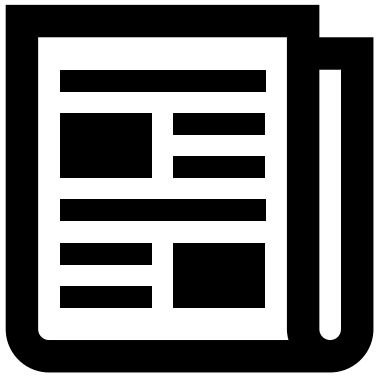
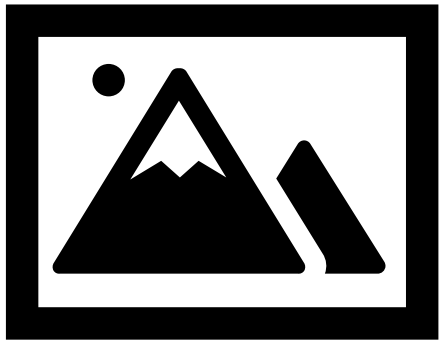
生成式學習的兩種策略：各個擊破還是一次到位

生成有結構的複雜物件



文句

由 token 所構成



影像

由像素所組成

語音

16k 取樣頻率，
每秒有 16,000 個取樣點

- 中文的 token 就是「字」
- 英文的 token 是 word piece　　unbreakable → un break able

生成影片

Source of examples:
https://imagen.research.google/video/

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

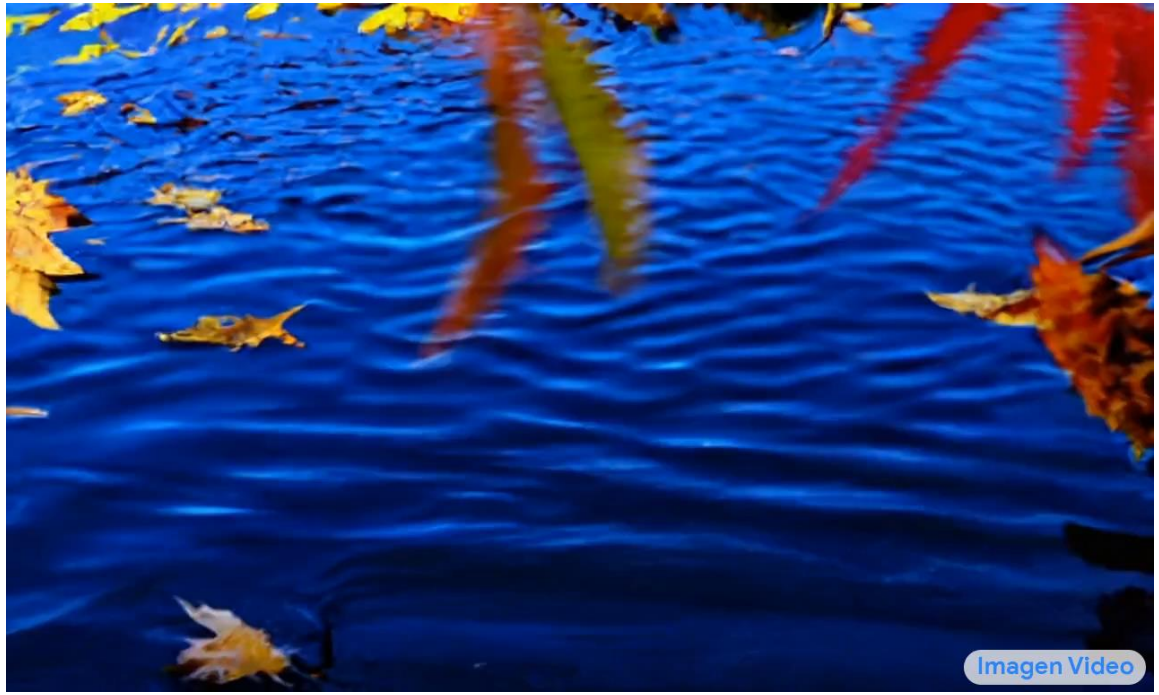
A teddy bear
washing dishes.

