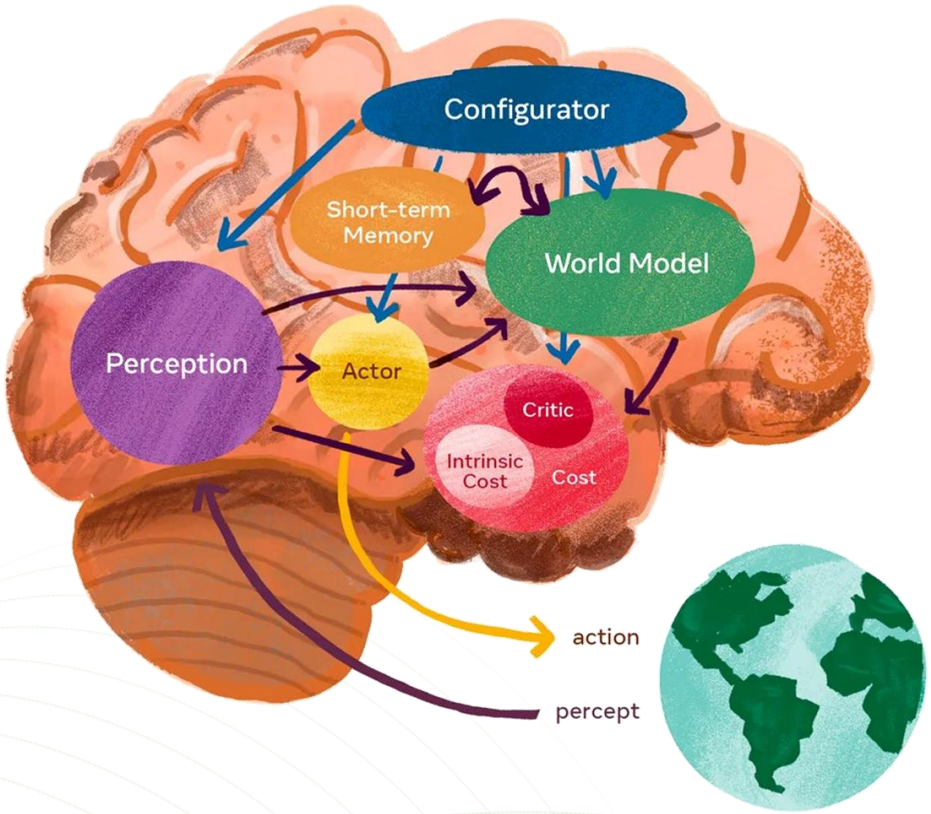


World Models
世界模型

World Models

All systems, designed to learn internal representations of how the world works.

- Yann Lecun

World Models - 世界模型

A foundational model that understand physical reality, anticipate outcomes, and plan efficient strategies.

Developments of World Models

https://www.youtube.com/watch?v=LY_J9-7BjOc

- **Jan 2023: Dreamer V3.**
  The algorithm uses an internal world model to plan and train an agent from simulated "dreams".
- **Sep 2024: World Labs.**
  World Labs is building "Large World Models" to enable spatial intelligence, allowing AI to perceive, generate, and interact with the 3D world.
- **Mar 2025: NVIDIA COSMOS.**
  The platform is designed to accelerate training for physical AI like robots and autonomous vehicles.
- **Jun 2025: Meta V-JEPA 2.**
  It enables "zero-shot" robotic planning, allowing an AI to perform new tasks without specific training.
- **Aug 2025: Google DeepMind Genie 3.**
  An interactive world model for generating real-time, persistent, and 3D environments from text or images.
- **Sep 2025: Meta Vision Language World Model.**
  A foundation model trained for language-based world modeling: perceive the environment through visual observations and predict world evolution using language-based abstraction.
- **Nov 2025: Marble World Lab**

World Models - 世界模型

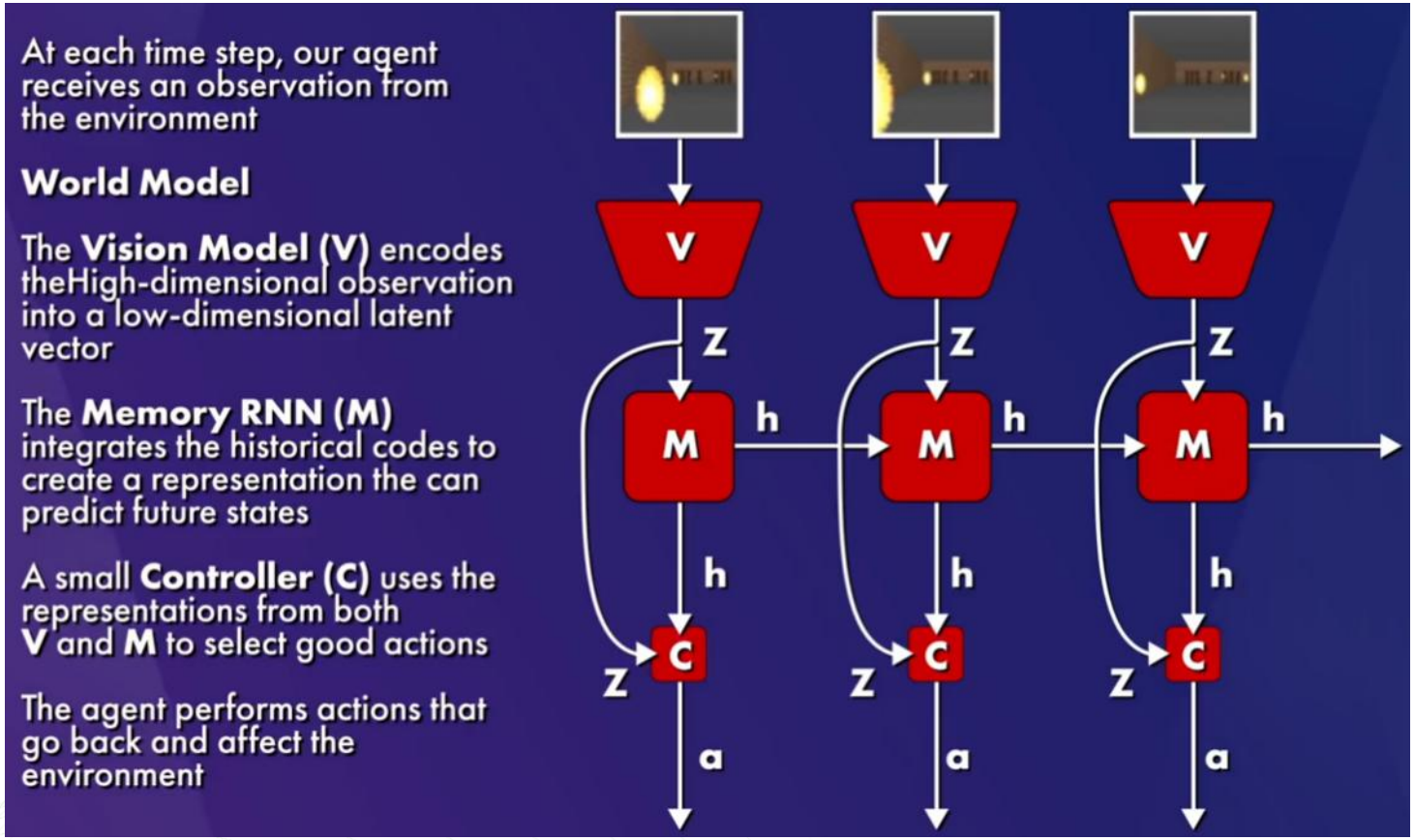
- AI could learn by training in its own simulated environment.
- Their model used a vision, memory and controller module to learn from its environment.
- Its vision model, a variational auto encoder, is generative, able to produce variations of the input data.

World Models vs. Language Models (世界模型 vs. 語言模型)

Aspect 面向	Large Language Models (LLMs) 大型語言模型	World Models 世界模型
Primary Data (主要資料來源)	Textual corpora (e.g., web text, books) 文字語料（如網頁文字、書籍）	Sensory data, simulations, and telemetry 感測資料、模擬與遙測數據
Architecture (架構)	Transformers with self-attention 具有自注意力機制的 Transformer	Hybrid architectures: encoders + latent dynamics (Transformers too!) 混合式架構：編碼器 + 潛在動態（也使用 Transformer！）
Objective (目標)	Predict next tokens 預測下一個語詞	Predict environment states; support decision-making 預測環境狀態，支援決策制定
Training Paradigm (訓練模式)	Self-supervised learning on text 文字的自我監督學習	Self-supervised or reinforcement learning 自我監督或強化學習
Applications (應用)	NLP tasks: translation, summarization, QA 自然語言任務：翻譯、摘要、問答	Robotics, control, simulation, model-based RL 機器人控制、模擬、基於模型的強化學習
Grounding (基礎對應)	Statistical, linguistic 統計與語言層面	Physical, causal 物理與因果層面

Chen

World Models - 世界模型



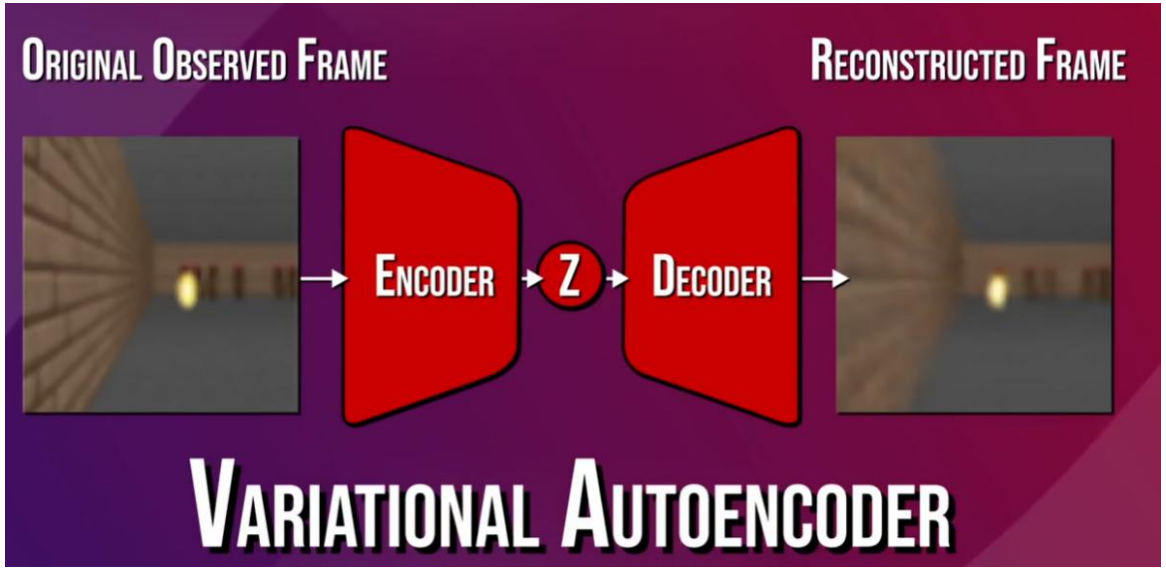
這一步能將複雜的感知資料壓縮成有用的特徵表示，以便後續的記憶與決策模組使用。

這個模組通常是一個 RNN（例如 GRU 或 LSTM），用於捕捉時間序列的動態關係。

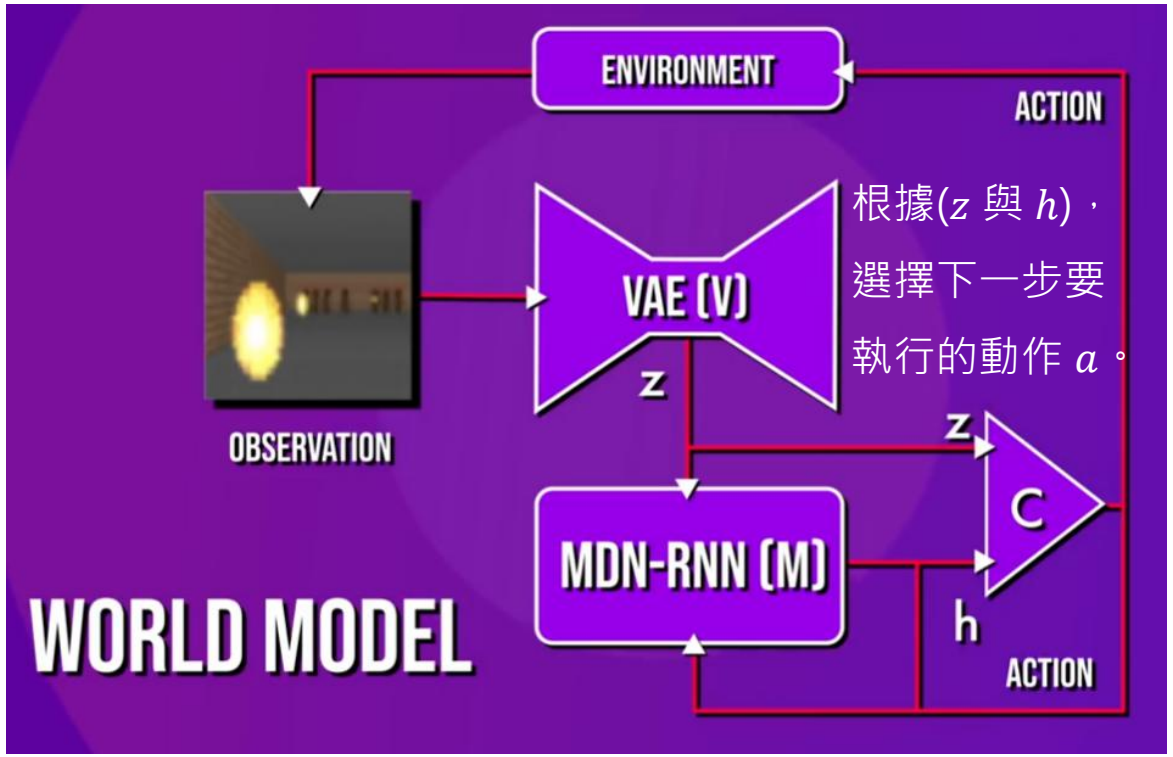
這是強化學習策略（policy）的核心部分。

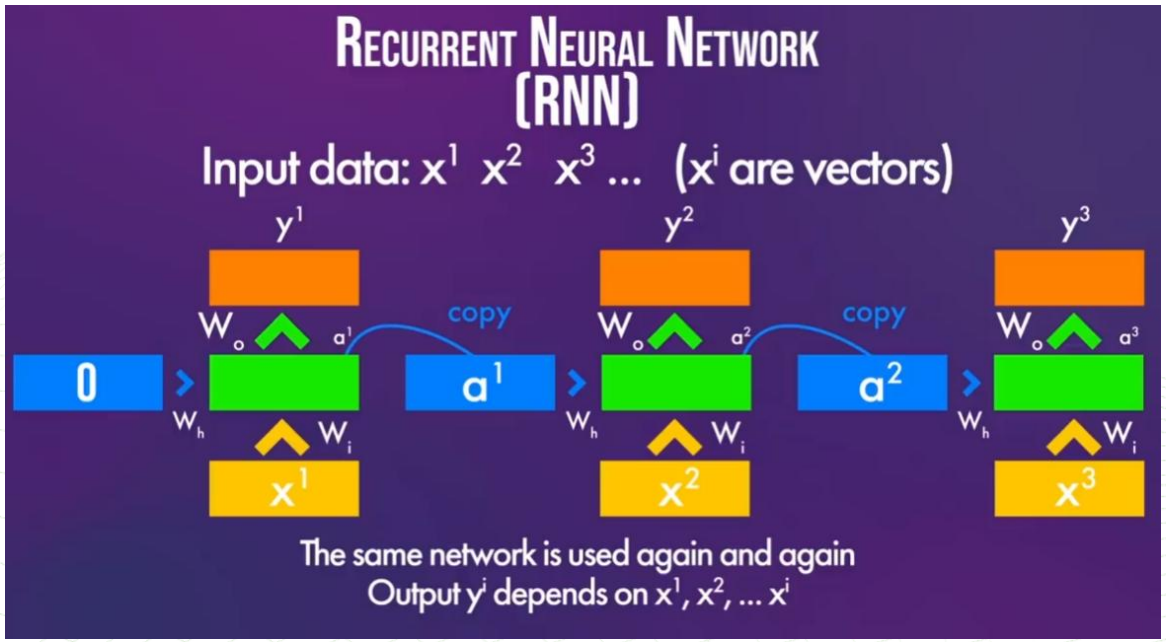
模組	輸入	輸出	功能
Vision Model (V)	觀測	潛在向量 z	特徵抽取
Memory RNN (M)	z, h_{t-1}	新的隱狀態 h_t	時間記憶與預測
Controller (C)	z, h_t	動作 a_t	決策控制

World Models - 世界模型



將高維度的觀測（例如影像畫面）**編碼**成低維度的潛在向量（latent vector）z





整合歷史的潛在代碼（latent codes）z，並建立一個能**預測未來狀態**的內部表示 h。RNN - Back Propagation through time

Jonathan Chen

https://www.youtube.com/watch?v=fkP7L2AeSVA

META : Yann LeCun: I'm right after all!

V-JEPA-2 World Model

Video Joint-Embedding Predictive Architecture

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

V-JEPA 2 is the first world model trained on video that achieves state-of-the-art visual understanding and prediction, enabling zero-shot robot control in new environments.

