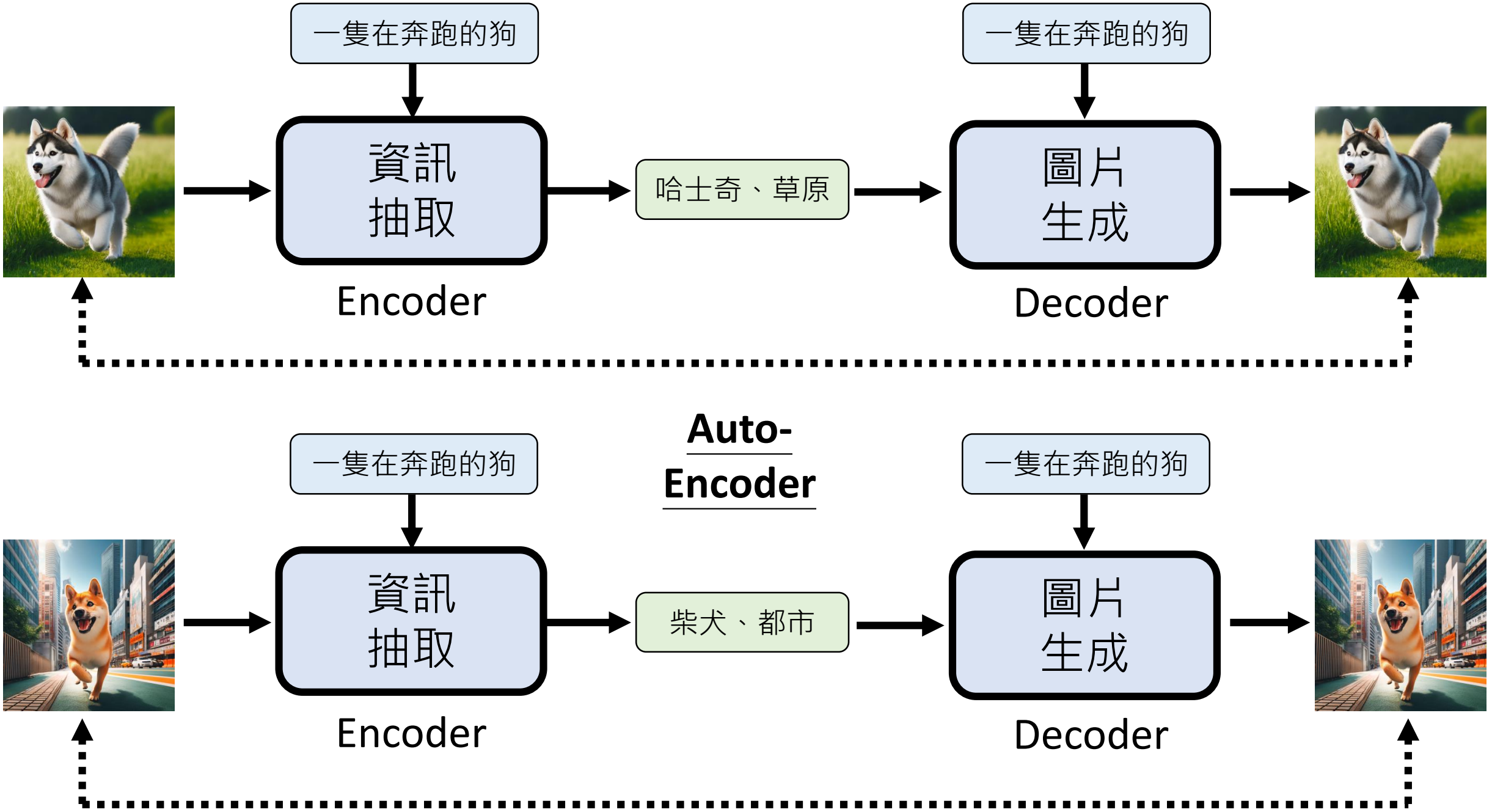


一隻在奔跑的狗 1 7 3

一隻在奔跑的狗 71 13 8

如何處理腦補問題


一隻在奔跑的狗
資訊
抽取
哈士奇、草原
一隻在奔跑的狗
圖片
生成
一隻在奔跑的狗
資訊
抽取
柴犬、都市
一隻在奔跑的狗
圖片
生成


一隻在奔跑的狗
資訊
抽取
Encoder
哈士奇、草原
一隻在奔跑的狗
圖片
生成
Decoder
Auto-
Encoder
一隻在奔跑的狗
資訊
抽取
Encoder
柴犬、都市
一隻在奔跑的狗
圖片
生成
Decoder

一隻在奔跑的狗

資訊
抽取

Encoder

哈士奇、草原

$\begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix}$

一隻在奔跑的狗

圖片
生成

Decoder

一隻在奔跑的狗

資訊
抽取

Encoder

柴犬、都市

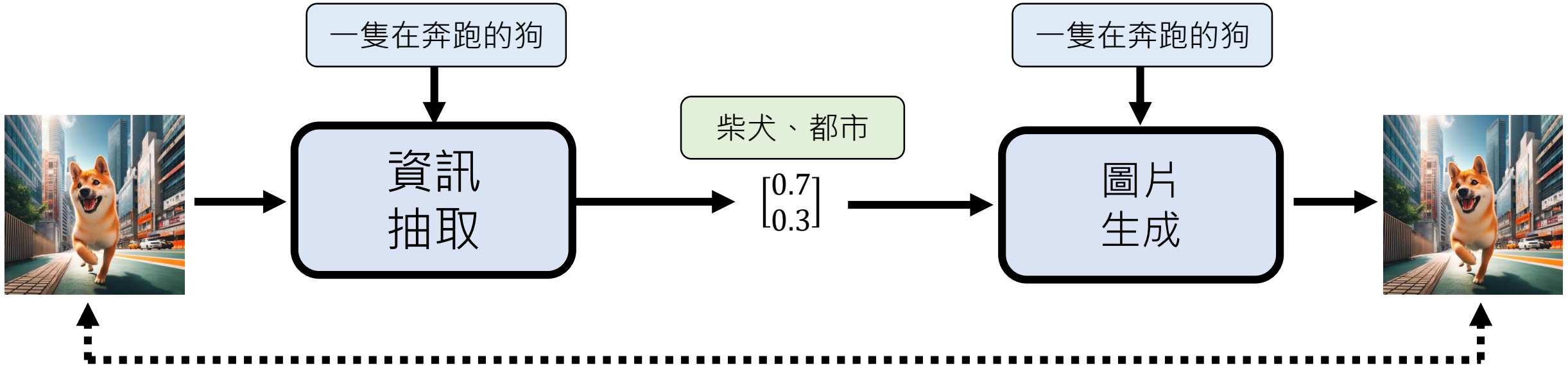
$\begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$