生成影片

Source of examples:
https://imagen.research.google/video/

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

A bunch of autumn leaves falling on a calm lake to form the text ‘Imagen Video’. Smooth.



生成語音

- InstructTTS: https://arxiv.org/abs/2301.13662

語調高昂，聲音宏亮，內心非常憤慨

聲音高，語氣嚴厲，大聲呵斥

鎮定從容，語氣平和，語調穩定

語氣中惆悵含有一絲苦澀

聲音難過，鬱鬱寡歡，傾訴的語氣中透露出疲憊落寞的情感

Source of examples: http://dongchaoyang.top/InstructTTS/

生成聲音

Source of examples: https://audioldm.github.io/

- Text-to-audio: https://arxiv.org/abs/2301.12503

Two space shuttles are fighting in the space.

Describe the sound of the ocean

ChatGPT: The steady crashing of waves against the shore,high fidelity, the whooshing sound of water receding back into the ocean, the sound of seagulls and other coastal birds, and the distant sound of ships or boats.

Describe what does a pop music sound

ChatGPT: Pop music that upbeat, catchy, and easy to listen, high fidelity, with simple melodies, electronic instruments and polished production.

A man is speaking in a *huge room*.

A man is speaking in a *small room*.

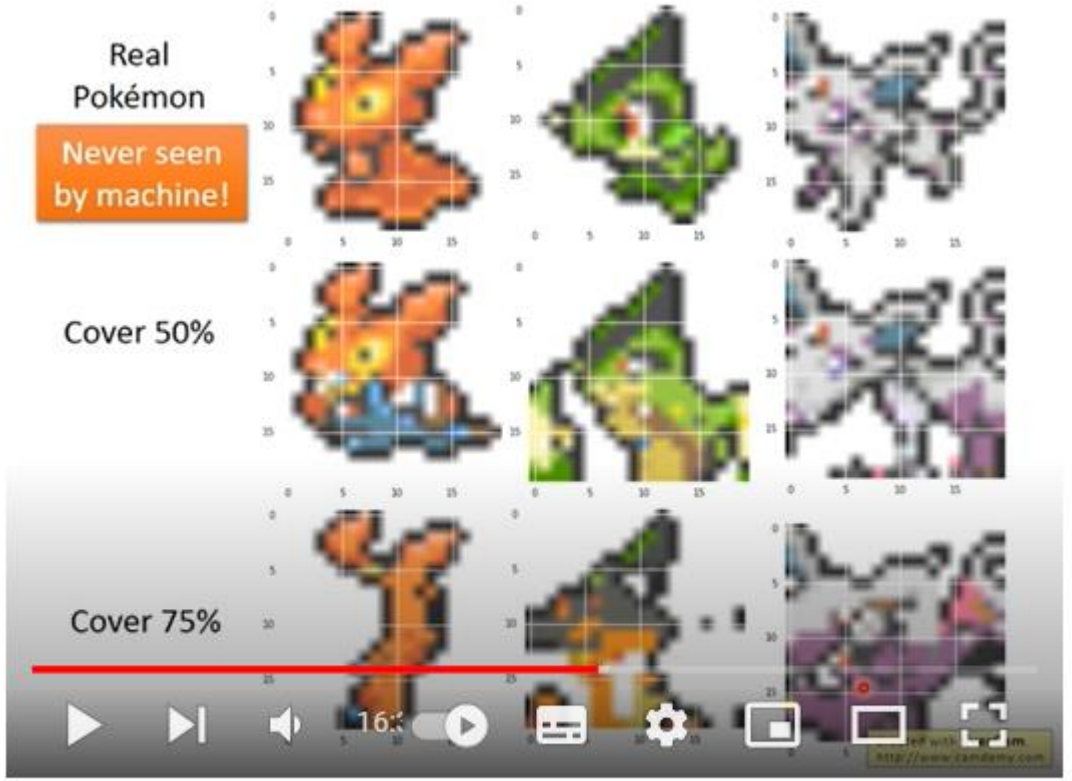
A man is speaking in a *studio*.