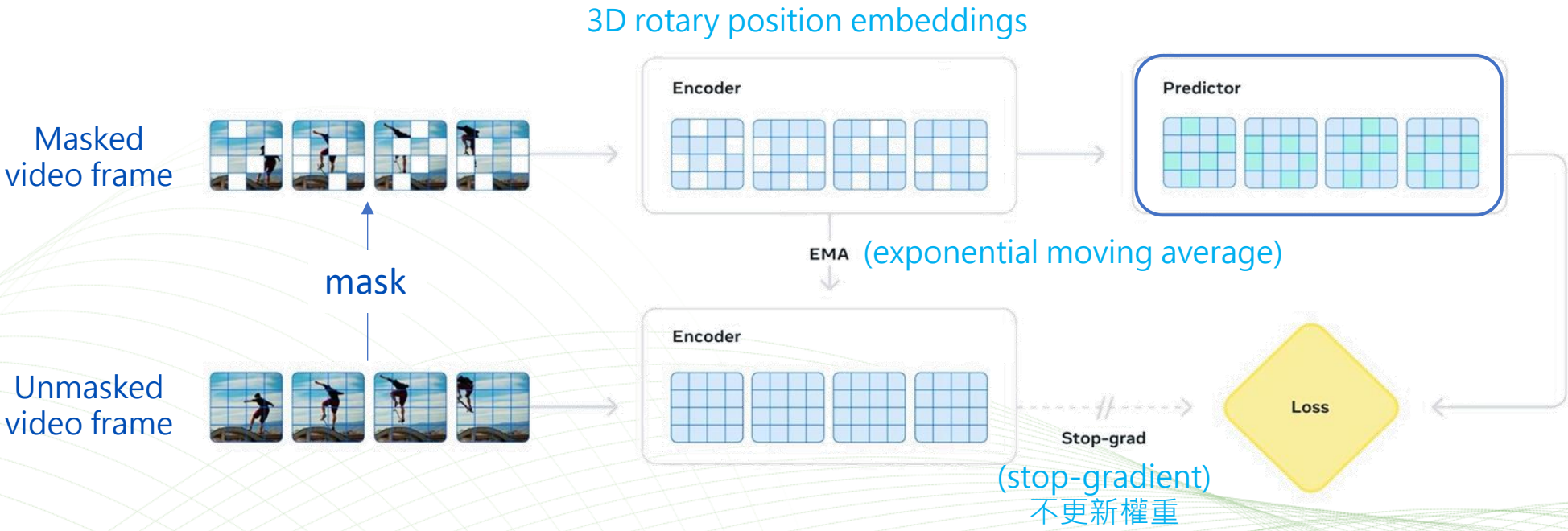
What Sets V-JEPA 2 Apart? (V-JEPA 2 特點)

V-JEPA 2, short for Video Joint Embedding Predictive Architecture 2, builds on massive data and sophisticated architecture. With 1.2 billion parameters and training on more than a million hours of videos and images, it achieves a remarkable understanding of physical dynamics and object interactions. Unlike earlier models, V-JEPA 2 unlocks **zero-shot robot planning**, enabling robots to tackle novel tasks and environments without prior exposure.

- **Self-supervised learning:** V-JEPA 2 learns directly from raw video, extracting knowledge about interactions and object behavior without the need for human-annotated labels.
- **Two-stage training:** The model undergoes broad visual pre-training, then is fine-tuned on robot action data. This enables accurate, action-conditioned predictions with minimal robot-specific information.
- **Joint-embedding architecture:** An encoder creates detailed world representations from video. A predictor then leverages these embeddings, along with action cues, to forecast future states.

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

V-JEPA learns by predicting the representation of masked parts of a video from the unmasked parts. It uses an encoder (a vision transformer with 3D rotary position embeddings) to extract video representations and a predictor to estimate masked patch representations. Training minimizes the L1 loss between the predicted and actual representations of masked patches, using stop-gradient and exponential moving average to avoid collapse.



The predictor's task is to use the visible parts (the context) to predict the representations of the hidden parts.

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

- 模型透過影片預訓練學習「時序資訊」與「物體交互關係」

- V-JEPA 是一種 self-supervised learning（自我監督學習）架構：是預測影片中被遮蔽或未來的影格特徵。

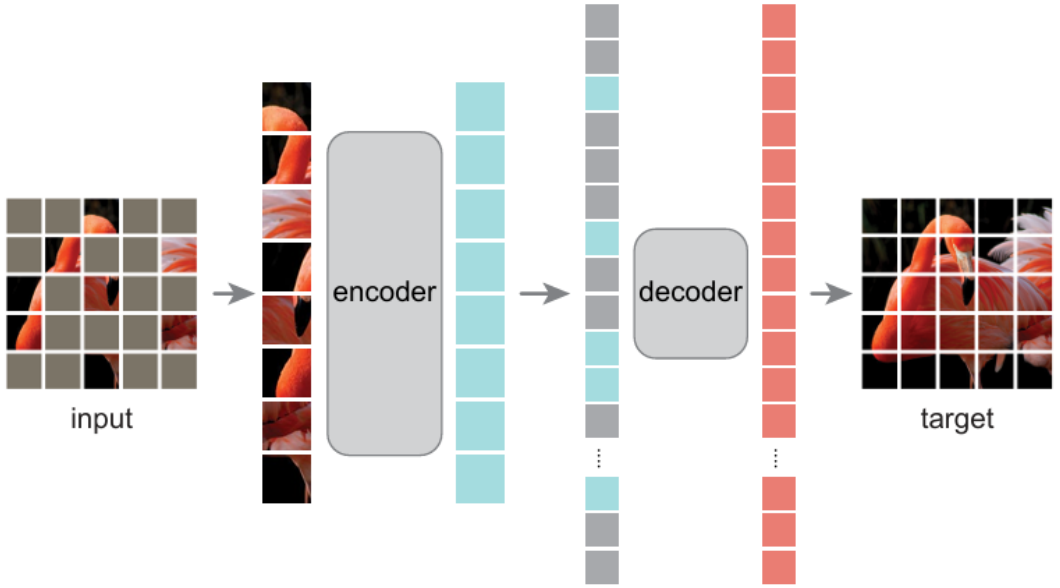
- 約 100 萬小時影片 + 100 萬張圖片
- 來源多樣（如動物、人類活動、物體互動等）
- 用來讓模型學到「視覺世界的基本概念與動態理解」

- 理解影片內容並回答問題
- 理解與預測動作
- 機器人操作與規劃

- 約 62 小時 的機器人資料
- 含有機器人的狀態與動作（例如手臂移動、抓取）
- 這部分用於後期「行動學習（Action Learning）」

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

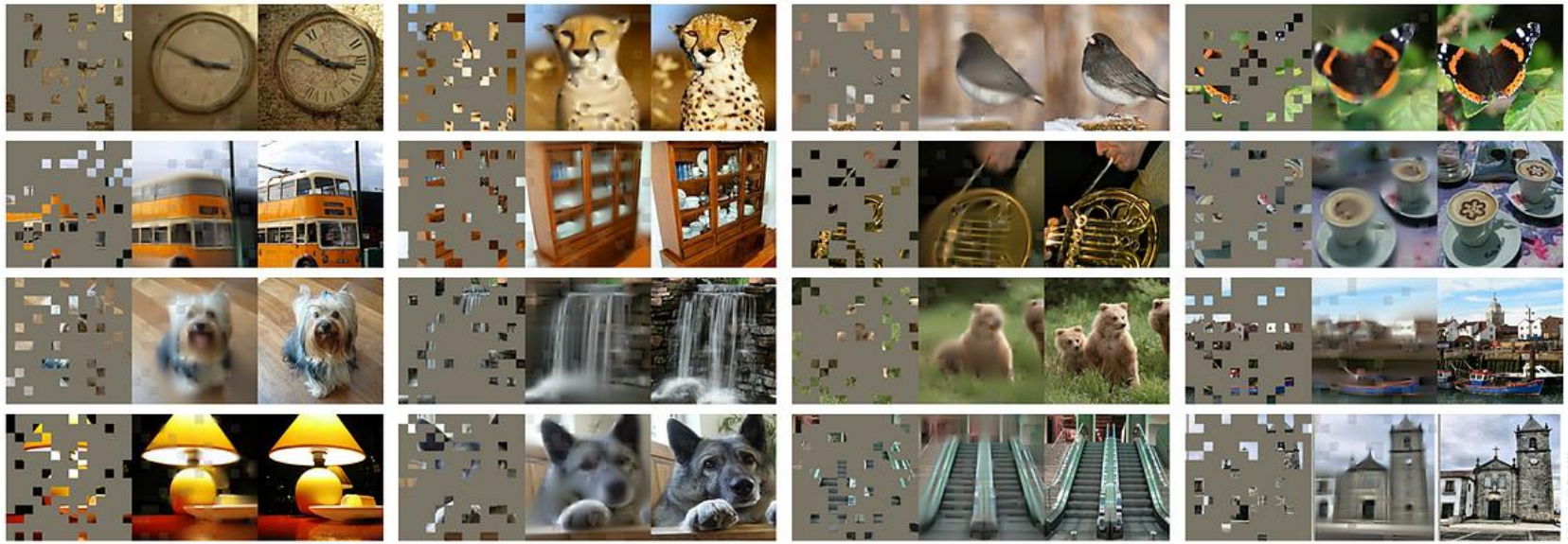
在預訓練期間，會遮蔽大量隨機影像片段（例如75%）。



遮罩標記在編碼器之後引入，完整的編碼patches與遮罩標記由一個小型解碼器處理，該解碼器以像素重建原始影像。

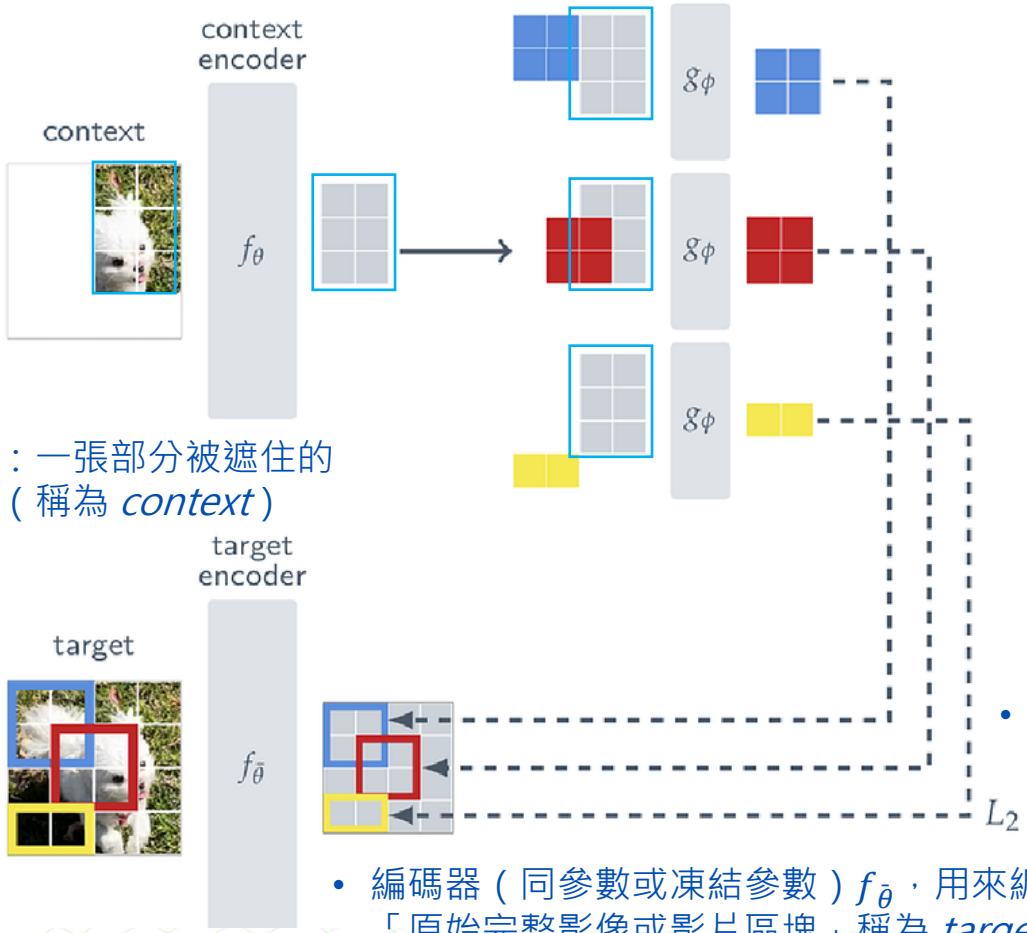
MAE (Masked AutoEncoder)

同I-JEPA，MAE以一定比例隨機 mask 掉圖片中的一些圖像塊(Patch)，然後重建他們的圖元值

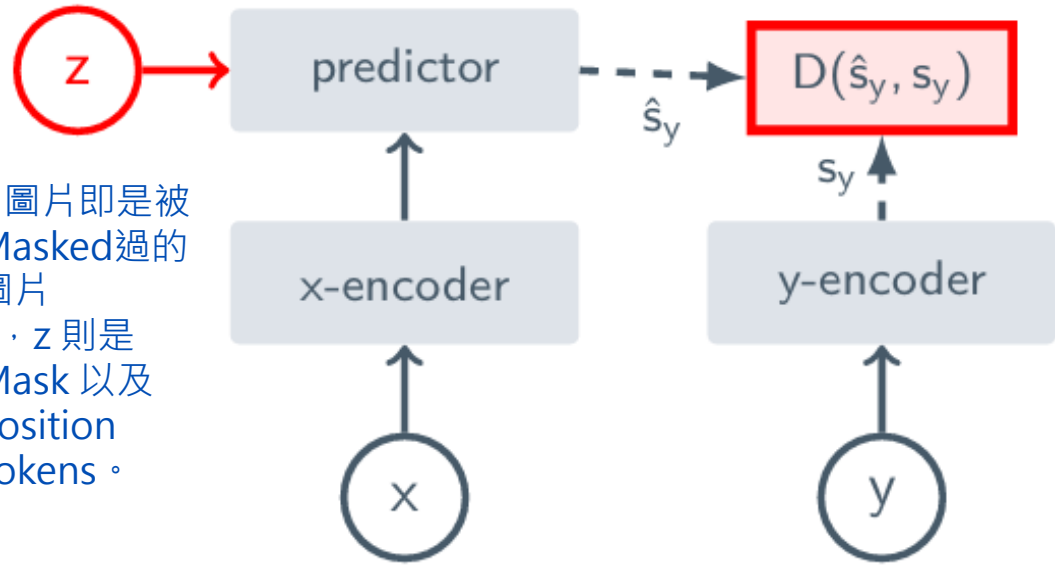


V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

- 經過 encoder f_θ 後，轉成特徵表示（feature embeddings）。
- 特徵代表模型對可見區域的理解
- 根據「上下文 (context)」去推測「被遮住的部分」

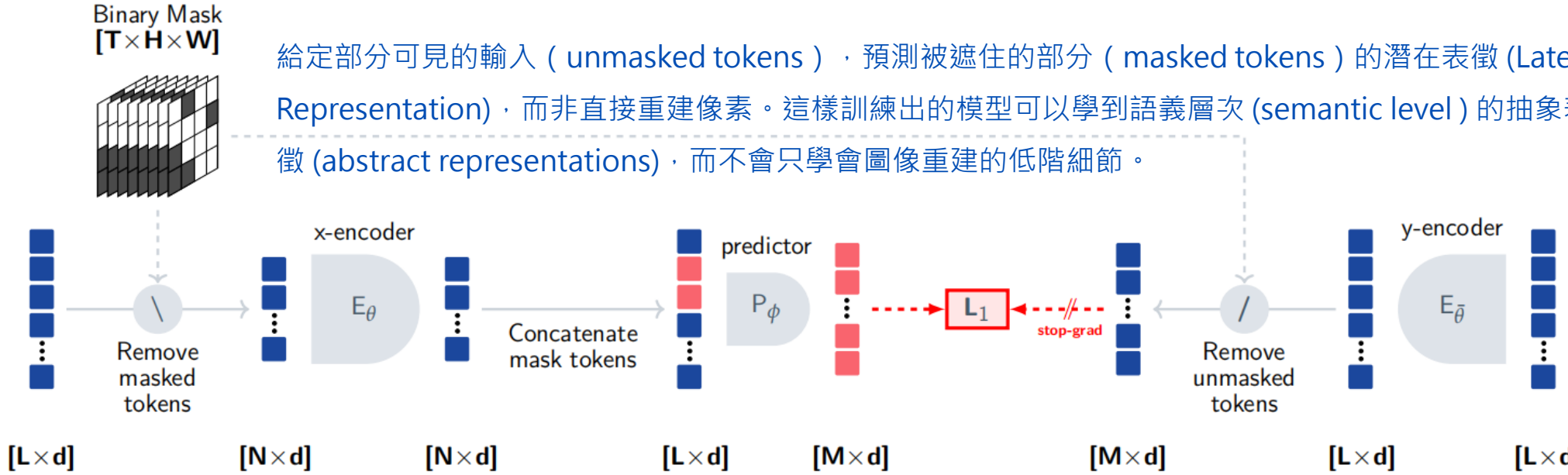


- 輸入：一張部分被遮住的圖片（稱為 *context*）
- 編碼器（同參數或凍結參數）$f_{\bar{\theta}}$，用來編碼「原始完整影像或影片區塊」稱為 *target*