一隻在奔跑的狗

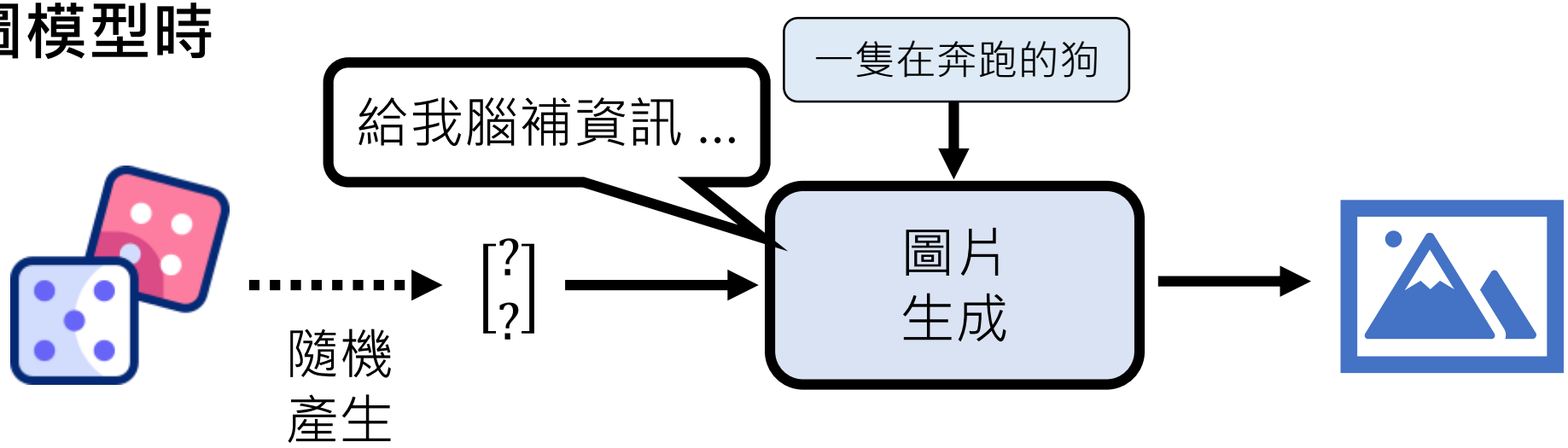
圖片
生成

Decoder

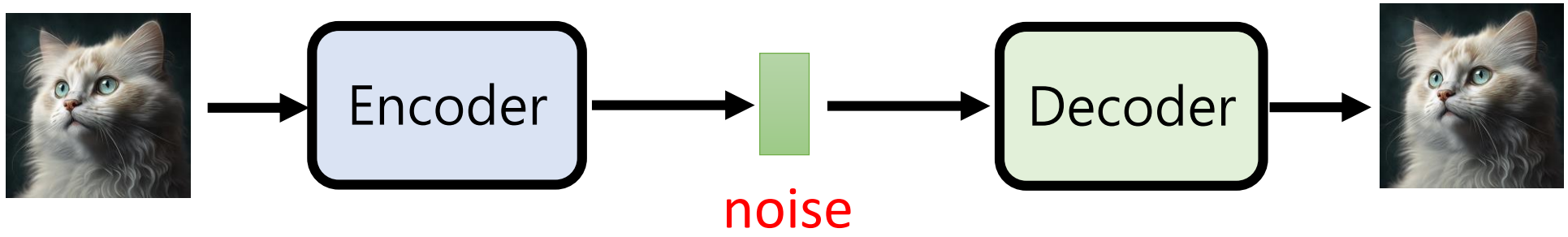
訓練文字生圖模型時



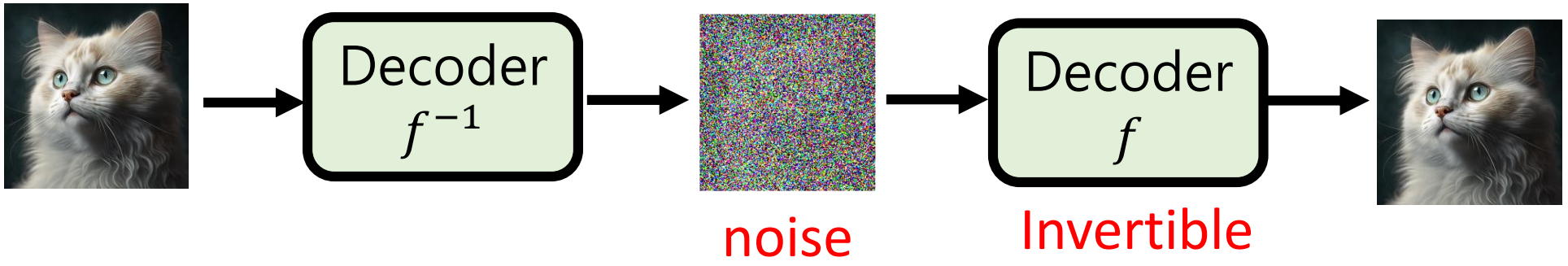
測試文字生圖模型時



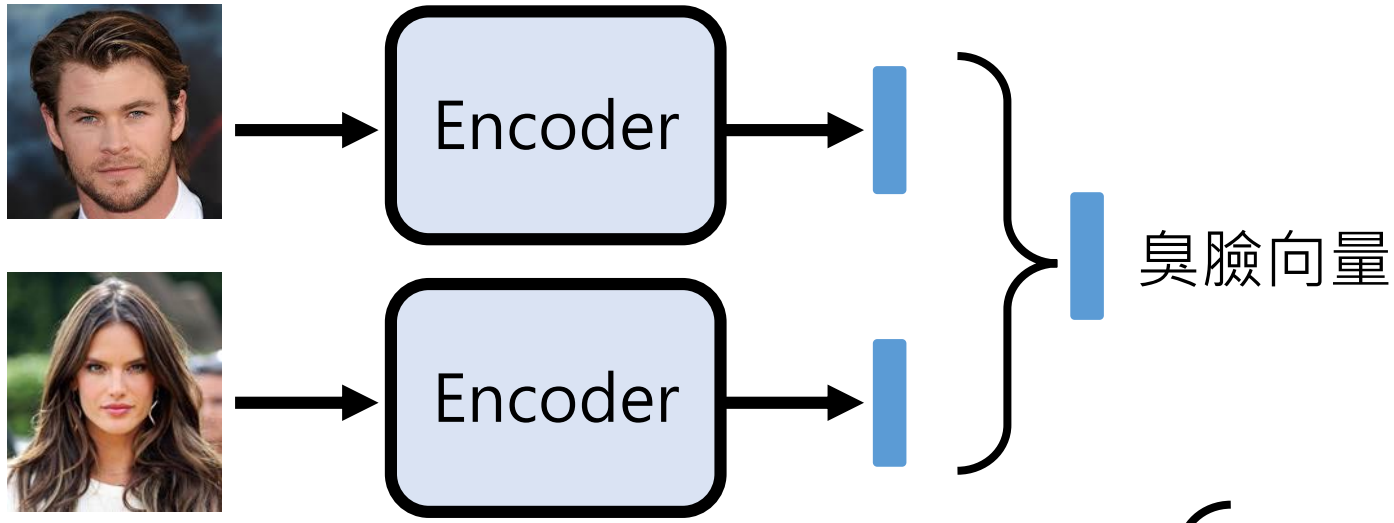
VAE

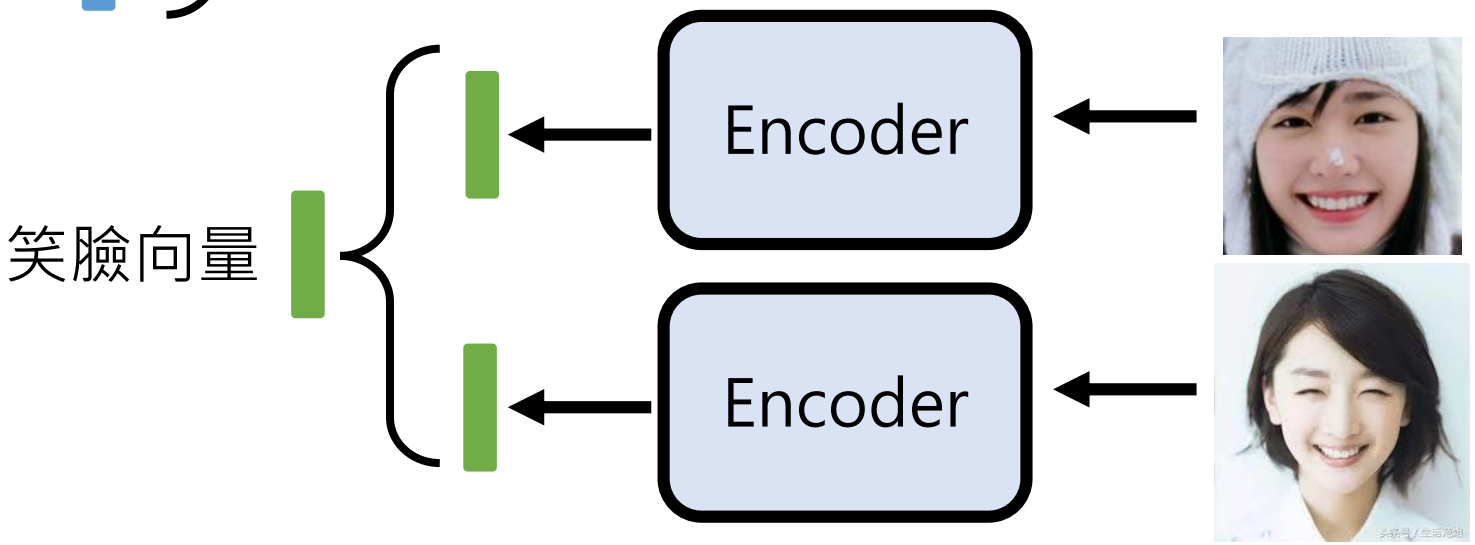


Flow-based



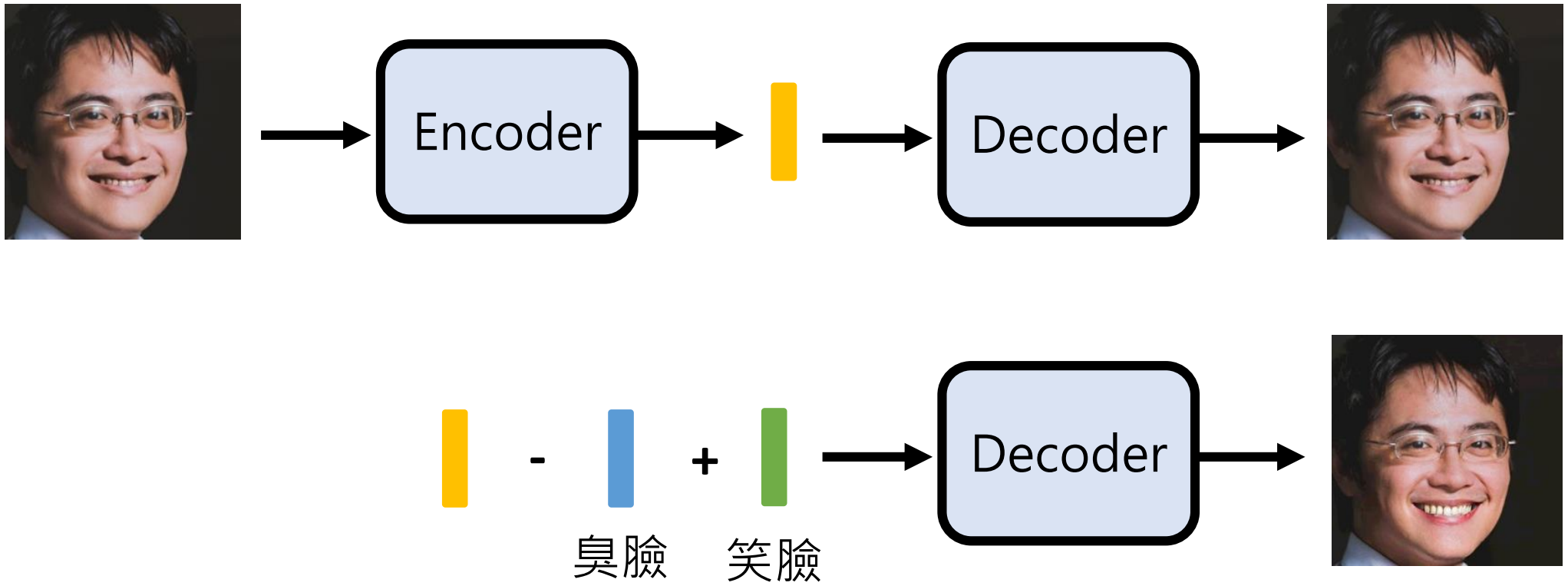
有資訊的雜訊 (noise)