策略一：各個擊破 Autoregressive (AR) model

機器學習是甚麼？ → 文字生成 → 機 器 學 …… [END]

一隻在奔跑的狗 → 影像生成 → 4 x 4 images

策略一：各個擊破



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

用一次只生一個像素的方式
讓機器自己畫寶可夢

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

(2016 年《機器學習》秋季班上課錄影)

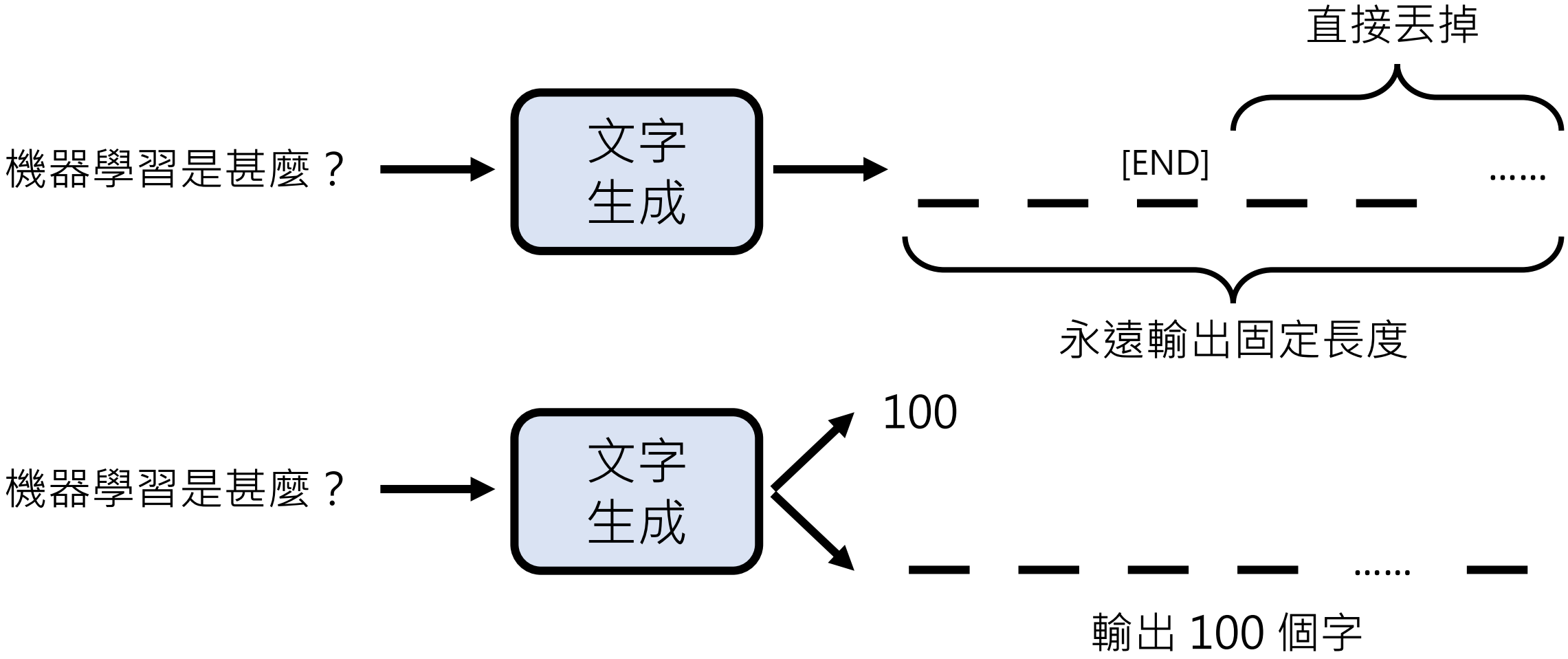
策略二：一次到位 Non-autoregressive (NAR) model

如何決定長度？

機器學習是甚麼？ → 文字生成 → 機 器 學 …… [END]

一隻在奔跑的狗 → 影像生成 → 4 x 4 images

策略二：一次到位 Non-autoregressive (NAR) model



各個擊破 vs. 一次到位 (生成速度)

各個擊破

每一個字(像素)都需要等前面的字(像素)生成

y_1 y_2 y_3 y_4 …….

一次到位

只要有足夠的平行運算能力，就可以很快

y_1 y_2 y_3 y_4 …….