- x 圖片即是被 Masked過的圖片
- y，z 則是 Mask 以及 Position Tokens。
- 模型比較「預測的特徵」與「真實特徵」的距離

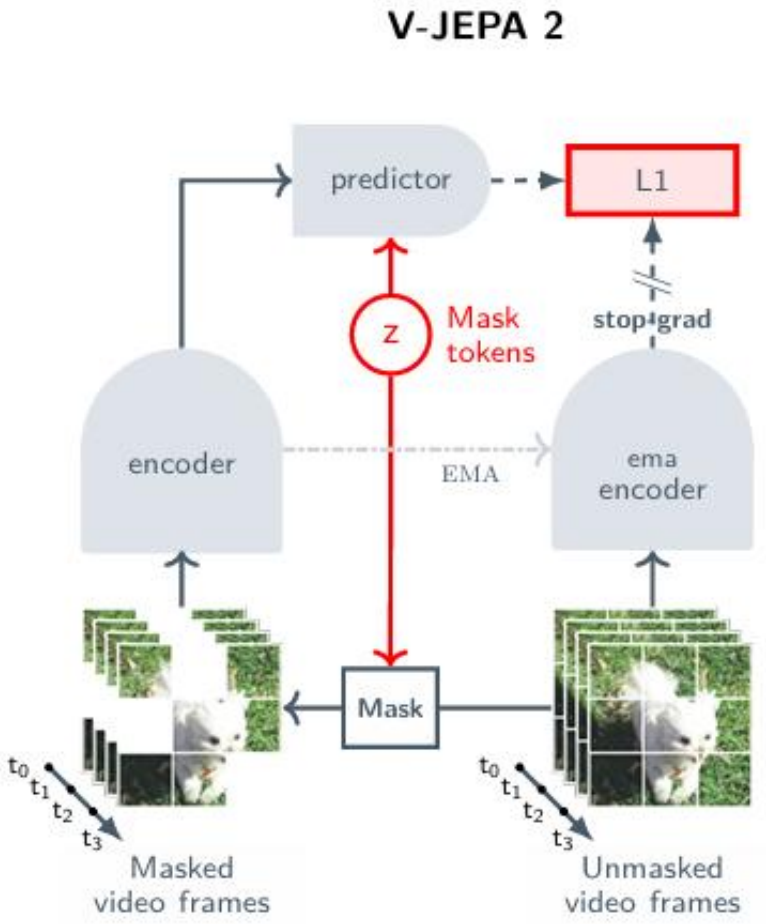
V-JEPA 2-AC (Video Joint-Embedding Predictive Architecture)



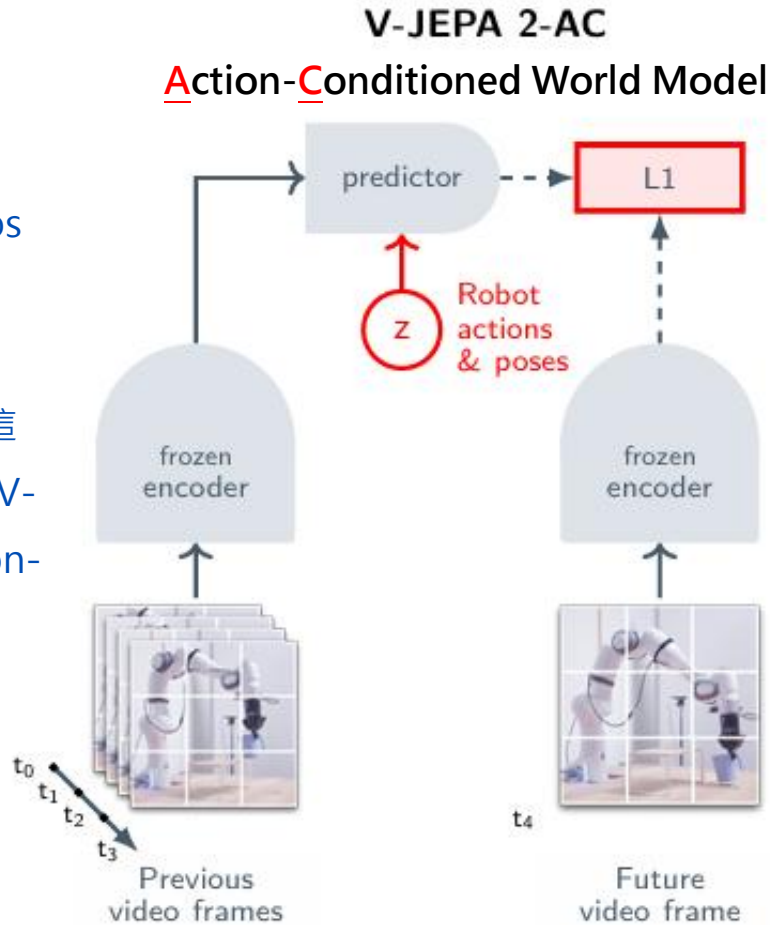
- **Binary Mask** [T × H × W] - 用來決定哪些影格、位置或patch會被masked，哪些會保留。黑色代表被mask掉的tokens，白色部分保留。
- **Remove masked tokens** → [L × d] → [N × d] -只留下未被mask的token
- **x-encoder** E_θ - 負責對未mask的輸入tokens編碼，得到上下文特徵 (context features)。輸出維度 [N × d]，其中 d 是embedding維度。
- **Concatenate mask tokens** → [L × d] - 在x-encoder輸出後，重新插入mask tokens(代表缺失的部分)。這樣恢復成原本的token長度 L。
- **predictor** P_ϕ - 一個小型網絡通常是MLP或transformer block，從已編碼的上下文中預測被遮蔽的tokens的embedding。輸出 [M × d]，其中 M 為mask掉的token數量。
- **y-encoder** E_θ - 是一個目標網絡 (target network)，通常透過 EMA (Exponential Moving Average) 從 E_θ 更新（這樣可以穩定訓練）它對完整輸入（包括masked部分）進行編碼，得到真實的目標embedding [L × d]。
- **Remove unmasked tokens** → [M × d] - 從y-encoder的輸出中，只保留被mask掉的部分（也就是模型需要預測的部分）。
- **損失函數** L_1 - 比較 predictor 的輸出（預測embedding）與 y-encoder 的對應部分（真實embedding）。

Jonathan Chen

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)



V-JEPA 2的架構，同 Nvidia Cosmos 一樣可以搭配 Robot Data做 Post-Training，這類 Model 則稱為V-JEPA 2-AC (Action-Conditioned)。

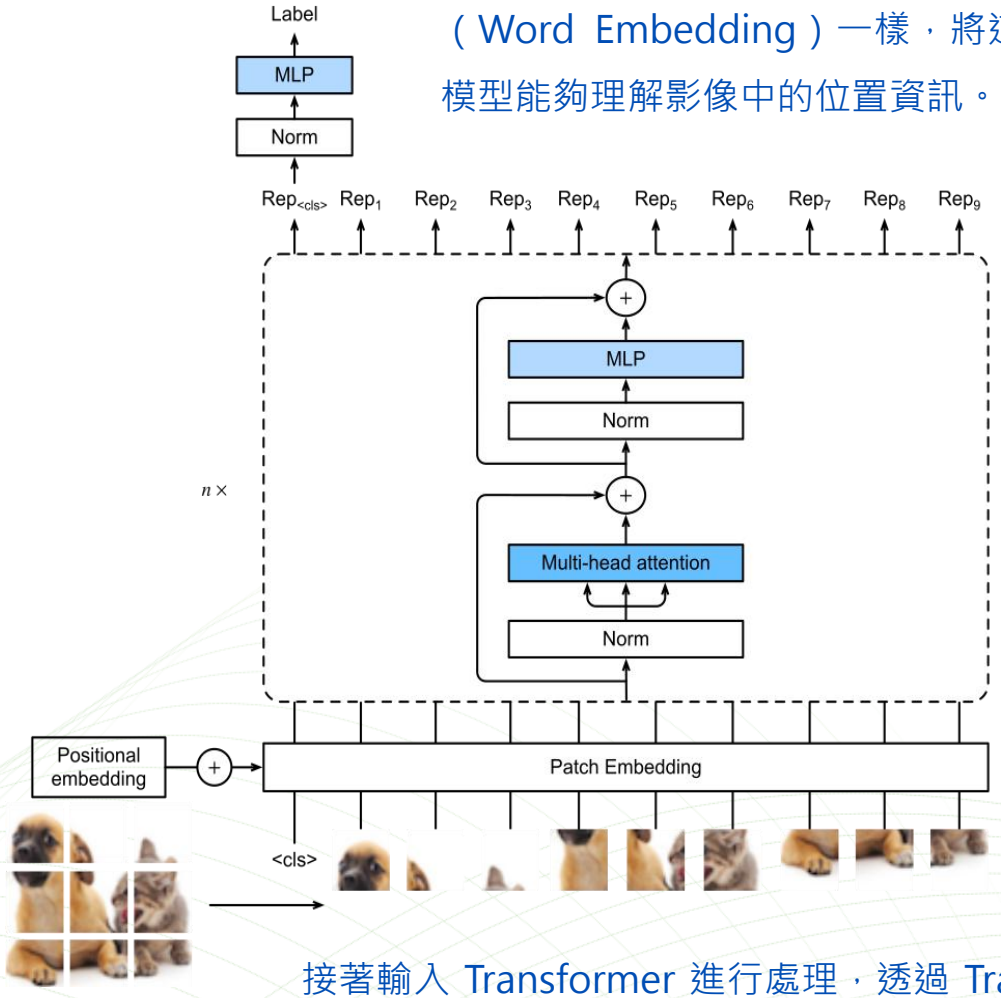


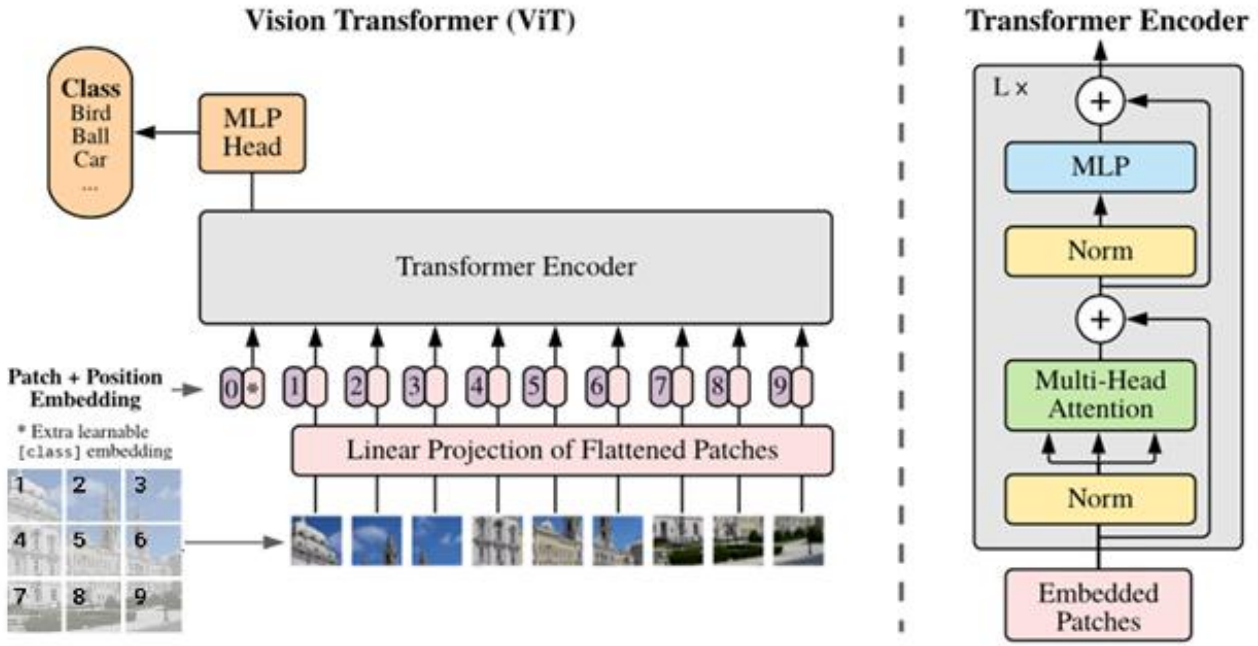
- V-JEPA 2 Architecture

  與V-JEPA相同，Target Encoder同樣是Context Encoder的EMA (Exponential Moving Average)。不同的地方在於，V-JEPA 2採用了 RoPE 的位置編碼，而非傳統的 sincos 位置編碼。

Vision Transformer ViT Architecture

ViT 的主要特點是將影像分割成固定大小的圖像區塊 （Patches），然後像自然語言處理（NLP）中的詞嵌入（Word Embedding）一樣，將這些圖像區塊展平成向量，並加上位置編碼（Positional Embedding），使模型能夠理解影像中的位置資訊。





接著輸入 Transformer 進行處理，透過 Transformer 的輸出結果，使用 MLP（多層感知機）進行分類，這作法類似於傳統 CNN 最後的全連接層（Fully Connected Layer）。

ViT (Vision Transformer)

Add & Norm 的主要目的是：

讓資訊在深層網路中更穩定地傳遞，避免梯度消失或爆炸，並加速收斂。

- Add (Residual Connection，殘差連接)
- Norm (Layer Normalization，層正規化)

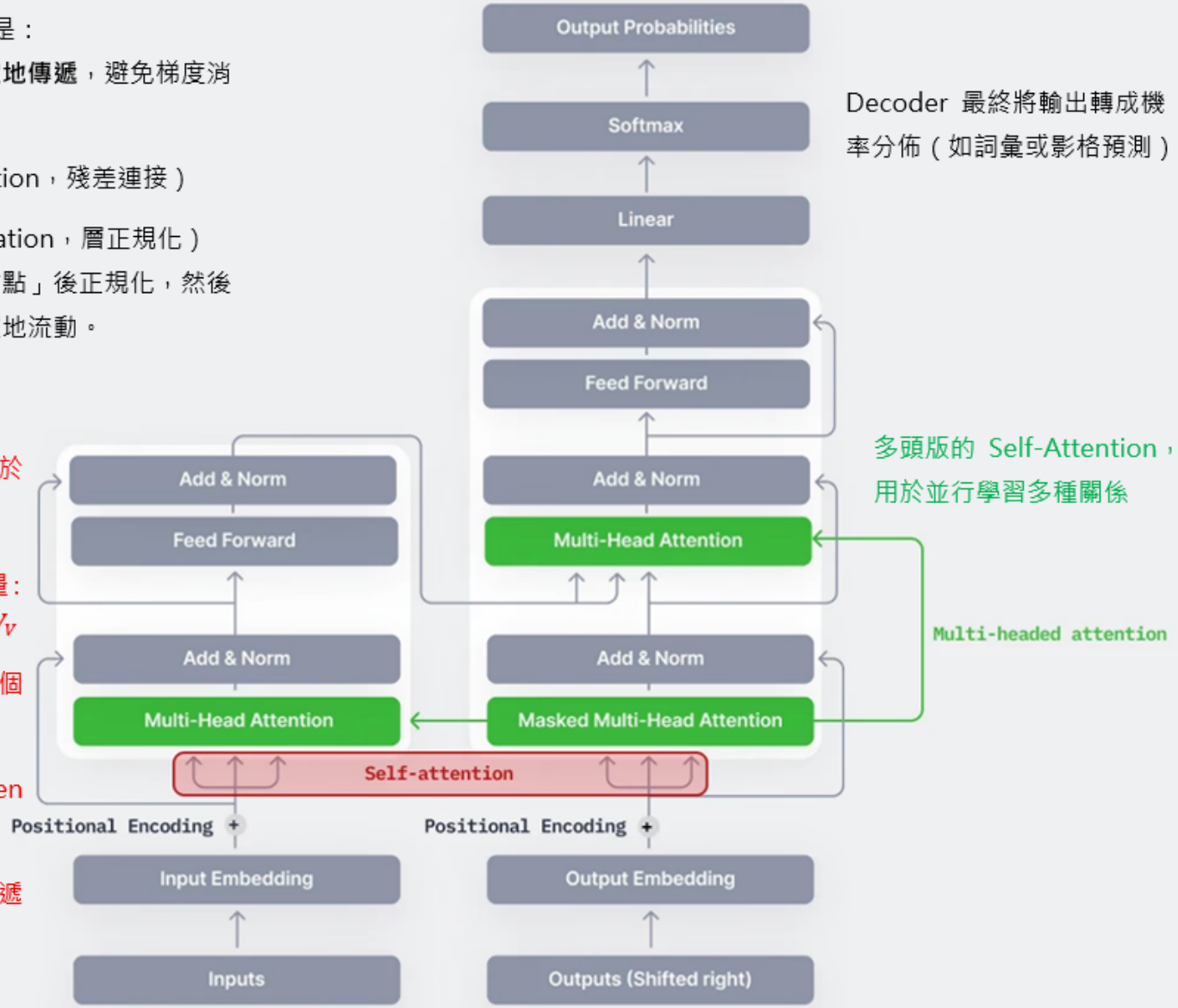
每層的輸入都被「修正一點點」後正規化，然後再餵給下一層，讓訊息穩定地流動。

Encoder 的內部注意力，用於捕捉輸入間的全局關聯。

對每個 token 生成三個向量：

$Q = XW_Q, K = XW_K, V = XW_V$

- **Query (Q)**: 代表目前這個 token 想要問的問題
- **Key (K)**: 代表其他 token 提供的資訊索引
- **Value(V)**: 代表實際要傳遞的資訊內容



Decoder 最終將輸出轉成機率分佈 (如詞彙或影格預測)