有資訊的雜訊 (noise)

臭臉向量　笑臉向量

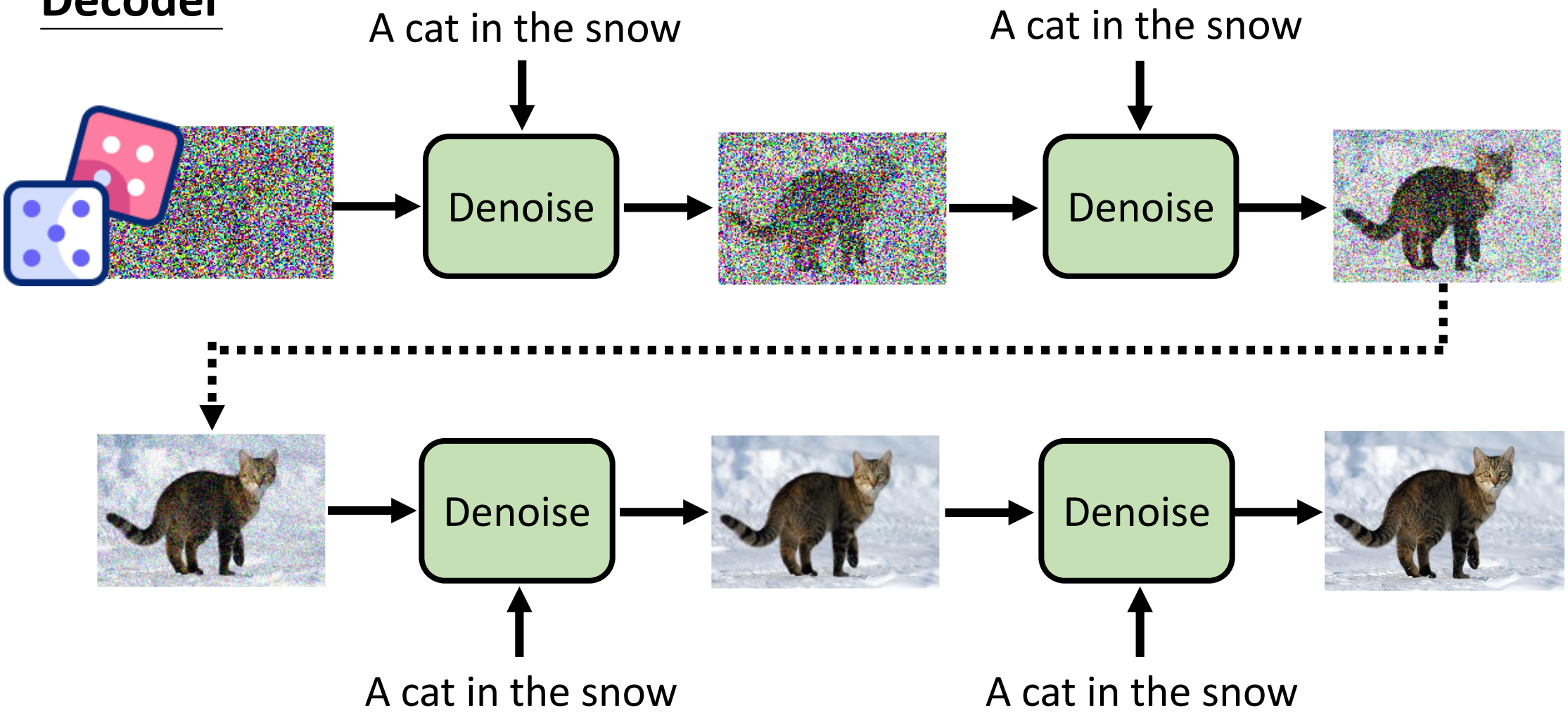


Powered by: https://openai.com/blog/glow/

此處是極度簡化後的講法，詳細說明請見：
https://www.youtube.com/watch?v=azBugJzmz-o&list=PLJV_el3uVTsNi7PgekEUFsyVllAJXRsP-