這就是為什麼影像生成常用「一次到位」

各個擊破 vs. 一次到位 (生成品質)

請問李宏毅的職業是甚麼？

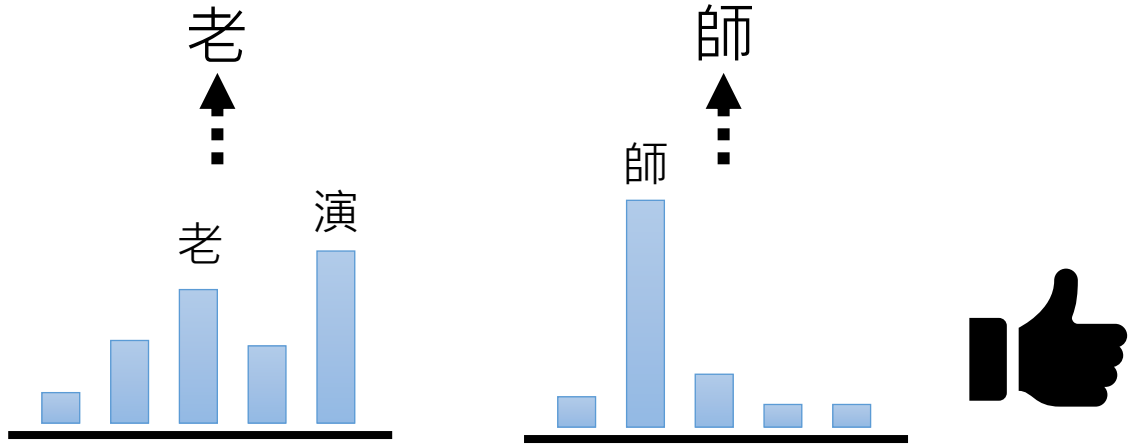
(可以回答演員或老師)

各個擊破 vs. 一次到位 (生成品質)

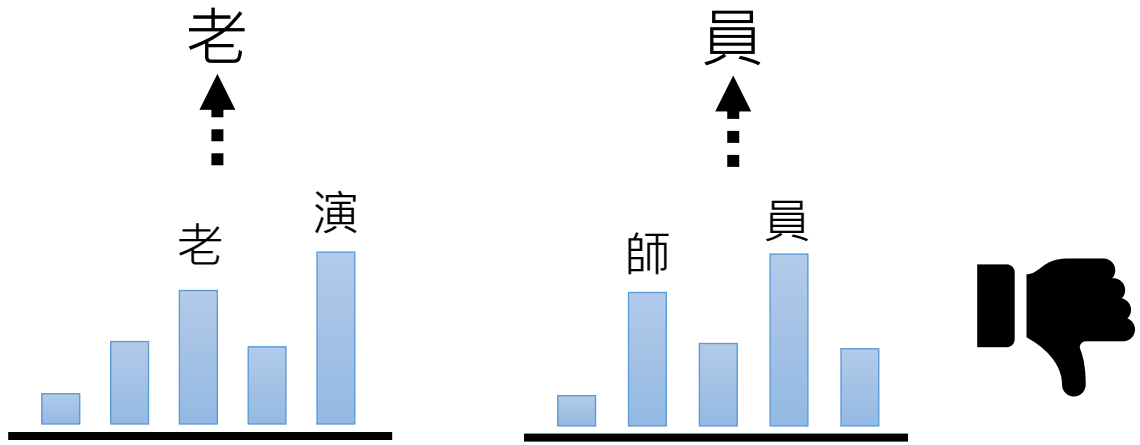
請問李宏毅的職業是甚麼？

(可以回答演員或老師)