多頭版的 Self-Attention，用於並行學習多種關係

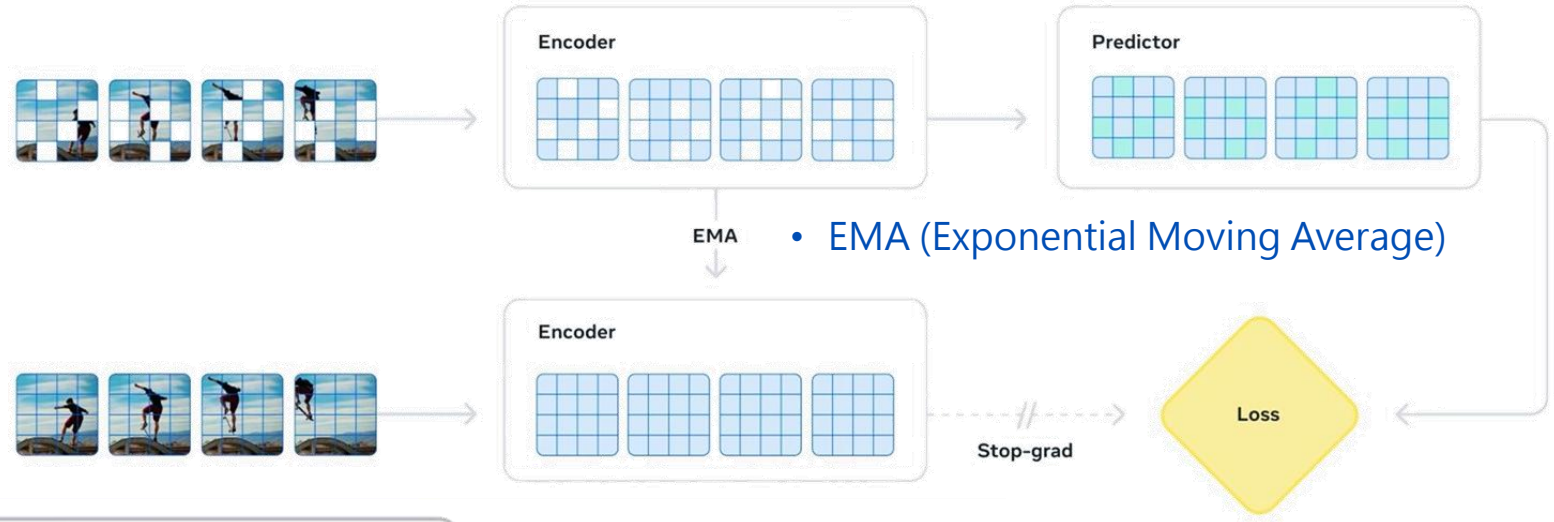
Multi-headed attention

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)

屬於 **self-supervised paradigm**：

- 它不需要「真實標籤」。
- 讓模型自己預測 **未來潛在表徵**（future latent representation）。
- 用 **Frozen Encoder + Stop Gradient** 生成穩定目標 Z_3。
- 用 **Loss = $||\hat{H}Z_3 - Z_3||$** 讓模型學會「理解與預測世界」。

→ 就是「讓模型預測世界的變化」，而不是靠人工標籤告訴它「對或錯」。



- EMA (Exponential Moving Average)

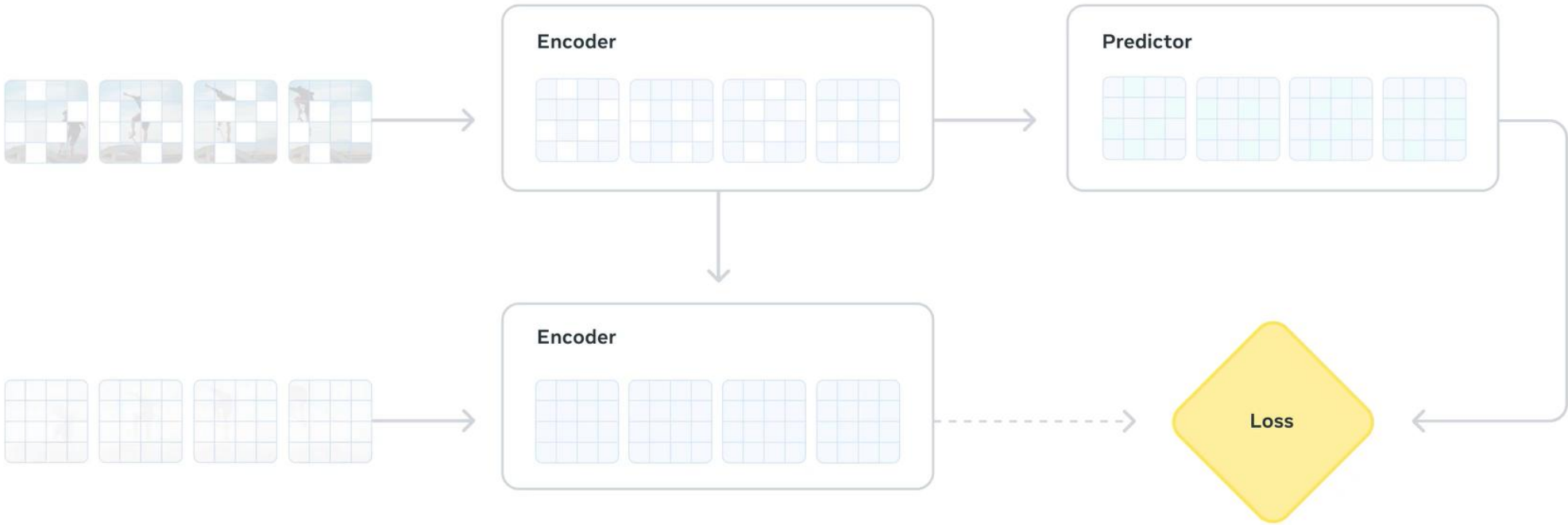
- 全新 AI 模型「V-JEPA 2」，能更精確理解三維空間與物體動態。

V-JEPA 2 就是為了模仿這種人類的「物理直覺」，並具備三大能力：

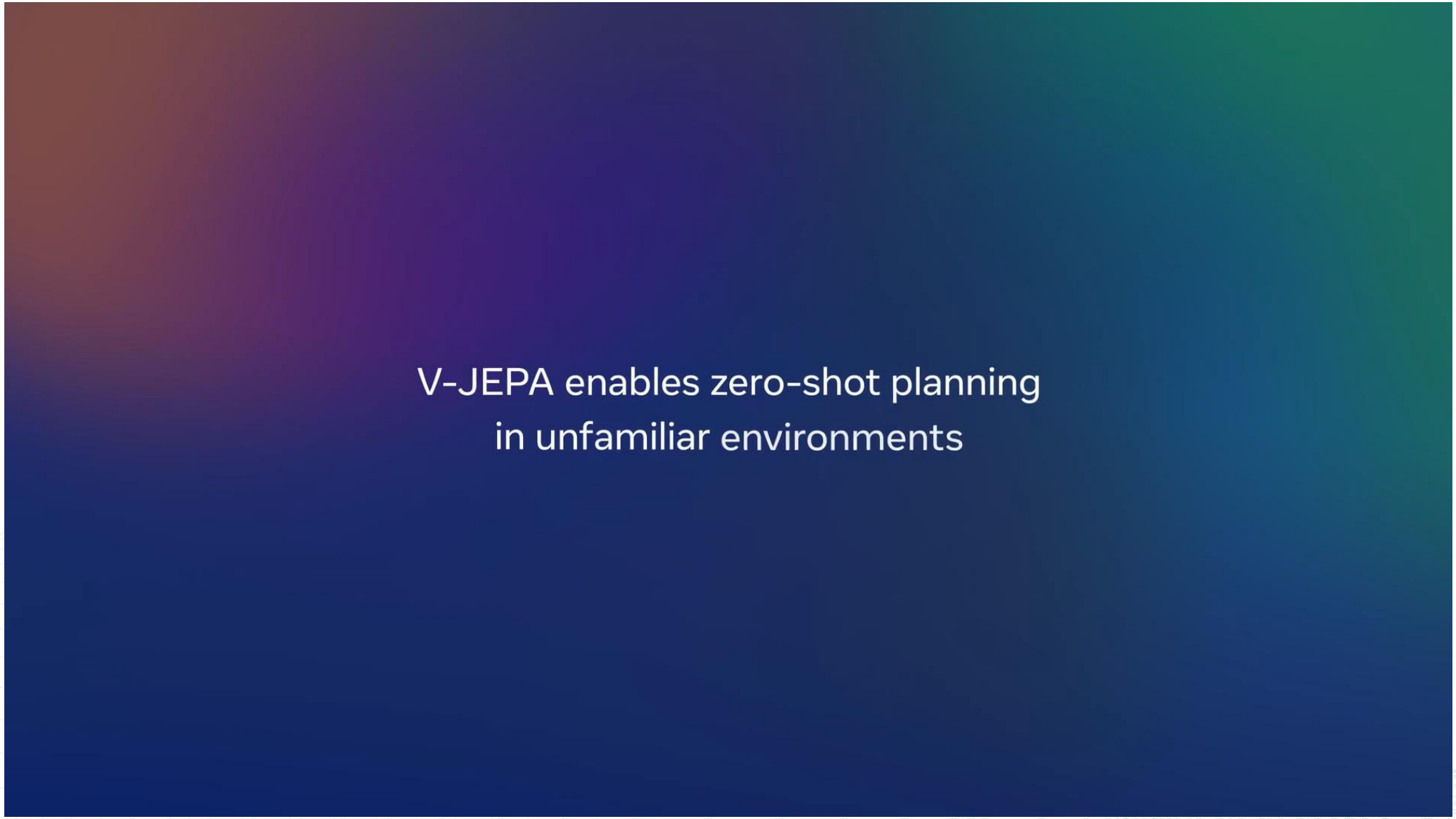
1. 理解（understanding）：掌握環境中物體與人之間的互動。
2. 預測（predicting）：預測特定動作可能產生的結果。
3. 規劃（planning）：設計出實現任務的行動流程。

Jonathan Chen

V-JEPA-2 (Video Joint-Embedding Predictive Architecture)



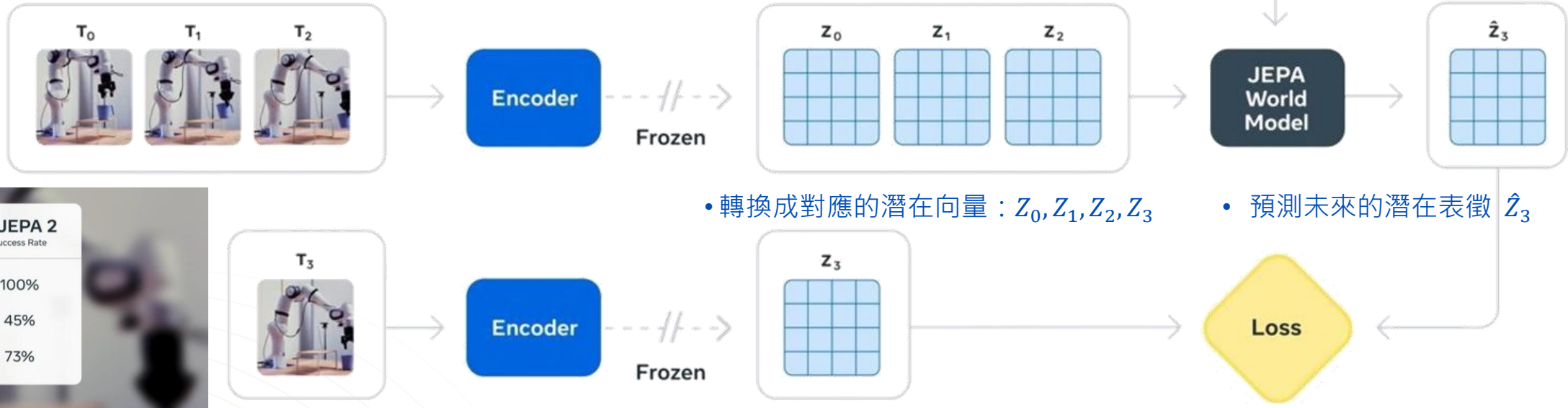
V-JEPA-2 (Video Joint-Embedding Predictive Architecture)



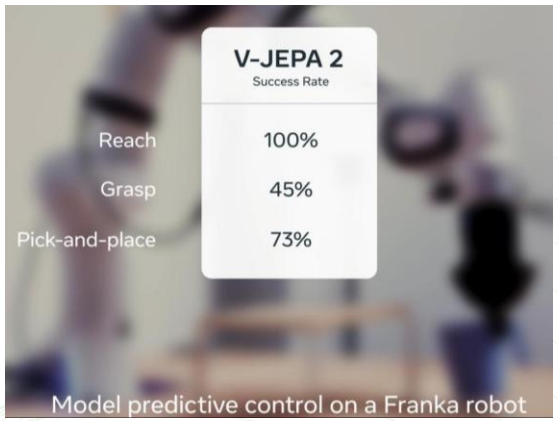
V-JEPA 2-AC (Video Joint-Embedding Predictive Architecture)

T_0, T_1, T_2, T_3 是機器人在不同時間點的觀察畫面（如相機影像）。

- 上排：T_0, T_1, T_2 → 代表觀察到的歷史狀態（過去的 frames）
- 下排：T_3 → 代表未來某一時間點的目標狀態



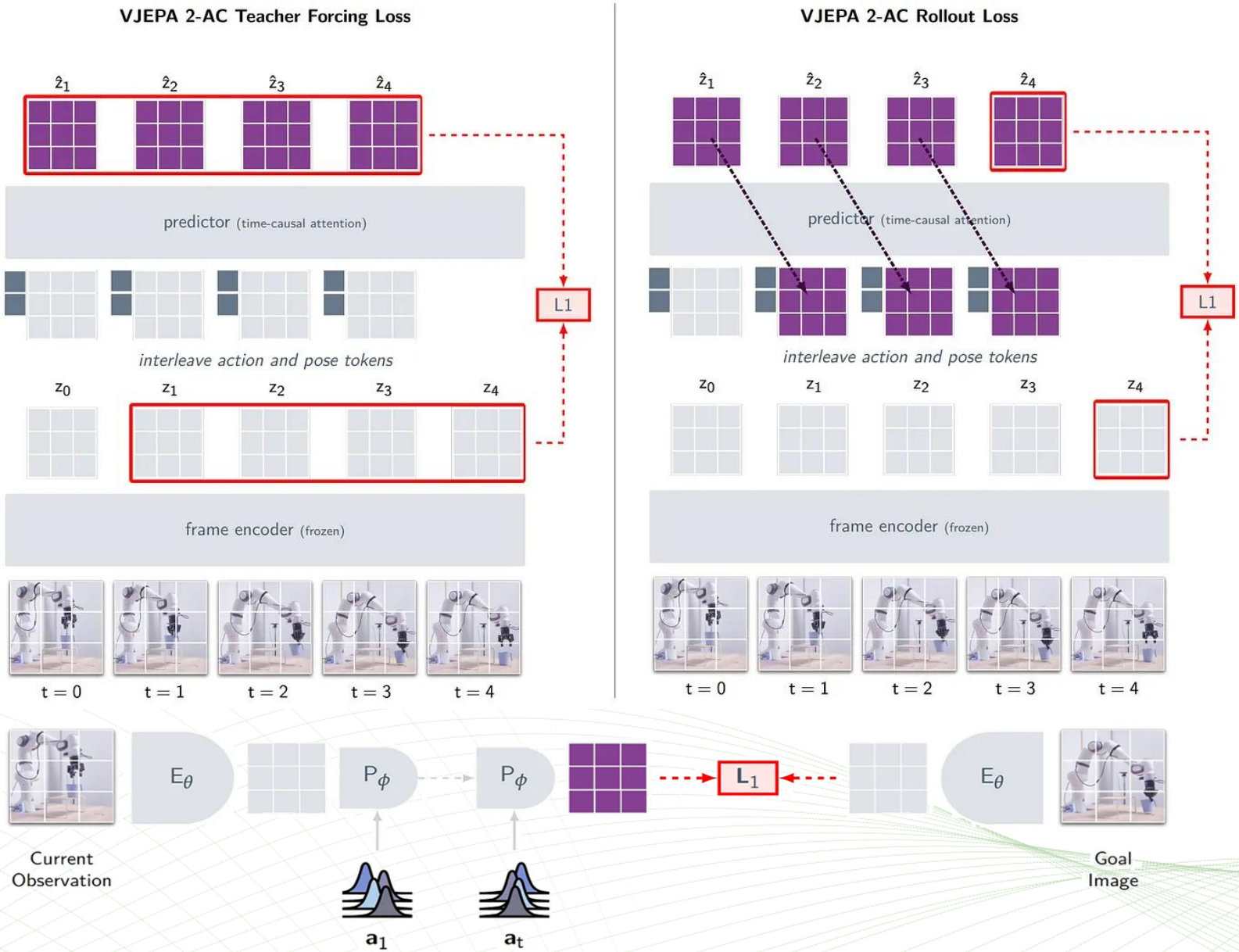
- 對應的機器人動作與姿勢 A_0, A_1, A_2
- 轉換成對應的潛在向量：Z_0, Z_1, Z_2, Z_3
- 預測未來的潛在表徵 $\hat{Z}_3$



- 「Frozen」表示這個影像特徵抽取模型不會在這階段更新權重（可能是事先訓練好的視覺模型，如 ViT）
- 輸出的預測 $\hat{Z}_3$ 會與實際的 Z_3 比較

- 模型可以部署在機械臂上，去執行物體操作類的任務，比如觸碰（Reach）- 100%、抓取（Grasp）- 45%、選擇和擺放物體（Pick-and-place）- 73%，而不需要大量的機器人資料或者針對性的任務訓練。

V-JEPA 2-AC (Video Joint-Embedding Predictive Architecture)



The rollout loss involves feeding the predictor's output back as input, allowing the model to be trained to predict several timesteps ahead.