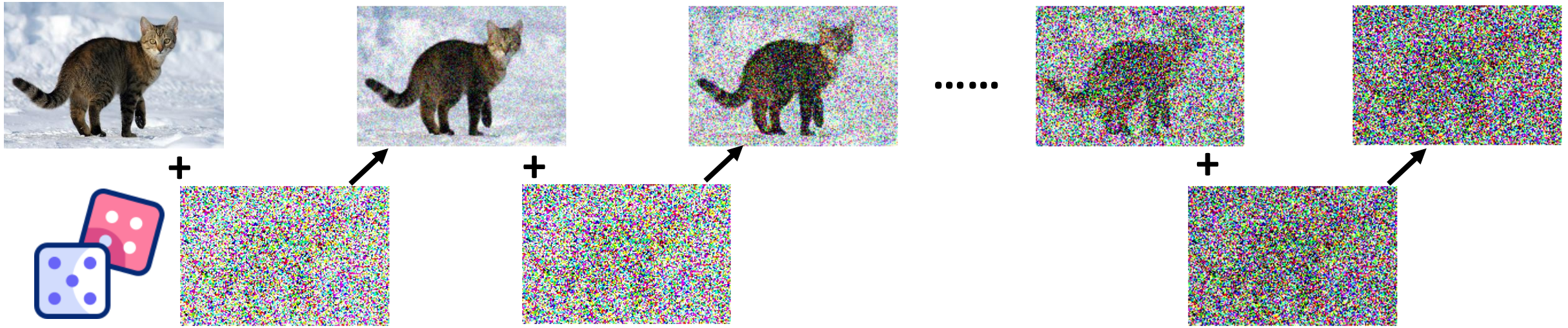
Diffusion Model

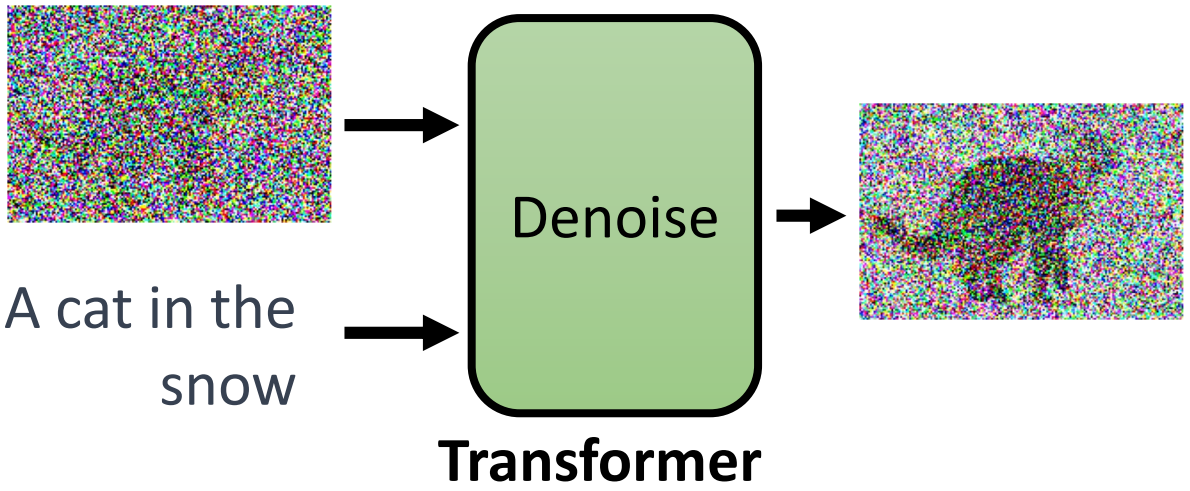
Decoder

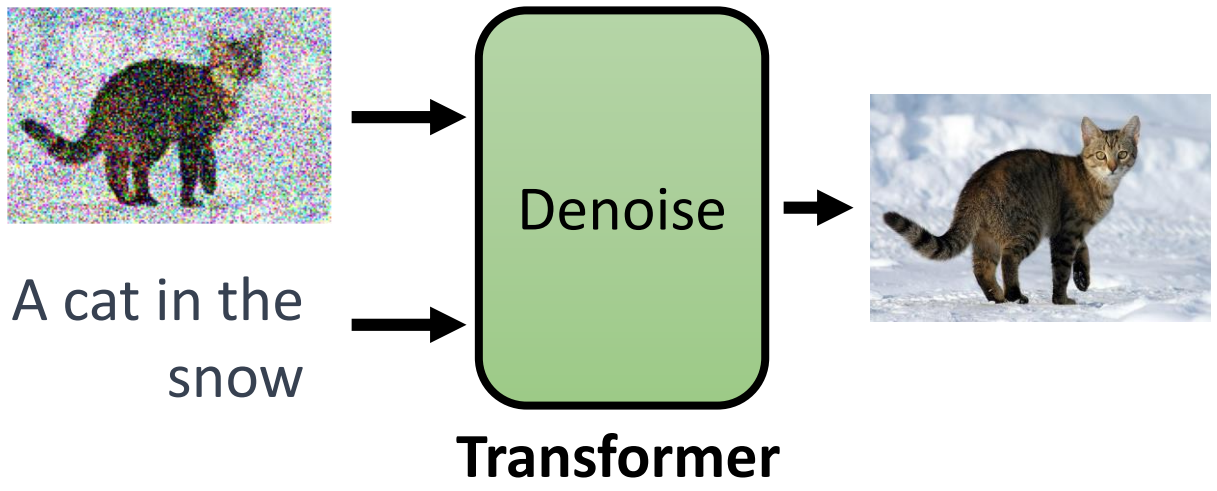


A cat in the snow

此處是極度簡化後的講法，詳細說明請見：
https://www.youtube.com/watch?v=azBugJzmz-o&list=PLJV_el3uVTsNi7PgekEUFsyVllAJXRsP-