各個擊破



一次到位



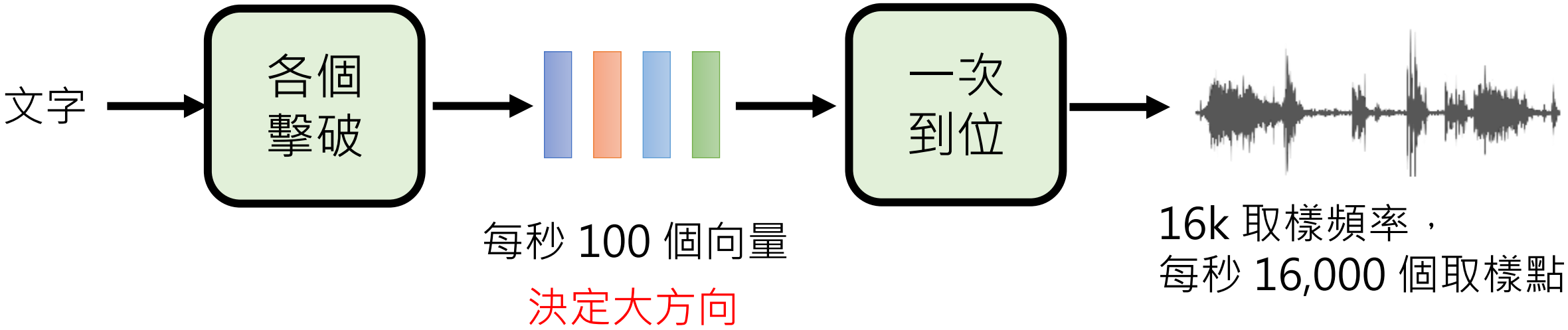
各個擊破 vs. 一次到位

	各個擊破 (Autoregressive, AR)	一次到位 (Non-autoregressive, NAR)
速度		勝
品質	勝	
應用	常用於文字	常用於影像

有沒有可能兩種策略截長補短？

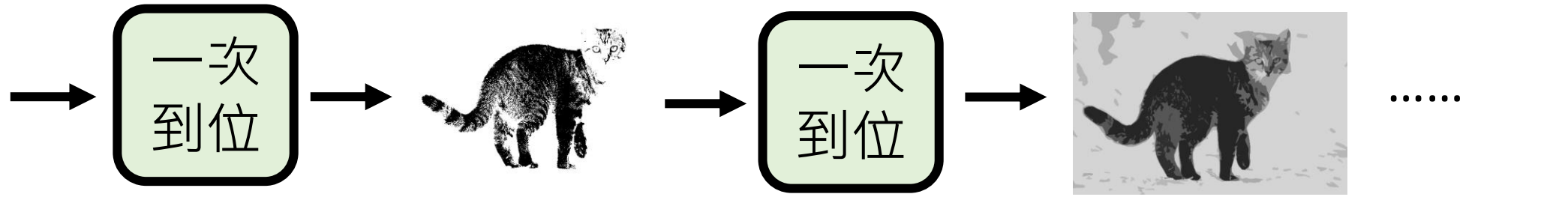
各個擊破 + 一次到位

- 以語音合成為例

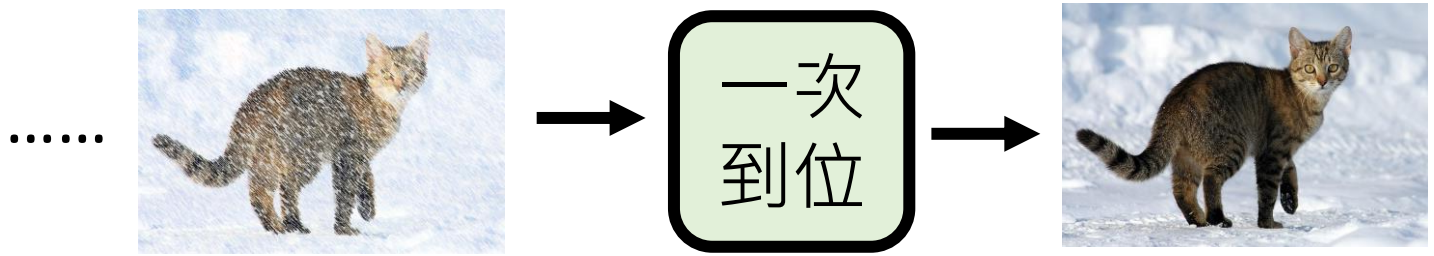


各個擊破 + 一次到位

一隻在
雪地的貓



「一次到位」改成
「 N次到位」



疑？這聽起來很像 Diffusion Model