Genie 3 World Model

Google Genie 3 世界模型

Genie : Generative Interactive Environments

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

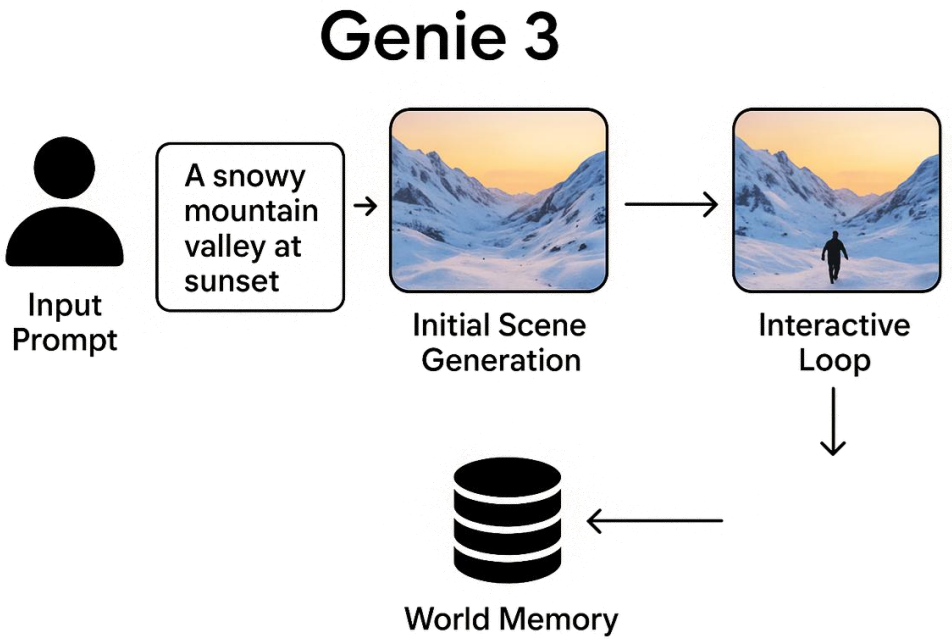
Genie 3: Creating dynamic worlds that you can navigate in real-time

Google DeepMind 2025/8/5 發表 Genie 3

Google DeepMind Genie 3

- DeepMind 想用 Genie 3 等世界模型，打造「可以學習、探索、行動」的 AI agents — 透過讓 agent 在模擬世界中學習，再把策略轉移 (transfer) 到真實世界。也就是說，這不只是創造世界，更是為「能動、能學習、能決策」的 AI 打底。

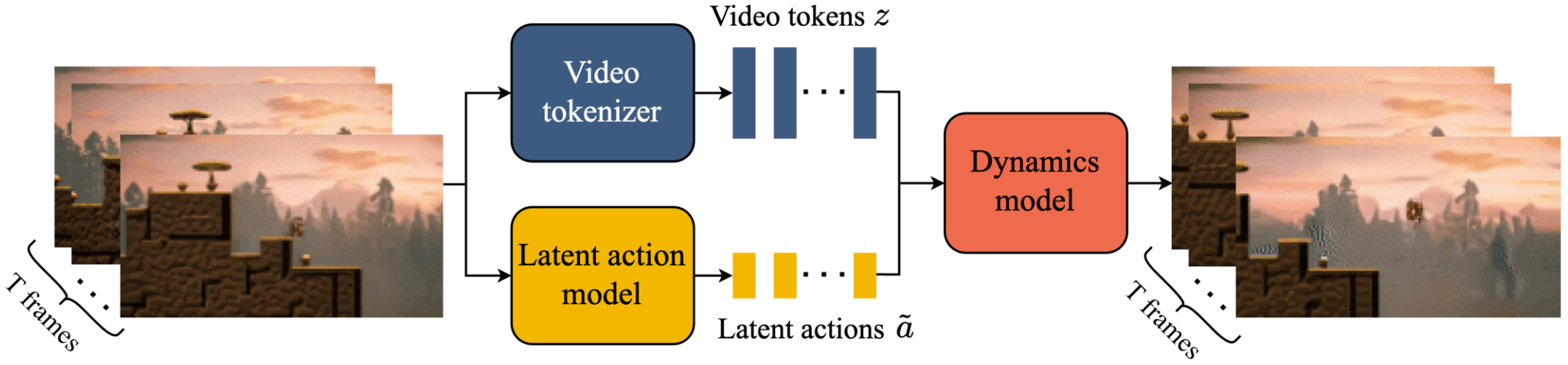
- 與傳統大型 language model (LLM) 不同，這類 world-model 給予 AI 一個「『物理 + 視覺 + 動作 + 因果』的世界」，讓 AI 有機會理解和預測世界的運作 (物體怎麼走、水怎麼流、光線怎麼變 ...) - 這對通用 AI (AGI) 來說，是非常重要的一塊。



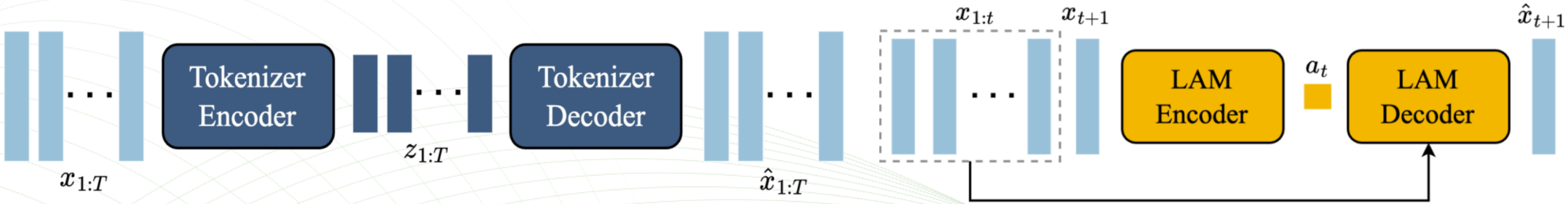
- **從靜態生成到「活世界」**：過去大部分生成式 AI (例如圖像生成、影片生成) 都只輸出一次性的靜態內容 (圖片/影片)；Genie 3 則把「世界 + 互動 + 連續性」整合起來 — 更接近「遊戲 + 模擬 + AI」。

- 同時，它對遊戲 / 模擬 / 內容創作 / 教育 / 訓練 (機器人、AI agent) 提供了一個「快速生成 + 高度可控 + 多樣性」的平台。對產業和研究都有潛力。

Genie : Generative Interactive Environments



Genie model training: Genie takes in T frames of video as input, tokenizes them into discrete tokens $\boldsymbol{z}$ via the video tokenizer, and infers the latent actions $\tilde{\boldsymbol{a}}$ between each frame with the latent action model. Both are then passed to the dynamics model to generate predictions for the next frames in an iterative manner.

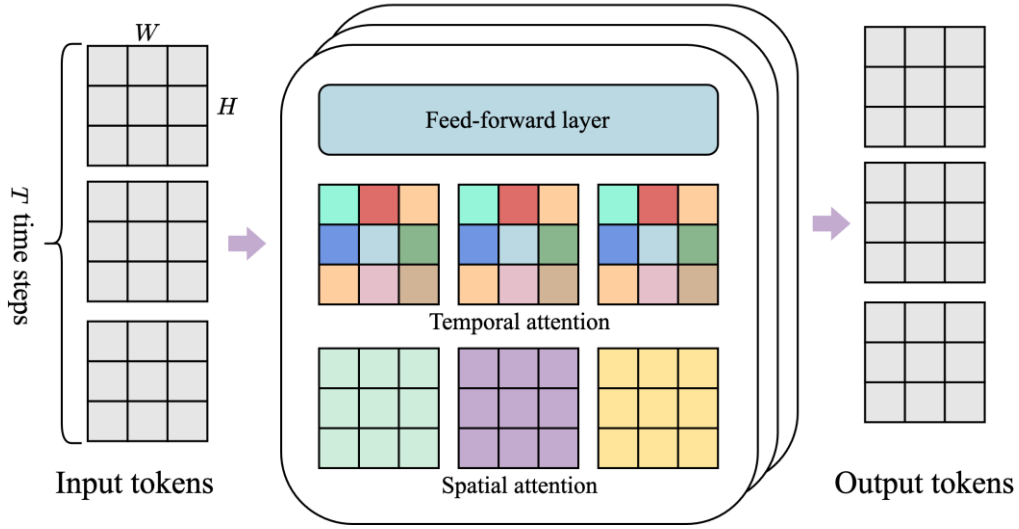


Video tokenizer: a VQ-VAE with ST-transformer.

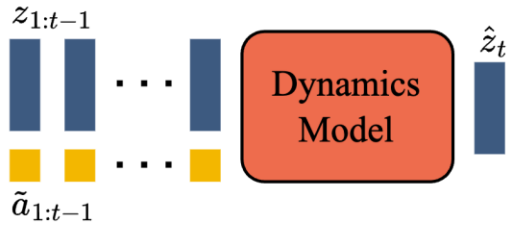
Latent action model: learns actions at unsupervised from unlabeled video frames.

Jonathan Chen

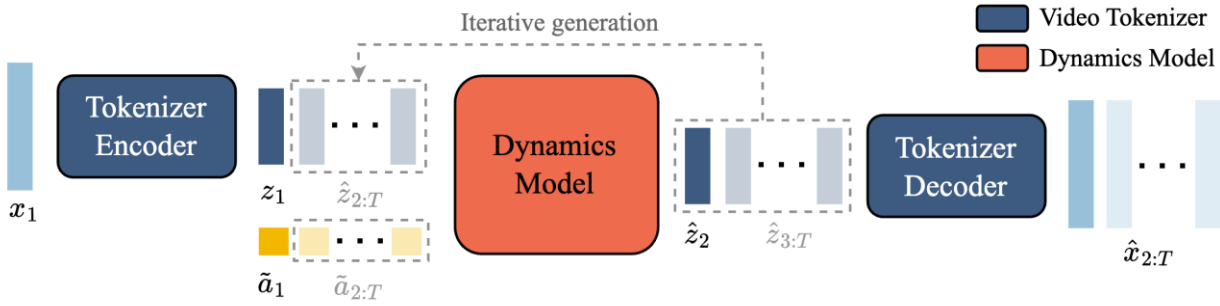
Genie : Generative Interactive Environments



ST-transformer architecture. The architecture is composed of L spatiotemporal blocks, each containing a spatial layer, temporal layer and feed-forward layer. Each color represents a single self-attention map, with the spatial layer attending over the $H \times W$ tokens from within a single time step, and temporal the same token from across the T time steps.



Dynamics model: takes in video tokens and action embeddings, and predicts future masked video tokens.



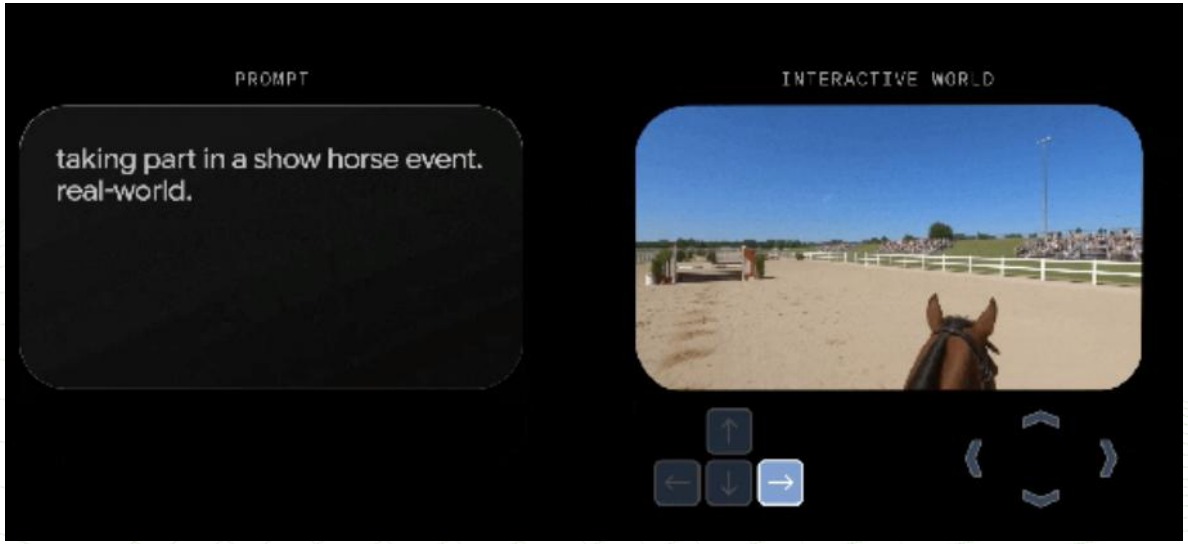
Genie Inference: the prompt frame is tokenized, combined with the latent action taken by the user, and passed to the dynamics model for iterative generation. The predicted frame tokens are then decoded back to image space via the tokenizer' s decoder.

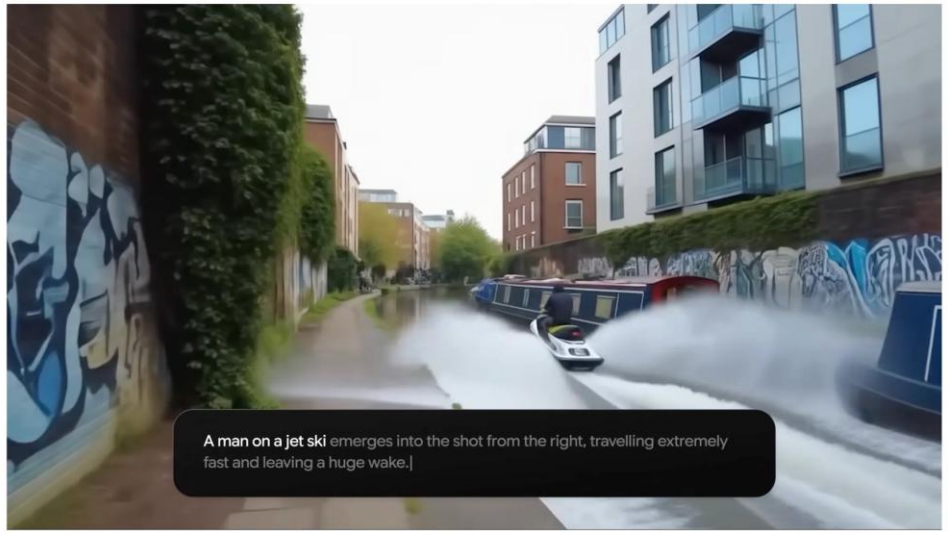
Jonathan Chen

Google DeepMind Genie 3

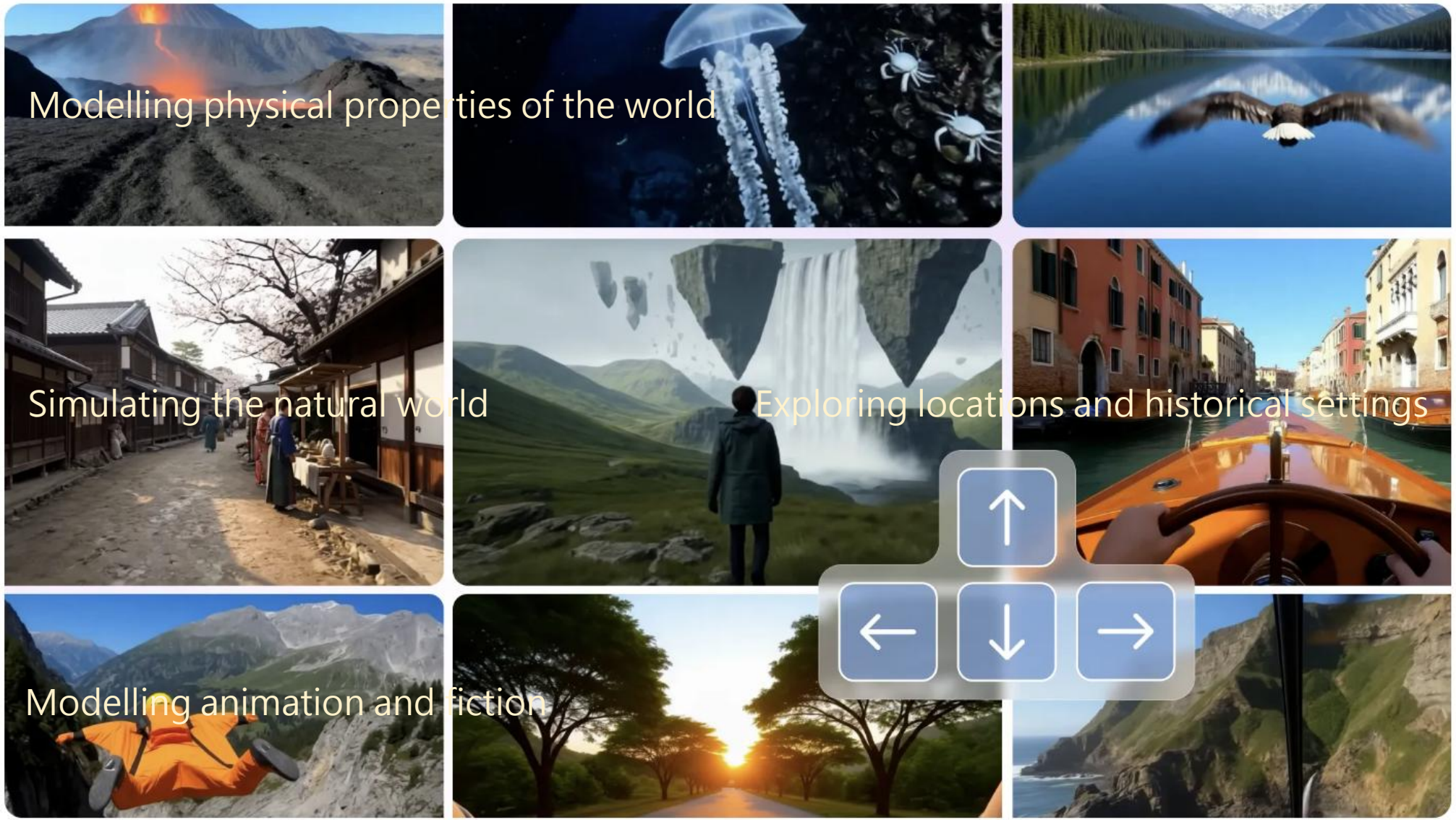
Genie 3支持的一些主要功能：

- **文本到3D世界生成：** 它可以將簡單的文本提示 prompt/context（例如，“一個機器人在街上行走”）轉換為具有基本移動控制的可玩3D環境。
- **可提示的世界事件：** 使用者可以通過鍵入新命令來動態更改環境（例如，在街道上添加雨水）。
- **視覺記憶：** Genie 3 可以記住環境中留下的物體，讓你稍後重新訪問它們，持續約一分鐘。
- **流暢且一致的視頻輸出：** 它可以保持 24 fps（幀每秒）在 720p 解析度下的視頻輸出，與 Genie 2 相比，參與度更高。