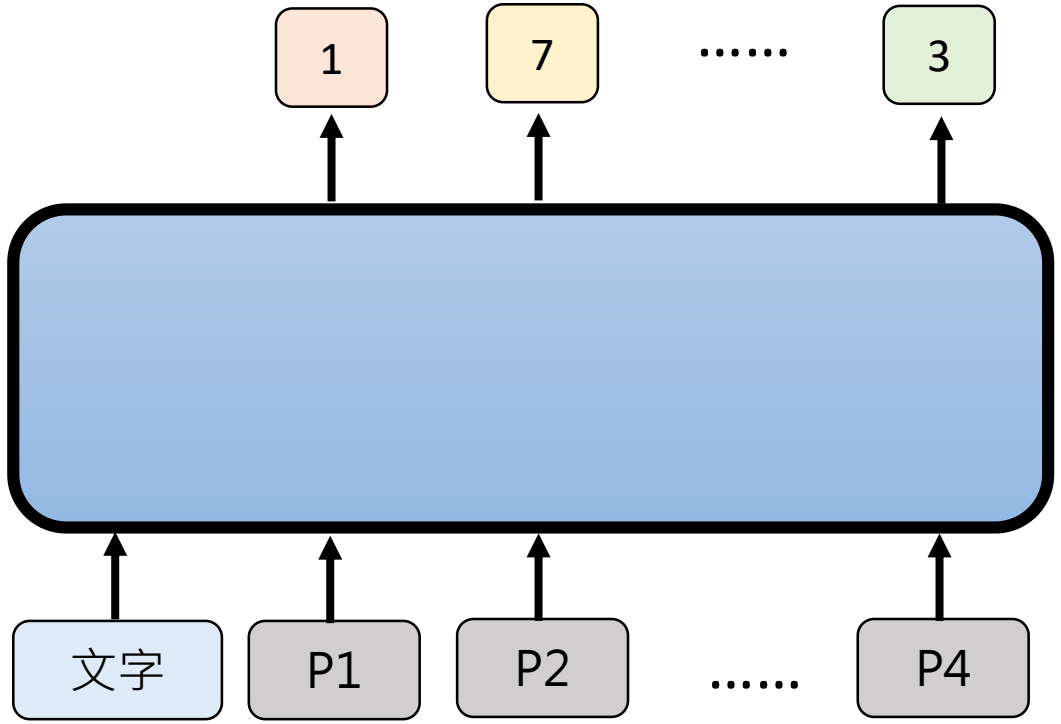


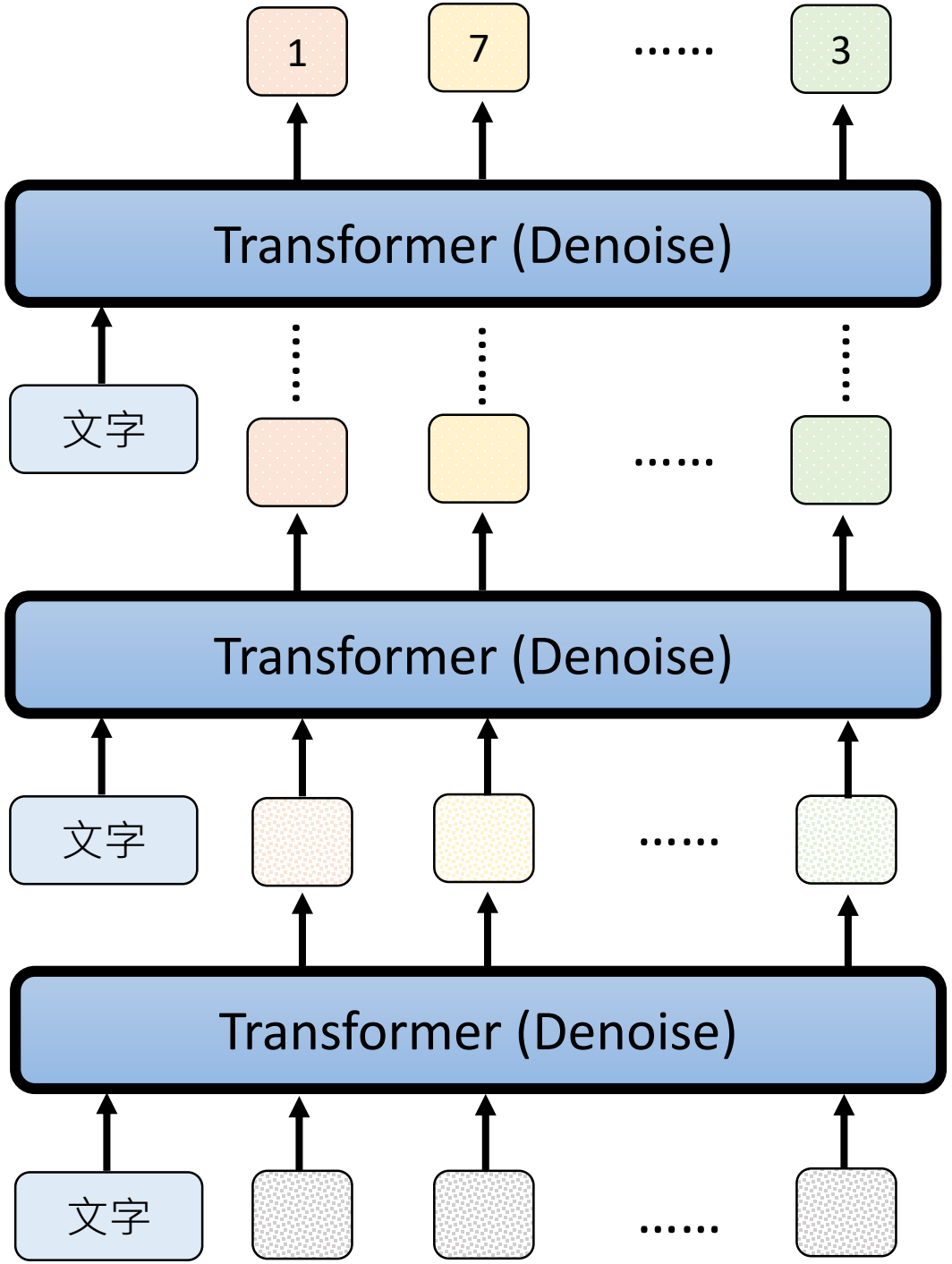




Diffusion Transformer

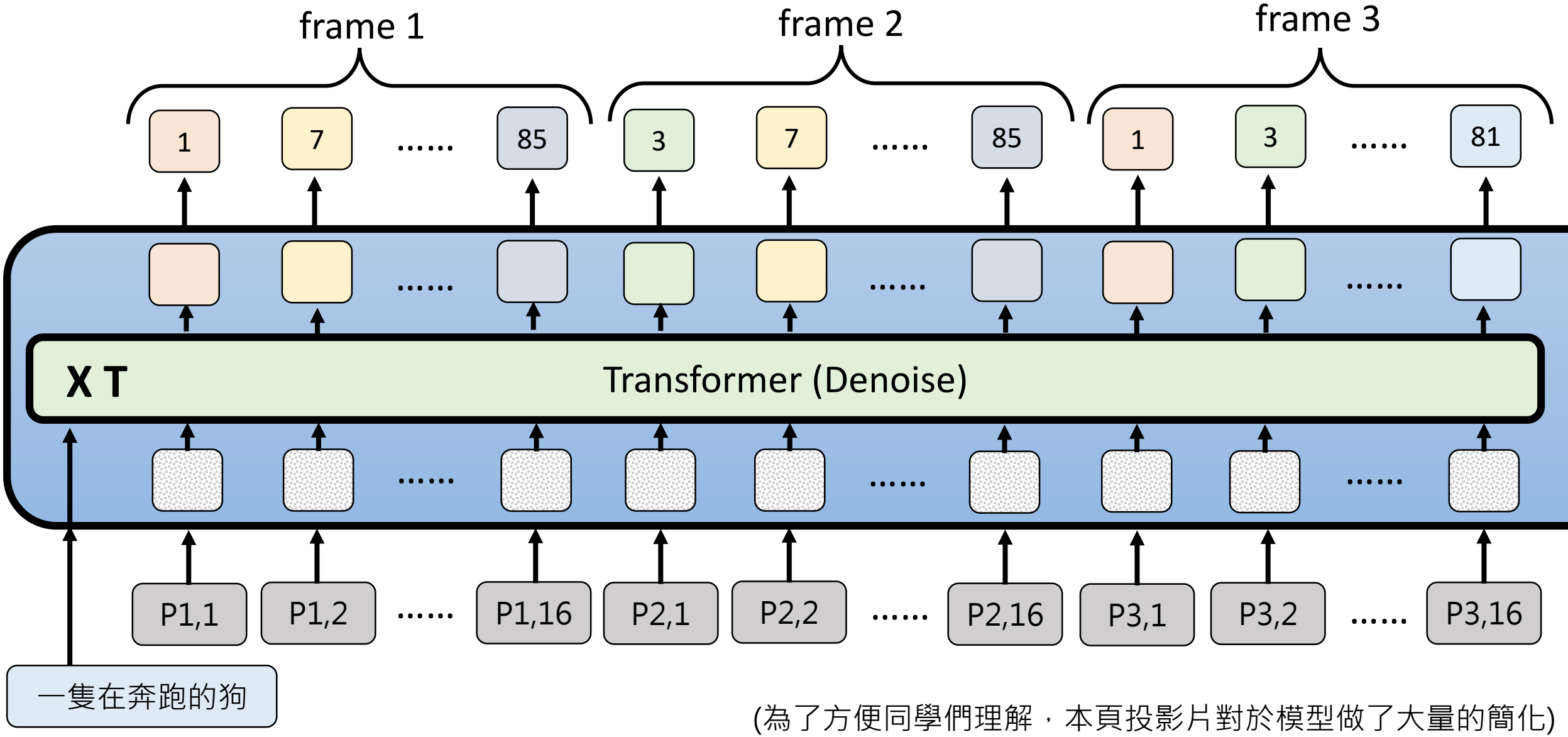
https://arxiv.org/abs/2212.09748





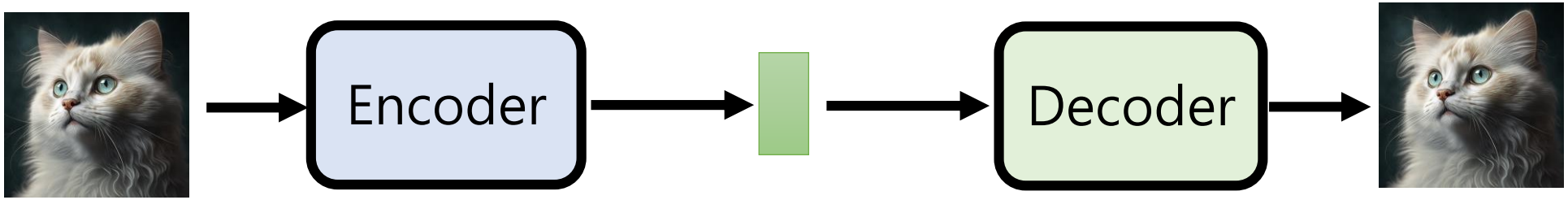
(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

文字生影片的模型

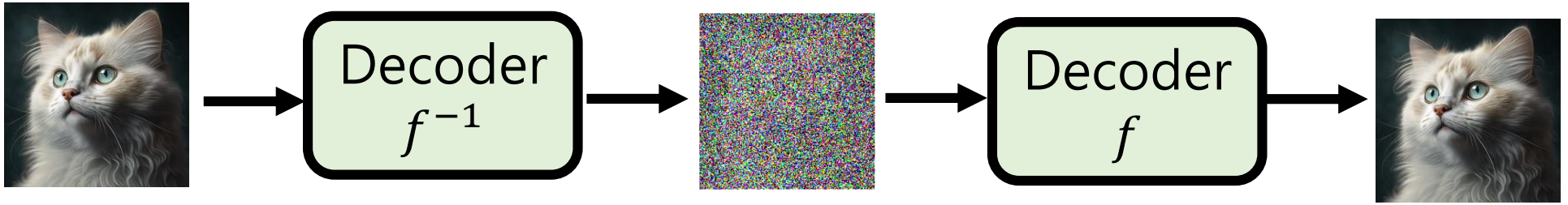


(為了方便同學們理解，本頁投影片對於模型做了大量的簡化)

VAE

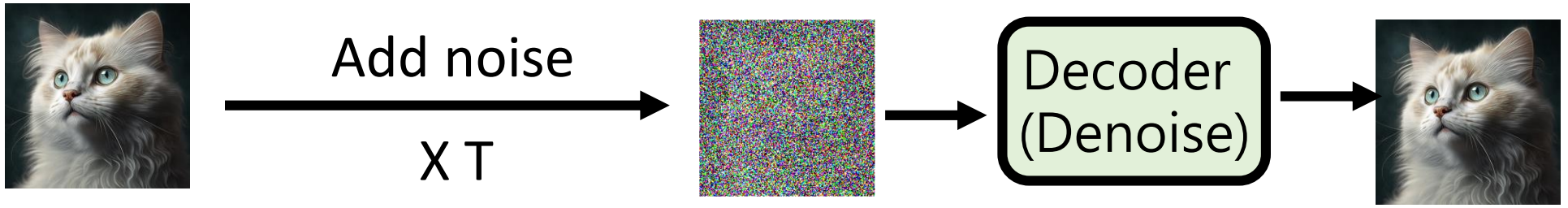


Flow-based



Invertible

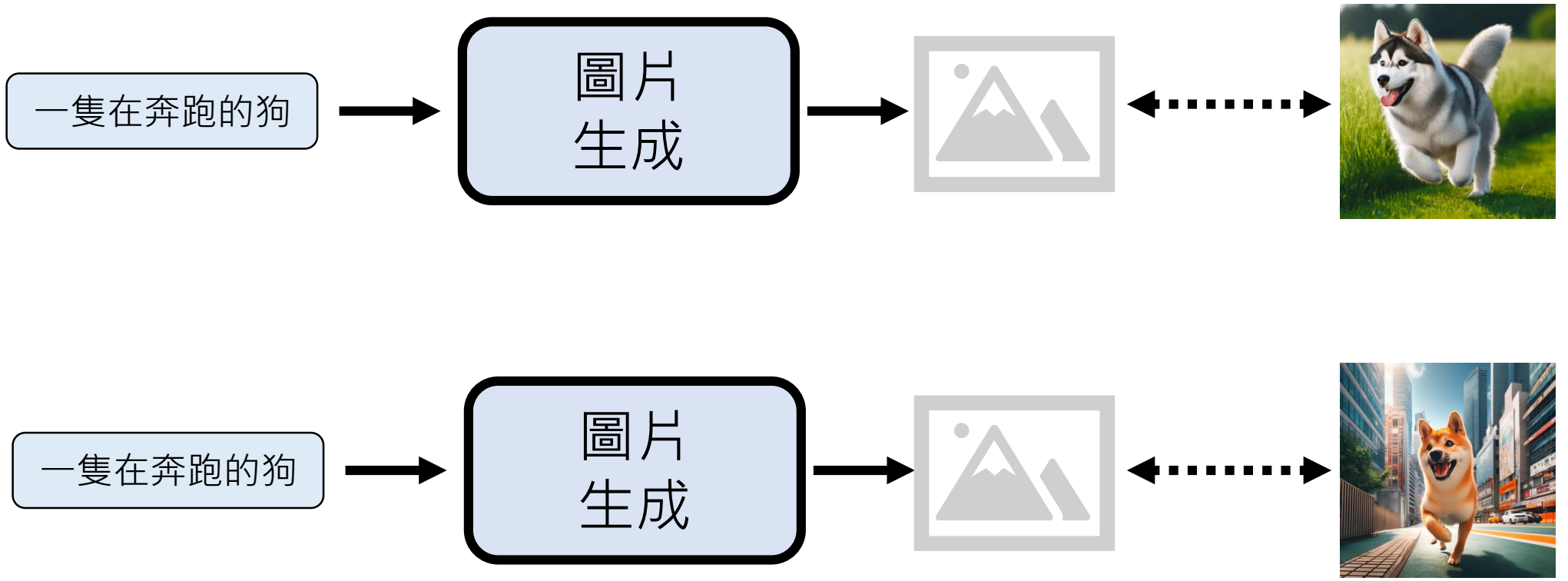
Diffusion



Forward Process

Reverse Process

Generative Adversarial Network (GAN)