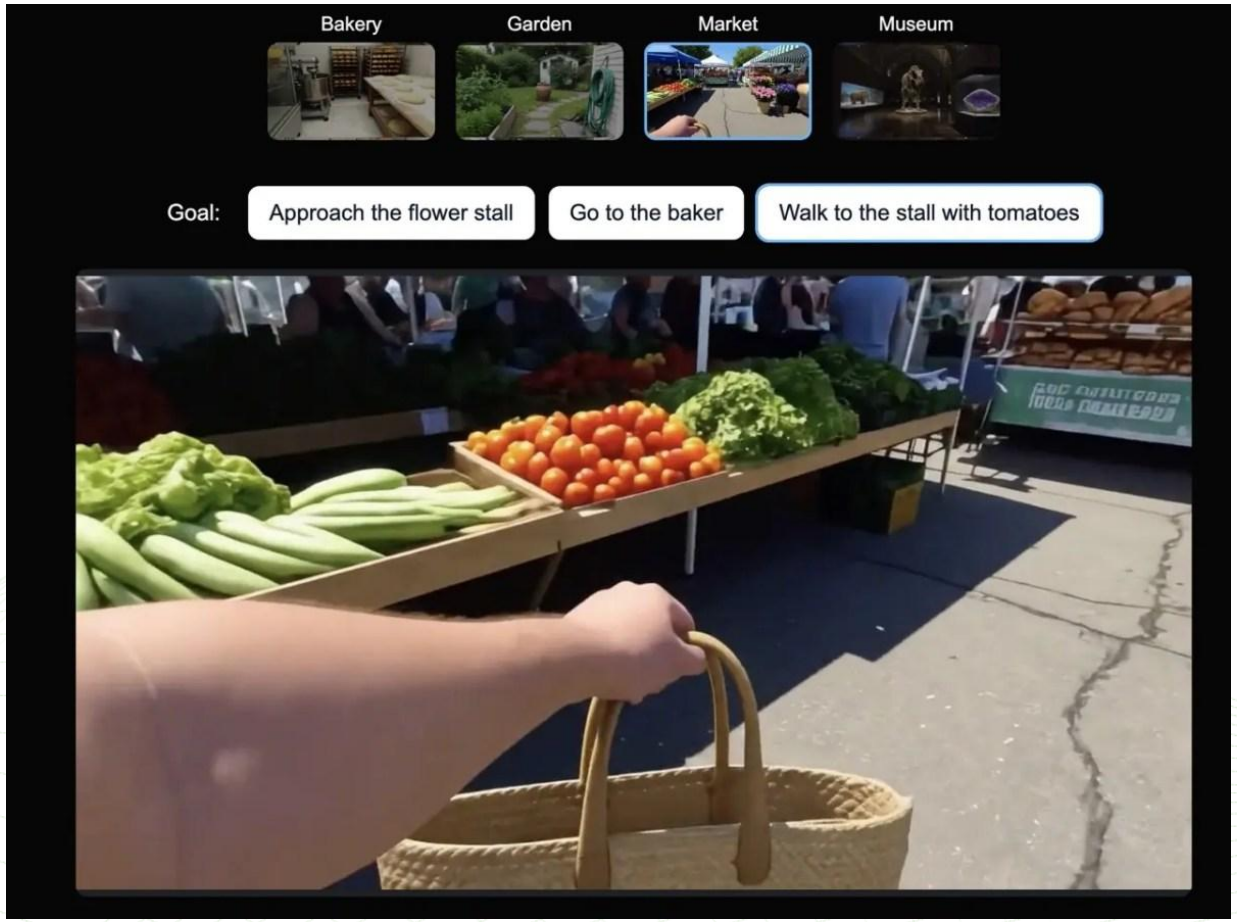


Google DeepMind Genie 3



Google DeepMind Genie 3

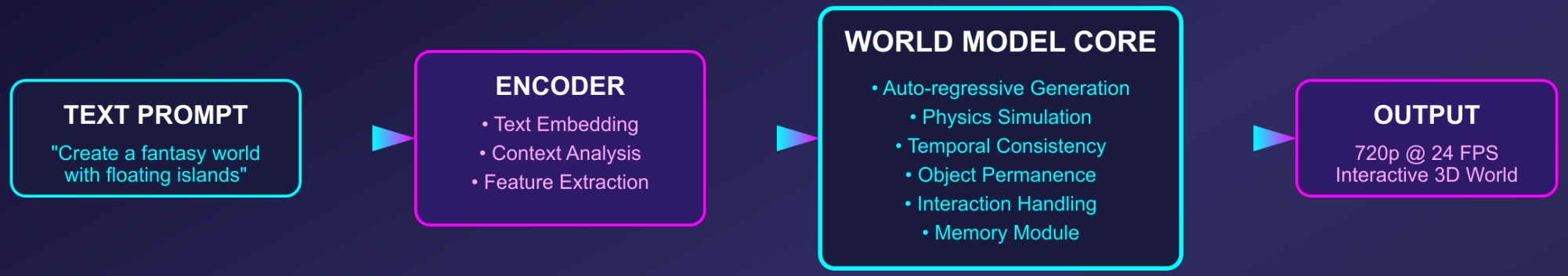
Google DeepMind 先前推出的 SIMA (Scalable Instructible Multiword Agent) 就是代表的例子。SIMA被設計成能在多種3D虛擬環境中接受指令，自己去觀察、規劃，並一步步完成任務。例如Demo中，可以指示它去市場買特定東西、在博物館找到某個展品，甚至完成多個步驟才能達成的複雜任務。(SIMA Agent 常受制於環境的一致性與可預測性)



Genie 3雖然已經能達到數分鐘的一致性，但仍舊還有許多不足：

- **執行行為有限**：現在的代理雖然能走動、觀察、與環境互動，但可直接執行的行為種類仍不多。許多世界事件需要透過指令觸發，而不是代理自己完成。
- **多代理互動不成熟**：讓多個AI在同一個世界中各自行動、相互影響，還是難題。要讓它們像真實人群那樣同時存在並互動，對世界模型來說仍是高難度的挑戰。
- **持續時間受限**：跨分鐘一致性已經是重大進步，但目前仍無法支撐數小時甚至數天的連續任務。對需要長期策略規劃的AI來說，這是一道天花板。
- **真實場景還原度有限**：即使能生成博物館或市場，這些空間與真實世界的地理與細節並不完全對應，因此在需要精準模擬的任務中會有落差。

Genie 3 Architecture Overview



Technical Specifications

Resolution:	1280 x 720 pixels (720p)
Frame Rate:	24 frames per second
Duration:	Multiple minutes (2-5 min typical)
Interaction:	Real-time user control
Memory:	Persistent object tracking
Prompting:	Dynamic world event control

Key Innovations

- ✓ First real-time interactive world model
- ✓ Emergent physics understanding
- ✓ Self-learned object permanence
- ✓ Consistent multi-minute generation
- ✓ Supports diverse world types
- ✓ Foundation for AGI agent training

Google DeepMind Genie 3

Genie 3 is an **11-billion-parameter auto-regressive transformer**.

Genie uses several clever techniques:

- **Multi-Scale Attention**: It pays attention to both the fine details in front of you (local consistency 局部一致性) and the overall structure of the world (global coherence 全域一致性), so a building in the distance doesn't suddenly change as you get closer.
- **Learned Physics**: Genie wasn't taught physics equations. Instead, it learned how objects should behave by watching over **200,000 hours of videos**, including gameplay and real-world footage. It figured out gravity, motion, and object interactions on its own. This is called **emergent physics**.
- **Smart Memory**: It has a multi-layered memory system to keep track of everything. Short-term memory handles immediate actions, while long-term memory ensures the world remains stable throughout your entire session.
- **Real-Time Speed**: Running on Google's powerful **TPU v5 infrastructure**, Genie can generate each new frame in just over 41 milliseconds, fast enough to deliver a smooth 24 frames per second (FPS) experience.

Genie 3 標誌著人工智慧歷史上的關鍵時刻。它**將 AI 從被動的內容產生者轉變為主動創造體驗的存在**。它讓建構虛擬世界的能力更加民主化，將創造的力量交到任何有想法的人手中。

https://www.youtube.com/watch?v=UslQB4LUuel

Introducing Marble by World Labs

Marble World Labs

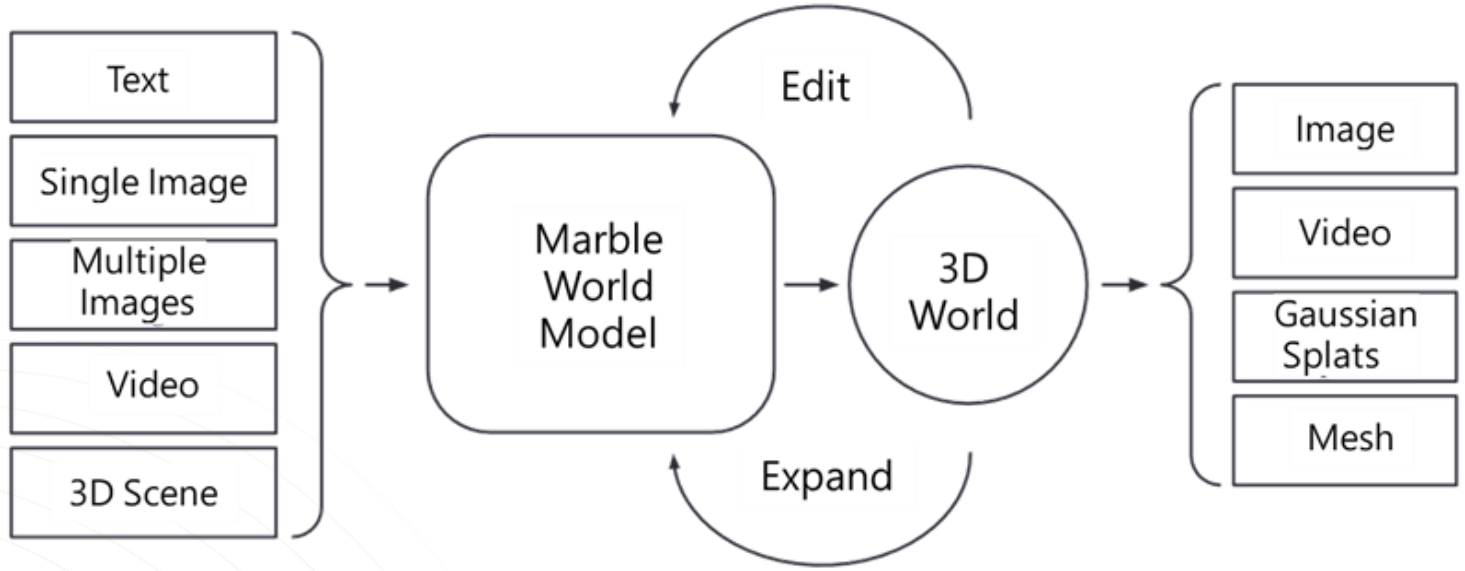
Marble 世界模型

Marble Highlights

Marble World Labs

Marble is the first of its kind - a next-generation world model making strides toward this vision. It can now create 3D worlds from a wide variety of input types, and lets users iteratively edit or expand worlds.

- Marble can create a full 3D world from a single image or a short text prompt.

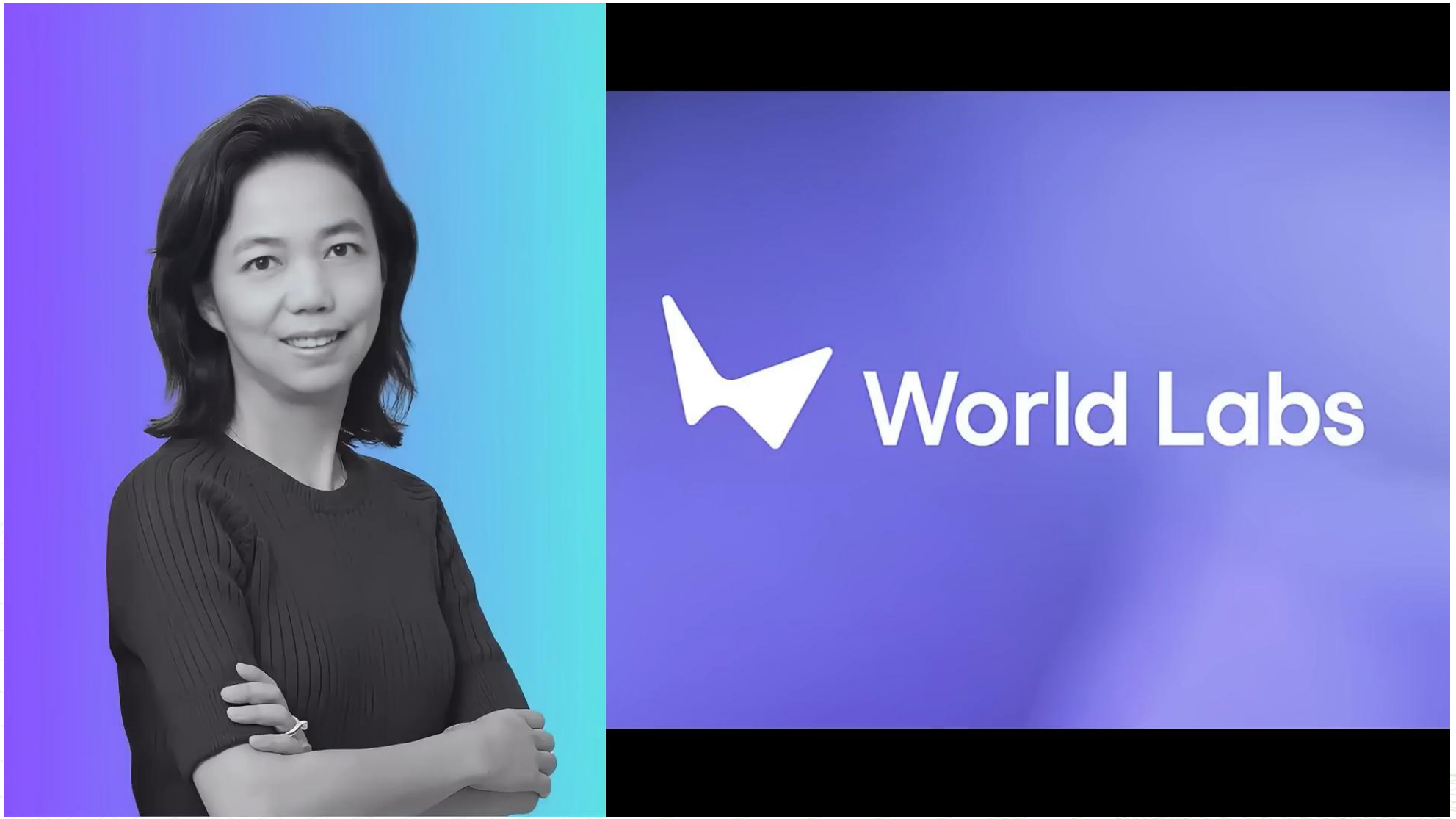


- After a world has been generated, Marble allows one - step expansion to make it larger.

- Marble can accept different prompt images (multi-image prompting) for different parts of the world, stitching them together into a unified 3D world.
- Marble includes AI-native world editing tools. Edits can be small and local: remove an object, touch up an area. They can also be more drastic: swap objects, change the visual style, or re-structure large parts of the world.

Marble World Labs

https://www.youtube.com/watch?v=UFyousBeB_Q&t=278s



Marble World Labs

3D Effects

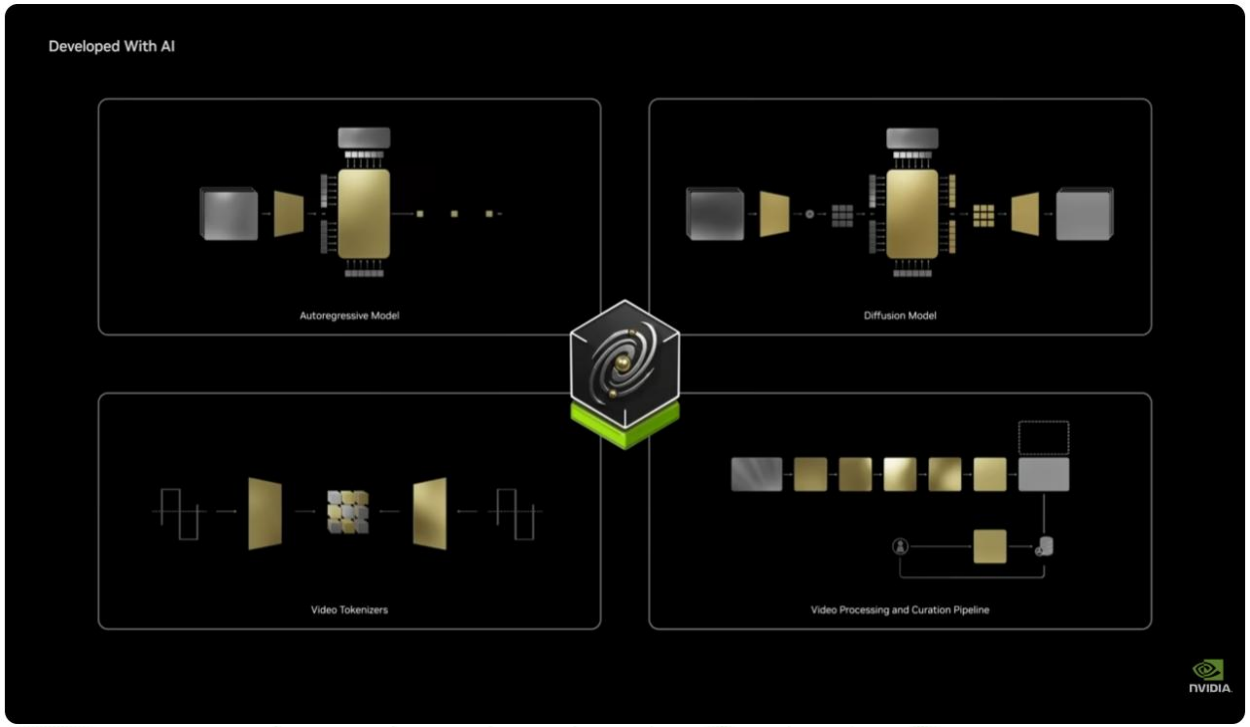
Most generative models predict pixels. Predicting a 3D scene instead has many benefits:

- **Persistent Reality**: Once a world is generated, it's there to stay. The scene won't change behind your back if you look away and come back. (3D 場景具有一致性與穩定性)
- **Real-Time Control**: After generating a scene, you can move around it in real-time. You can linger on the details of a flower, or peek around a corner to see what is revealed. (可以即時在場景中自由移動，近看細節或繞到角落觀察新內容，模型不只是靜態輸出，而能支援互動式探索。)
- **Correct Geometry**: Our generated worlds obey basic physical rules of 3D geometry. They have a sense of solidity and depth that contrasts with the dream-like nature of some AI-generated video. (具備真實的立體感與深度)




AI 教母 李飛飛 – Spatial Intelligence

https://www.youtube.com/watch?v=y8NtMZ7VGmU&t=59s



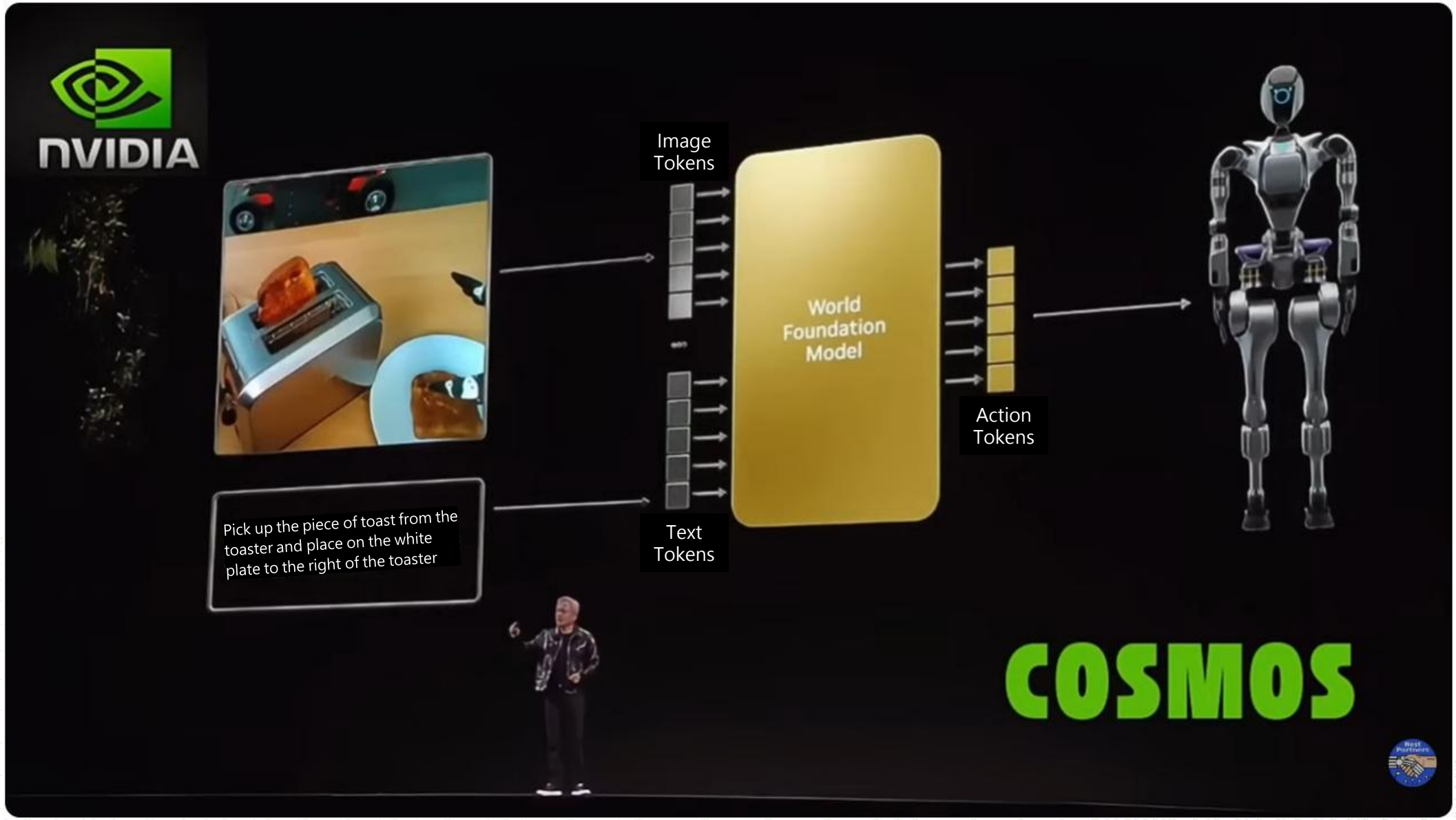
NVidia Cosmos (WFM)

Cosmos 世界基礎模型

https://www.youtube.com/watch?v=9Uch931cDx8

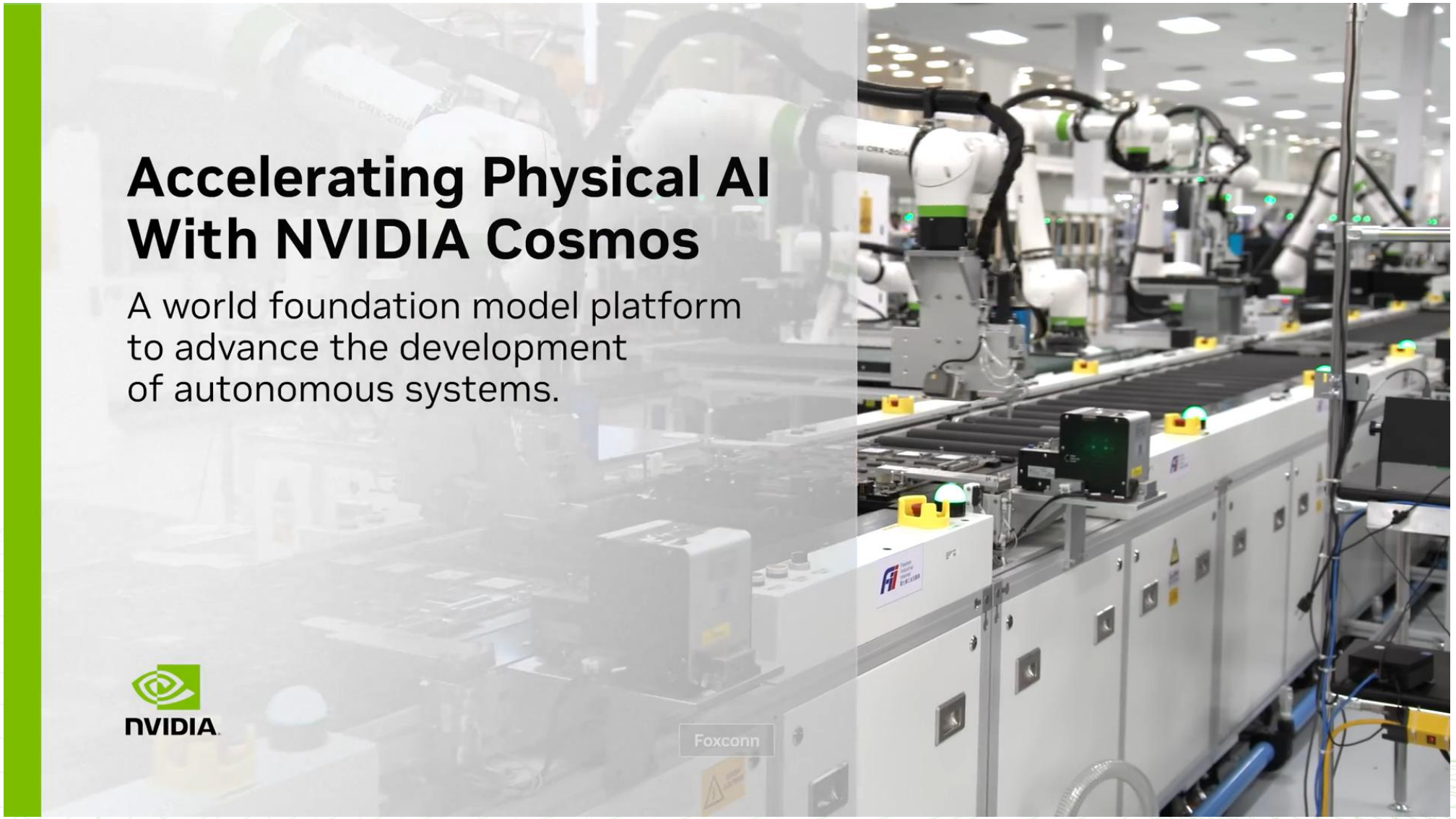
NVIDIA Cosmos: A World Foundation Model Platform for Physical AI

NVidia Cosmos (WFM)



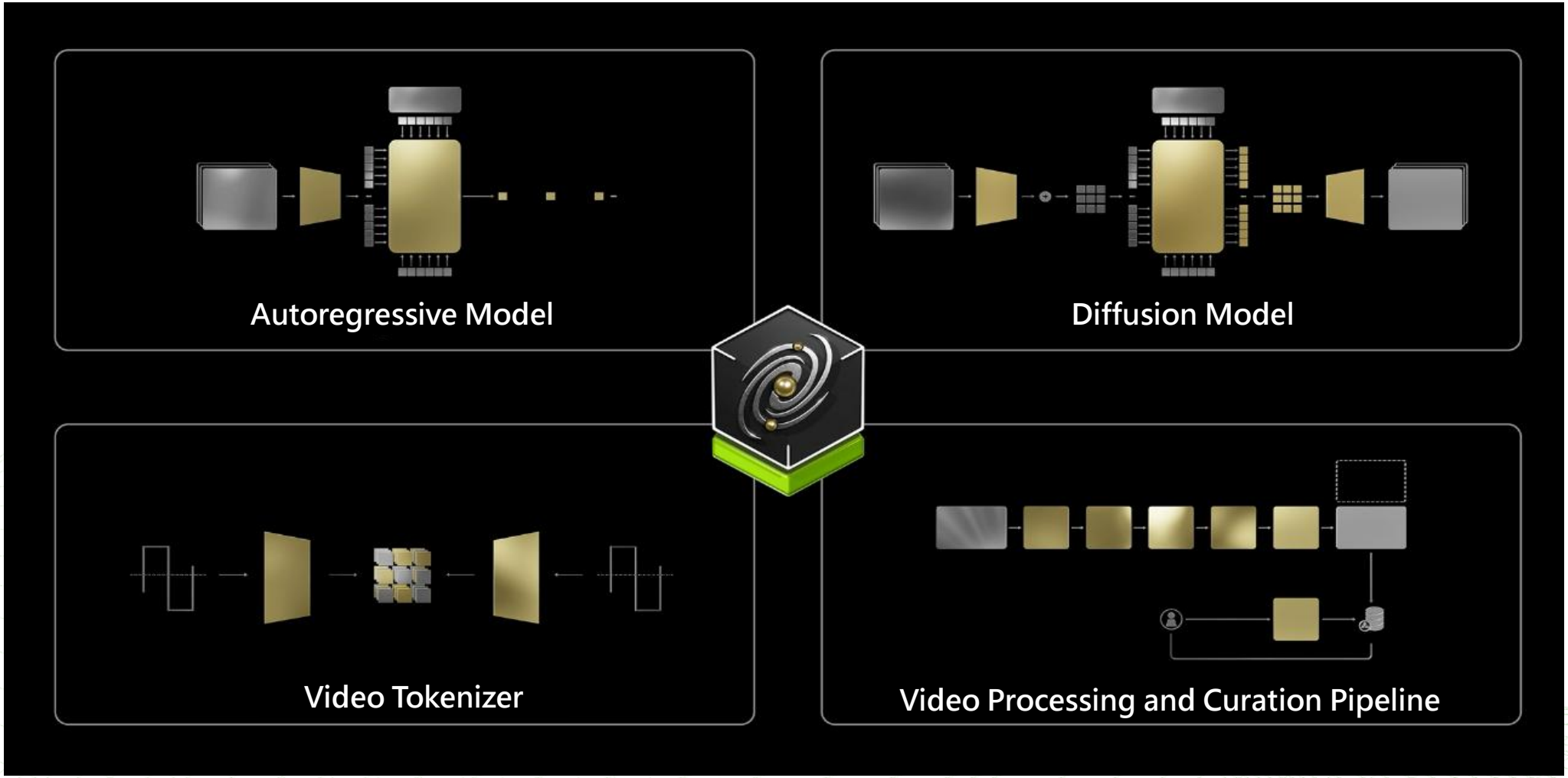
NVidia Cosmos (WFM)

https://www.youtube.com/watch?v=9Uch931cDx8



NVidia Cosmos (WFM)

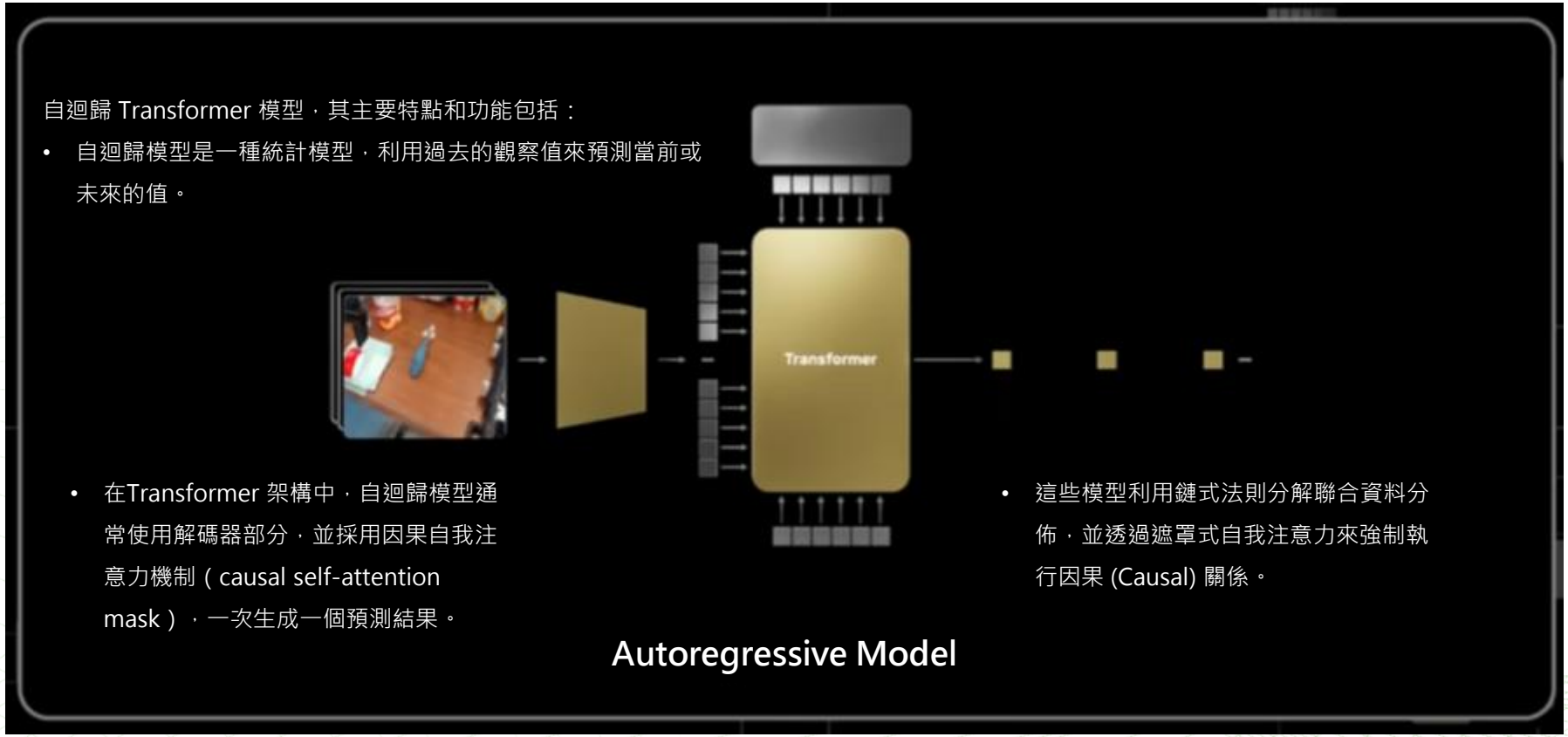
NVIDIA Cosmos, a platform of state-of-the-art generative world foundation models, advanced tokenizers, guardrails and an accelerated video processing pipeline — all designed to accelerate physical AI development.