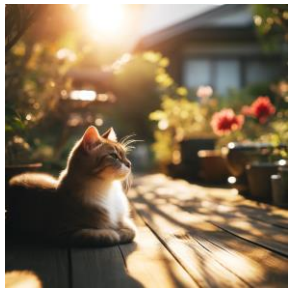


一隻在奔跑的狗

Discriminator

一隻在奔跑的狗

Discriminator



陽光下的貓

Discriminator

陽光下的貓

Discriminator

一隻在奔跑的狗

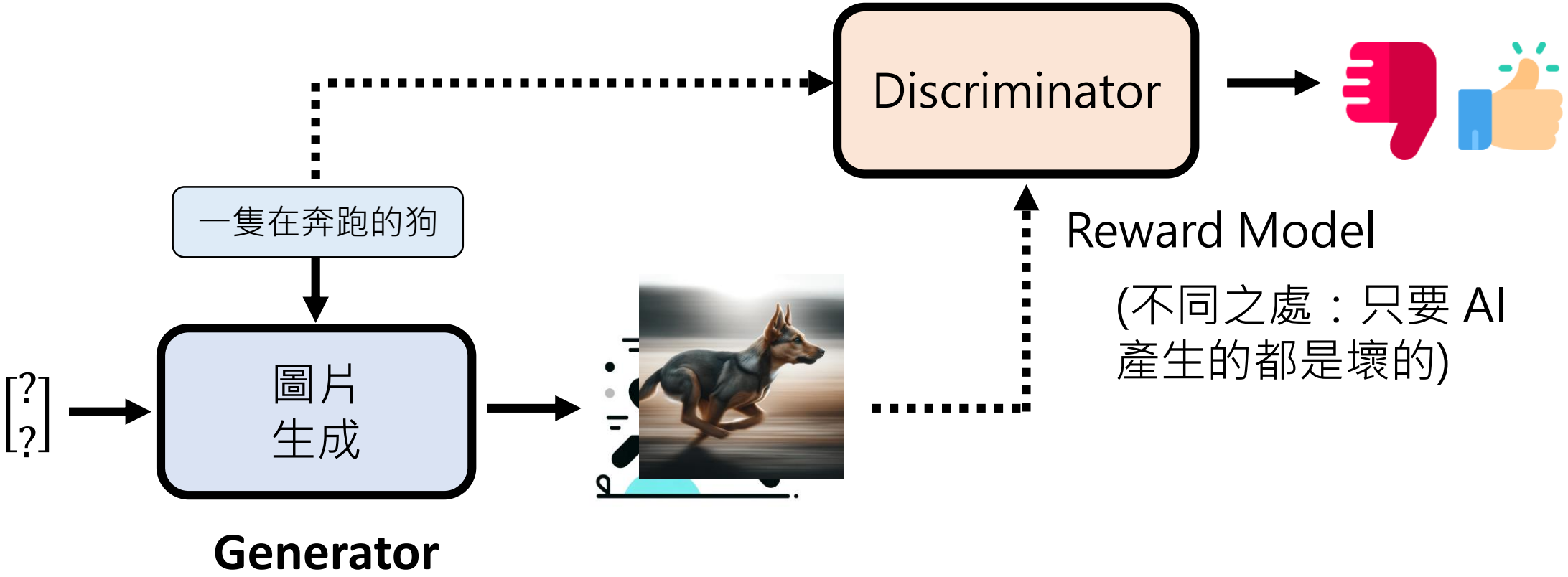
圖片生成

陽光下的貓

圖片生成

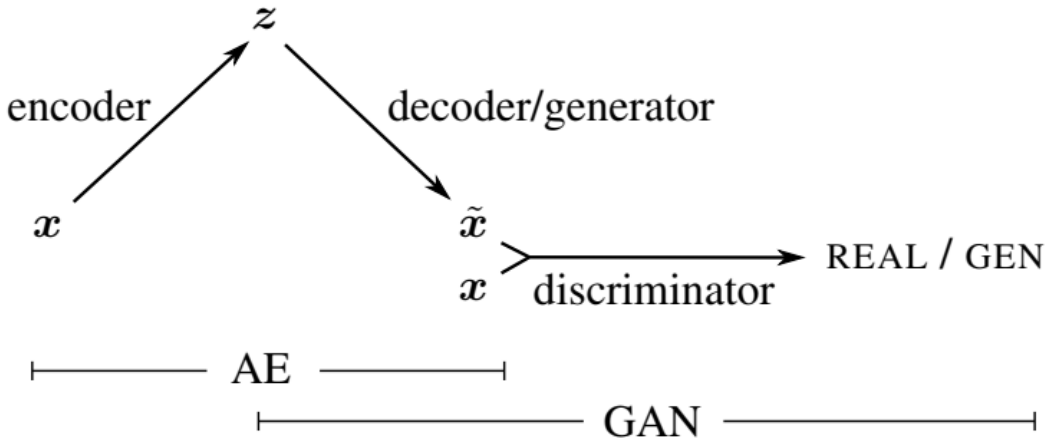
Generative Adversarial Network (GAN)

Discriminator 和 Generator 會交替訓練



比較 RLHF

GAN 是個外掛



VAE + GAN

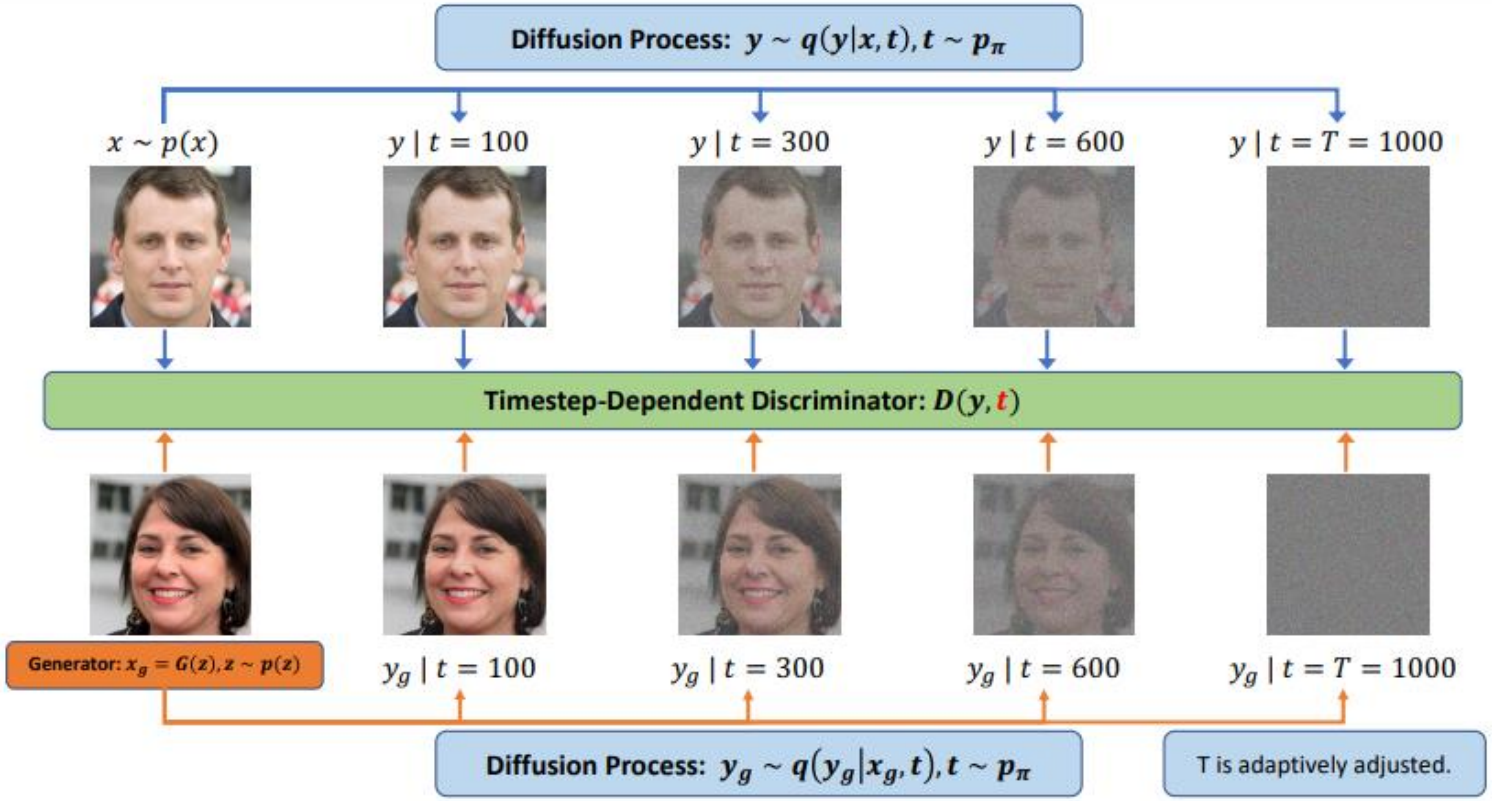
https://arxiv.org/abs/1512.09300

Flow + GAN

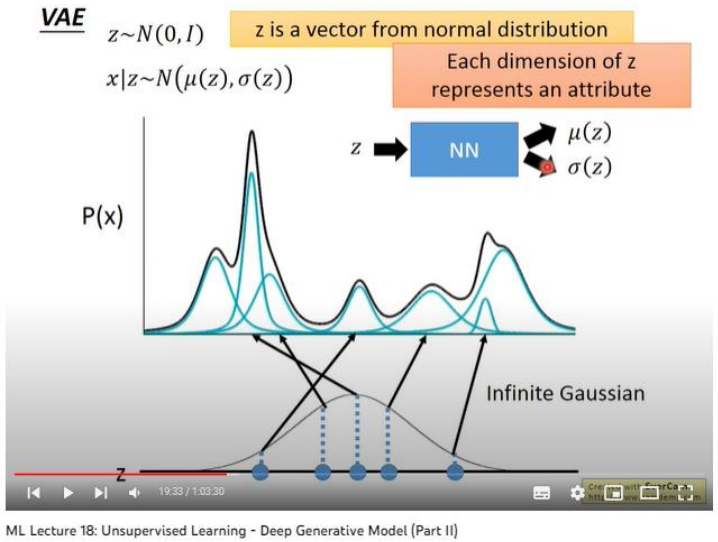
https://arxiv.org/abs/1705.08868

Diffusion + GAN

https://arxiv.org/abs/2206.02262



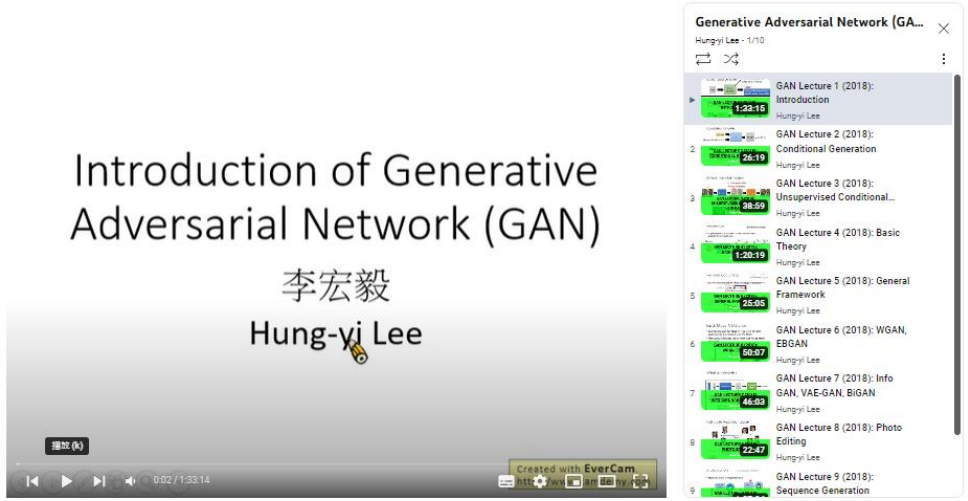
VAE



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

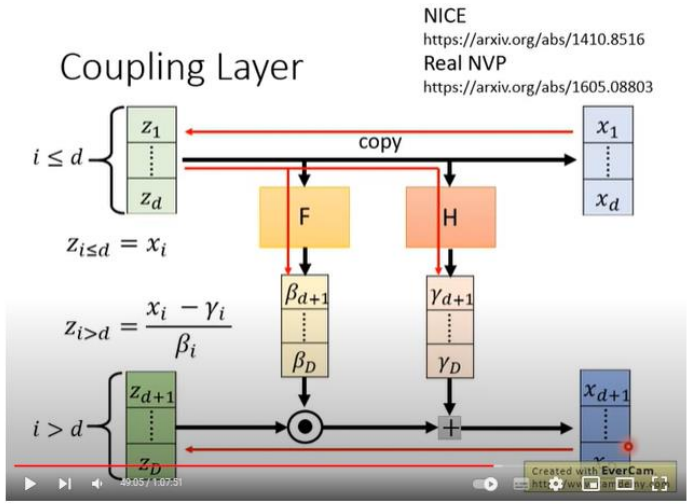
https://youtu.be/8zomhgKrsmQ (2016 機器學習)

GAN



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

Flow-based



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)

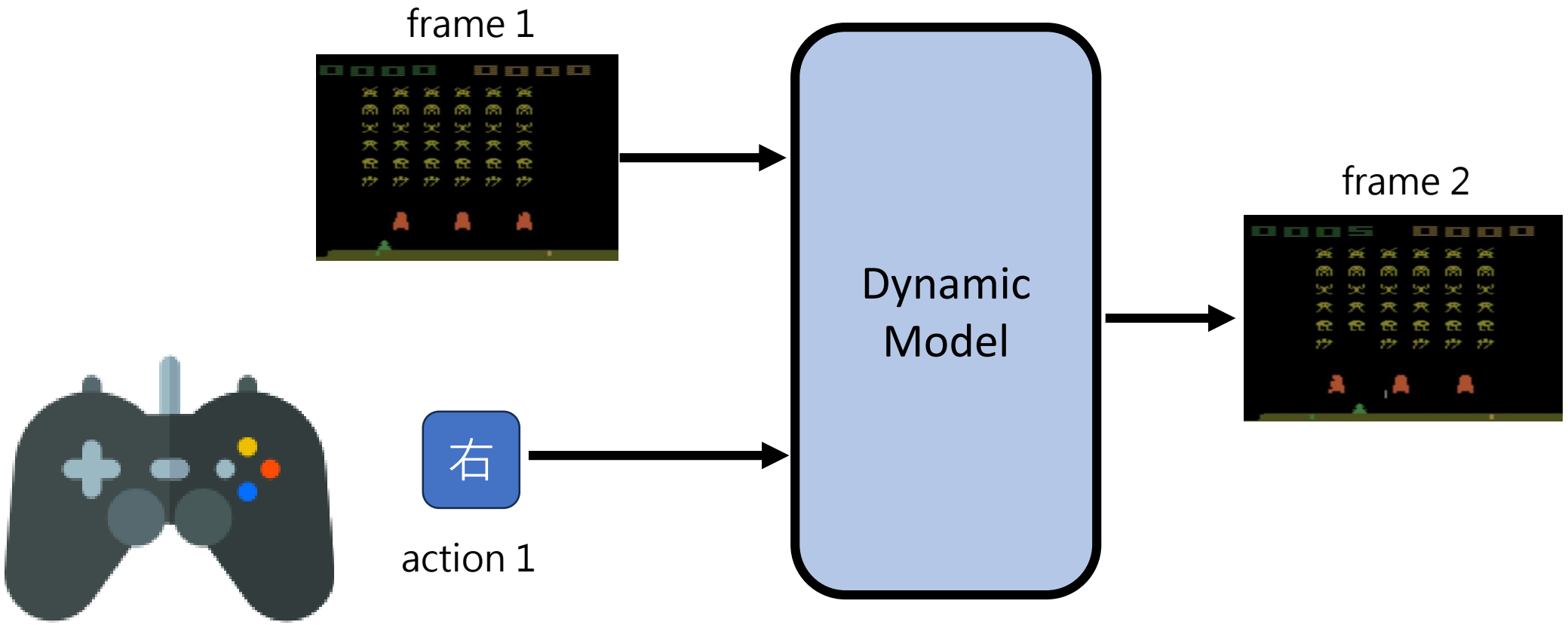
有沒有可能跟生成的影像有更強的互動

例如：直接操控這個人要走去哪裡

直接做個開放世界遊戲？

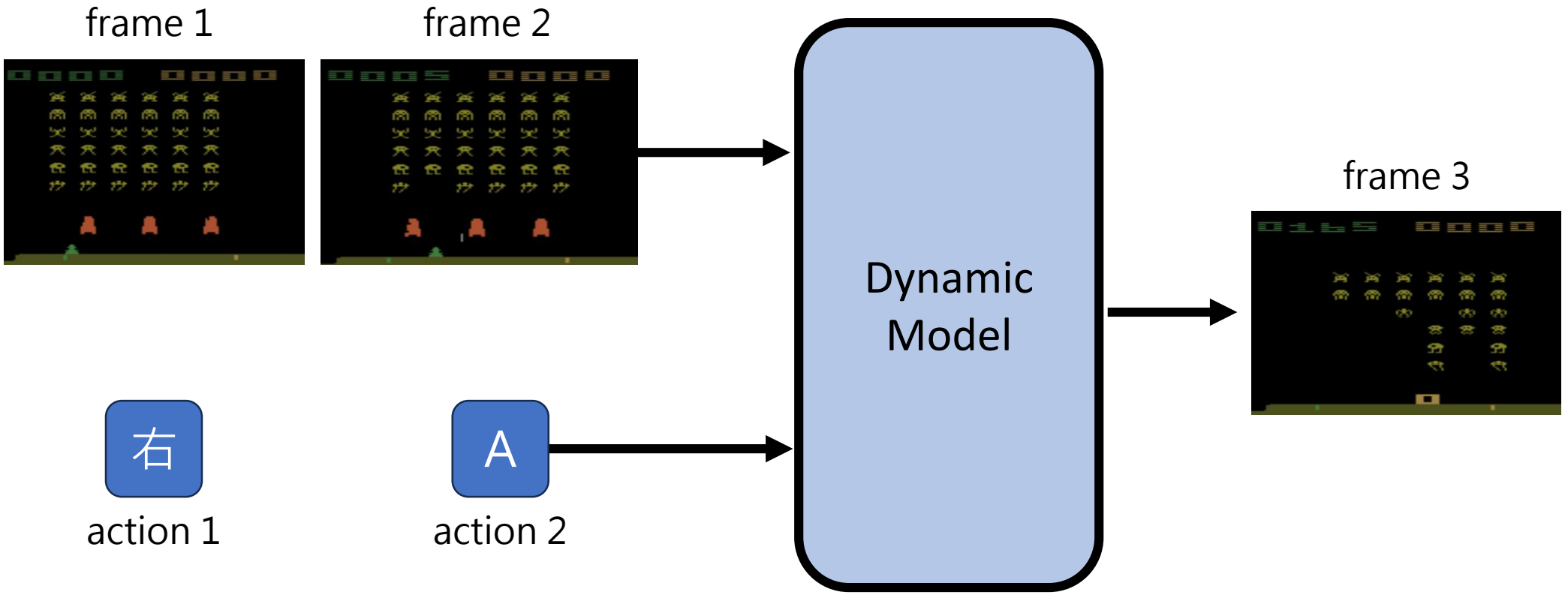
Genie: Generative Interactive Environments

https://arxiv.org/abs/2402.15391



Genie: Generative Interactive Environments

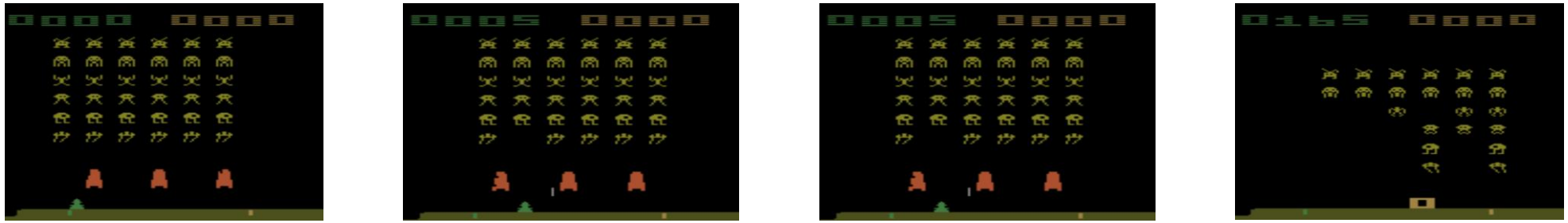
https://arxiv.org/abs/2402.15391



Genie: Generative Interactive Environments

- 訓練資料？

可以蒐集大量遊戲影片



右 A 無 上

沒有使用者輸入的動作

一隻在奔跑的狗

資訊
抽取

Encoder

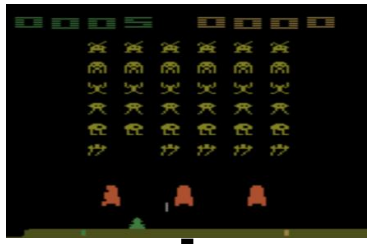
哈士奇、草原

$\begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix}$

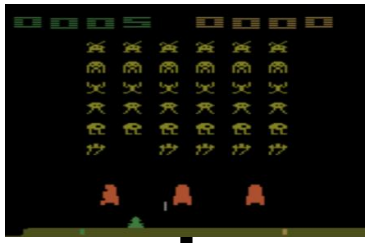
一隻在奔跑的狗

圖片
生成

Decoder



之前遊
戲畫面



之前遊
戲畫面

下一個畫面

動作
抽取

Latent Action

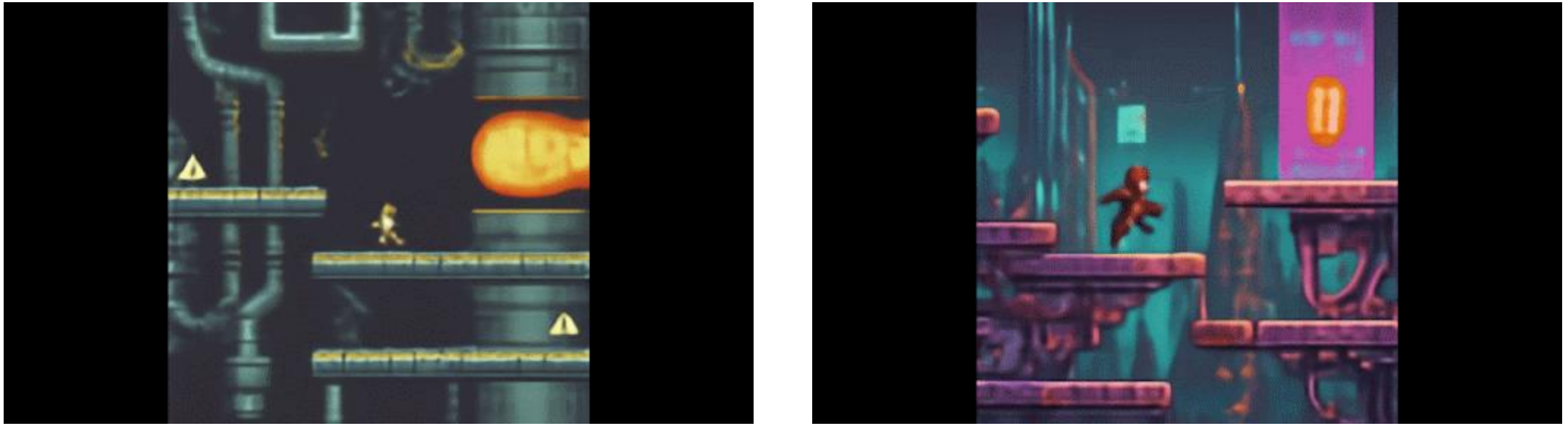
1

編號 1 - 8

圖片
生成

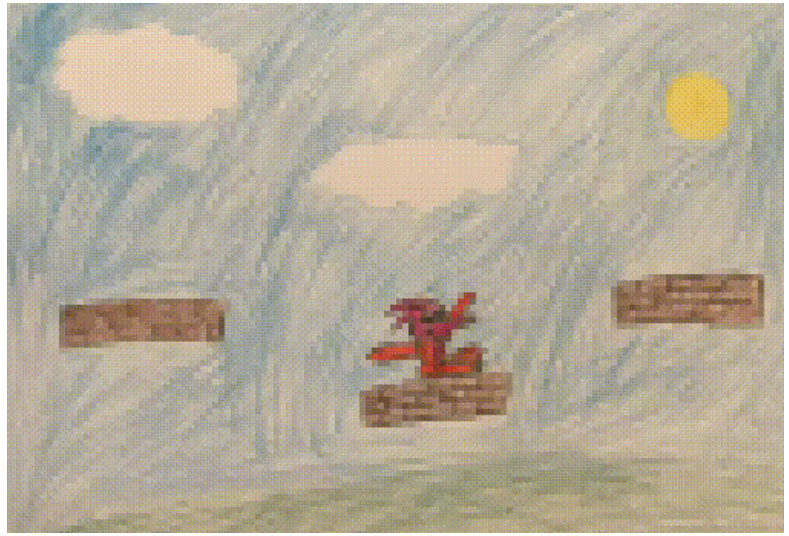
下一個畫面

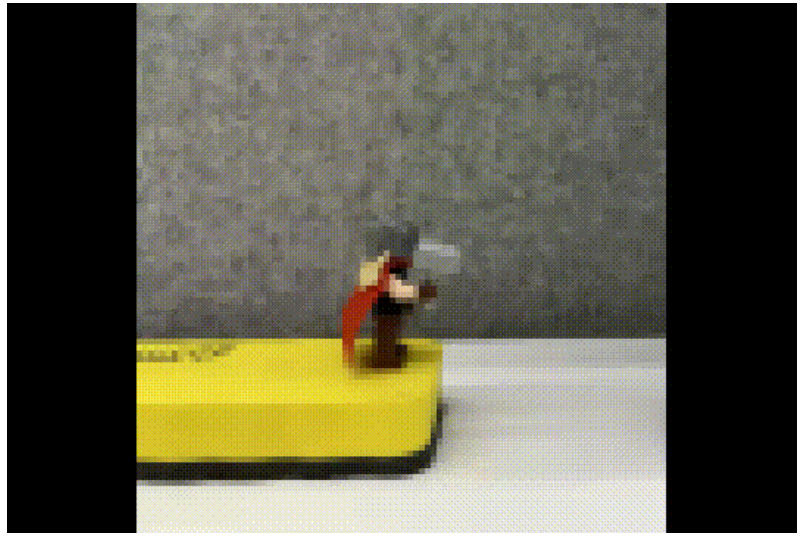
Genie: Generative Interactive Environments



Latent action: 6, 6, 7, 6, 7, 6, 5, 5, 2, 7

Source: https://sites.google.com/view/genie-2024/home





Source: https://sites.google.com/view/genie-2024/home

Sora 技術報告標題

https://arxiv.org/abs/2309.17080
https://wayve.ai/thinking/scaling-gaia-1/