NVidia Cosmos (WFM)

➢ Autoregressive Model（自回歸模型）

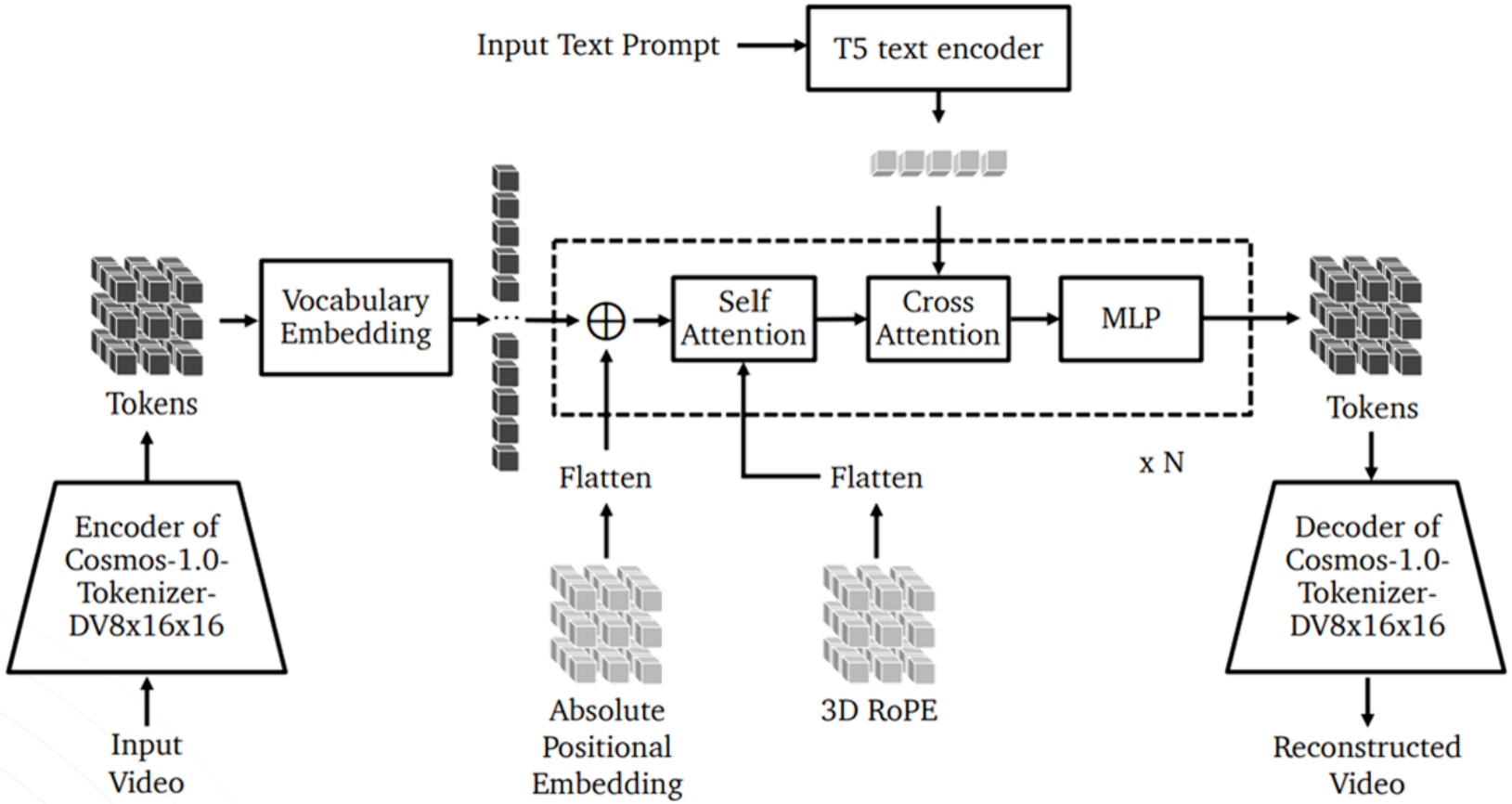
- 功能：負責預測影片中的下一個片段（token 或 frame）。
- 運作方式：根據先前的影片序列，逐步生成後續的內容。
- 角色：讓模型具備「從前文預測未來」的能力。



Autoregressive Model（自回歸模型）

Cosmos 自迴歸模型以更精確且更快速的未來影片幀預測，利用輸入文字、圖片及過去影片幀作為上下文。

自回歸模型是基於之前的生成內容案預設順序來逐段生成視頻。

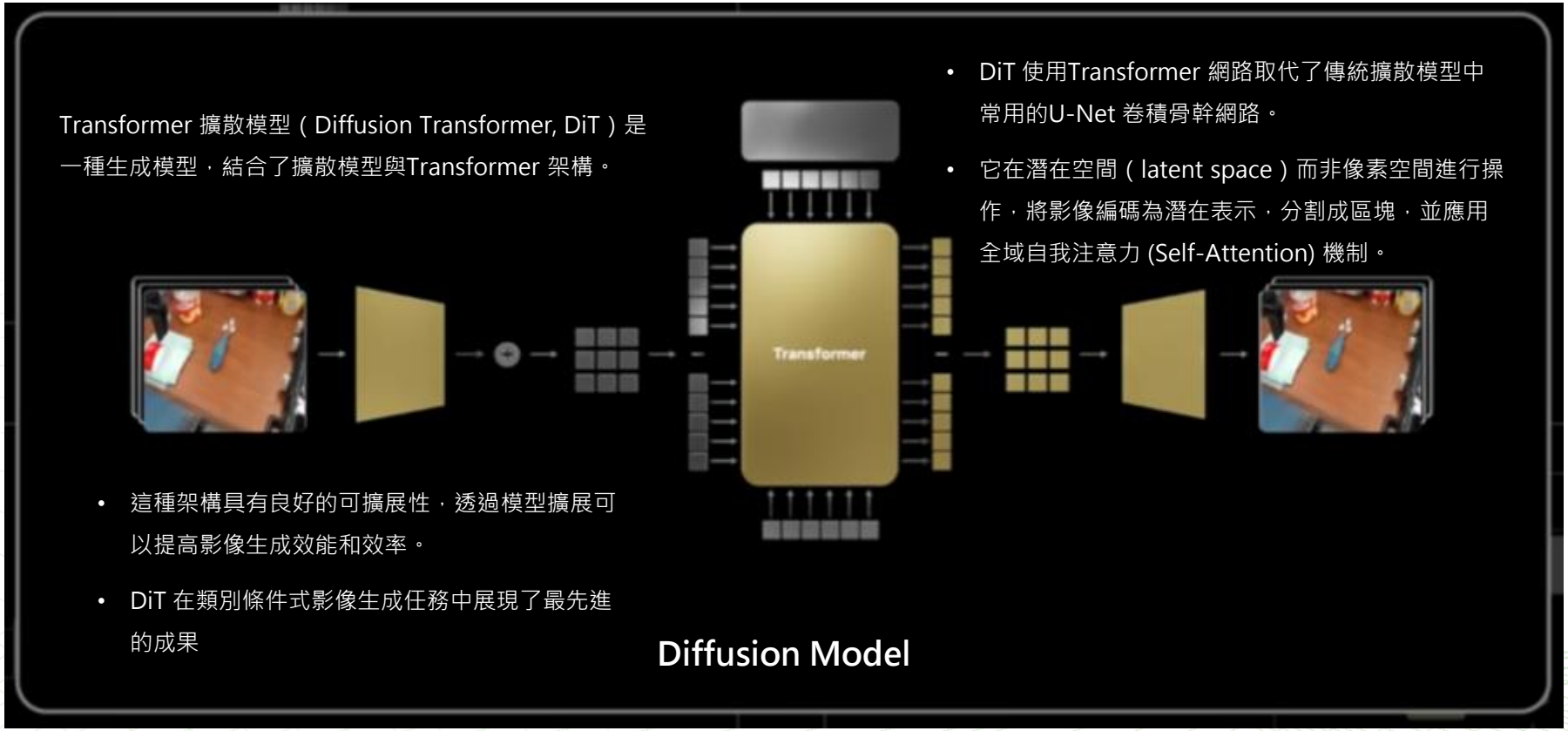


- 模型使用T5 文字編碼器處理文字輸入提示。
- 影片輸入透過Cosmos-1.0-Tokenizer-DV8x16x16 編碼器轉換為標記（Tokens）。
- 核心部分包含自我注意力（Self Attention）、交叉注意力（Cross Attention）和 MLP 層，並應用了3D ROPE 和絕對位置嵌入。
- 最終，標記透過解碼器重建為影片輸出。

NVidia Cosmos (WFM)

➢ Diffusion Model（擴散模型）

- 功能：用於生成高品質的影片或影像。
- 運作方式：先從隨機噪音開始，逐步「去噪」還原出真實的畫面。
- 角色：補充 autoregressive 模型，用於影像細節生成與修復。

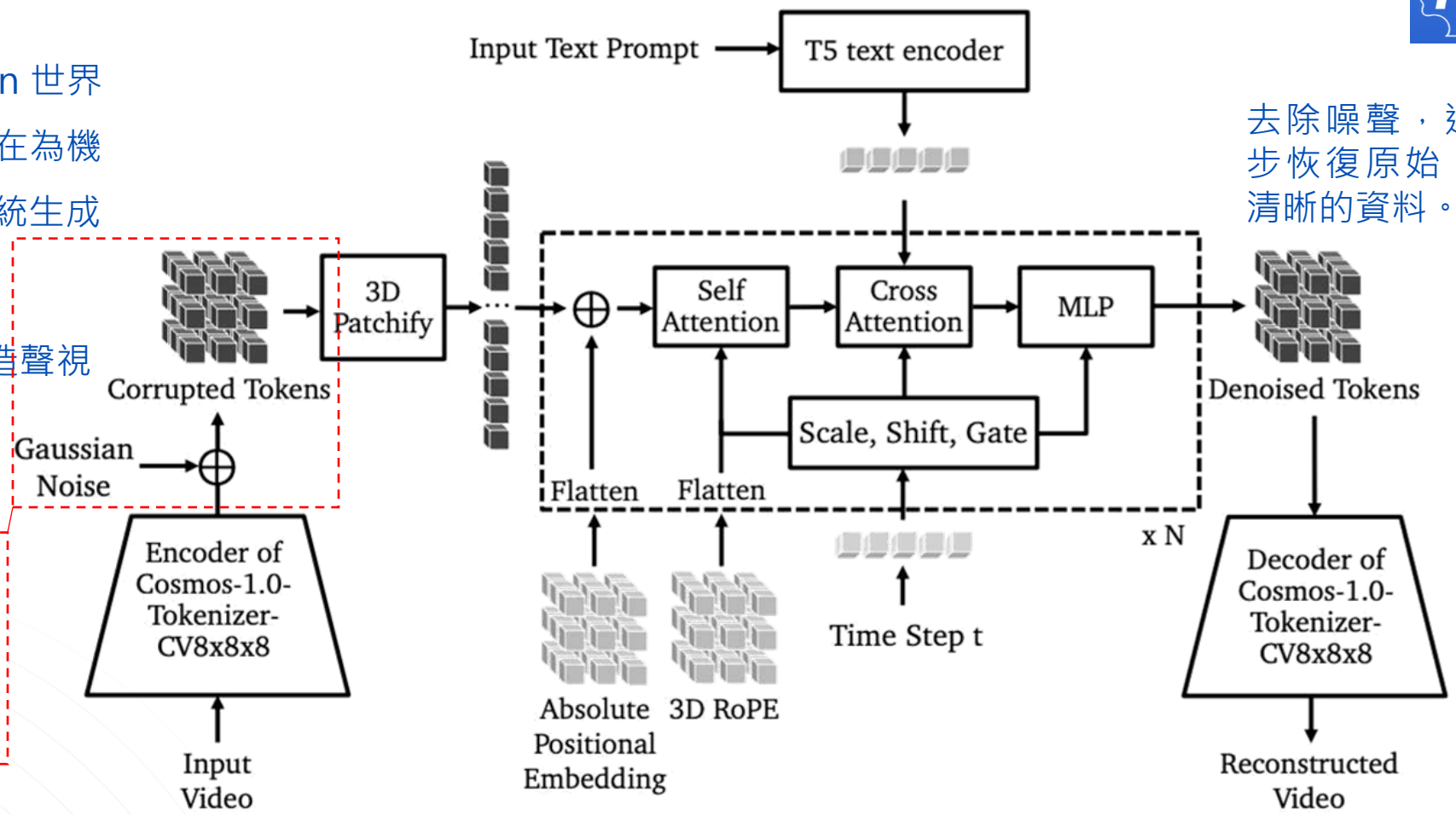


Diffusion Model（擴散模型）

NVIDIA Cosmos-1.0-Diffusion 世界基礎模型的架構圖。 該模型旨在為機器人學和自動駕駛等物理AI 系統生成逼真的合成數據。

擴散模型是通過逐步去除高斯造聲視頻中的躁聲來生成視頻。

「Corrupted Tokens」是指在訓練時被加入噪聲的影片或影像資料的數字表示形式，是模型學習去噪的輸入。

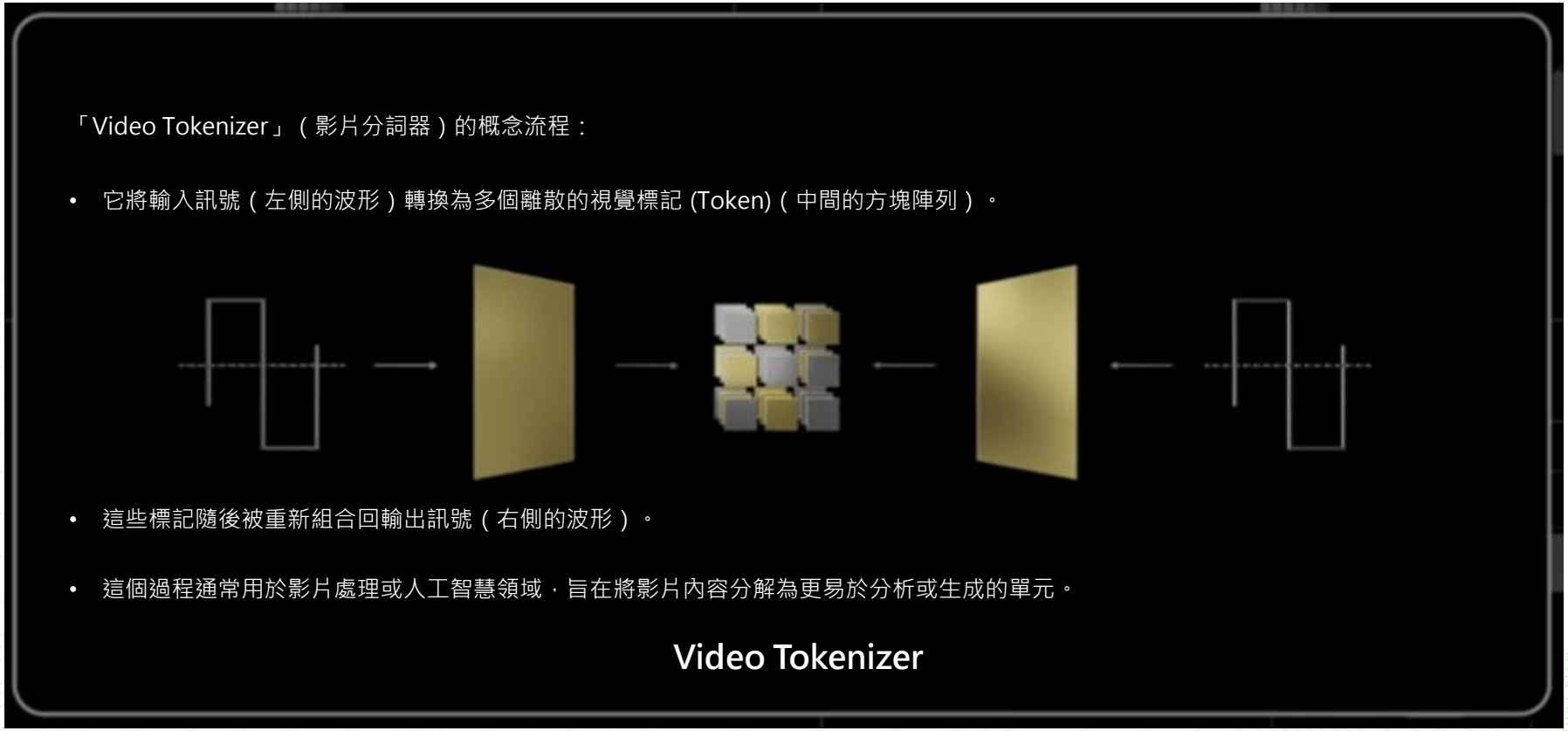


- **模型類型**: 擴散式Transformer 模型，用於潛在空間中的影片去噪。
- **輸入**: 接受輸入影片（透過編碼器轉換為標記）和文字提示（透過T5 編碼器）。
- **核心組件**: 包含交錯的自注意力 (Self-Aattention)、交叉注意力 (Cross Attention) 和MLP 層。
- **時間資訊**: 使用自適應層歸一化嵌入時間步長 (Time Step) 資訊，以進行去噪。
- **輸出**: 輸出重建的影片。

NVidia Cosmos (WFM)

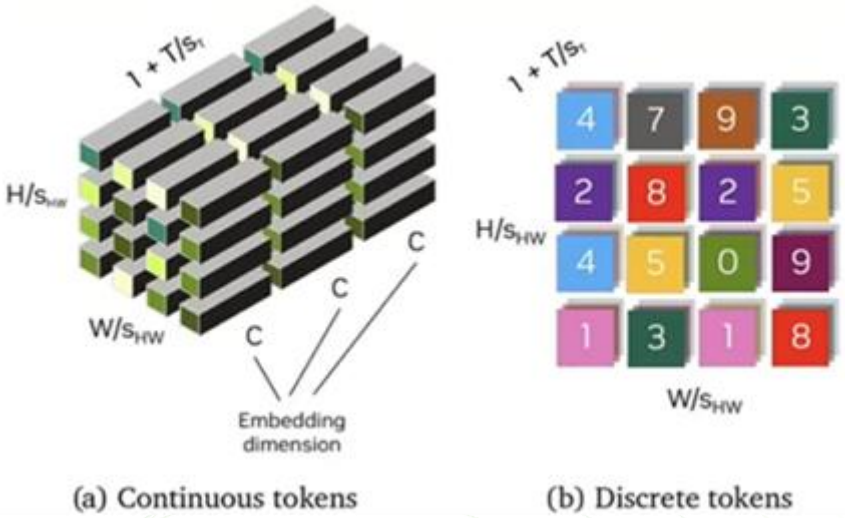
➢ Video Tokenizer（影片分詞器）

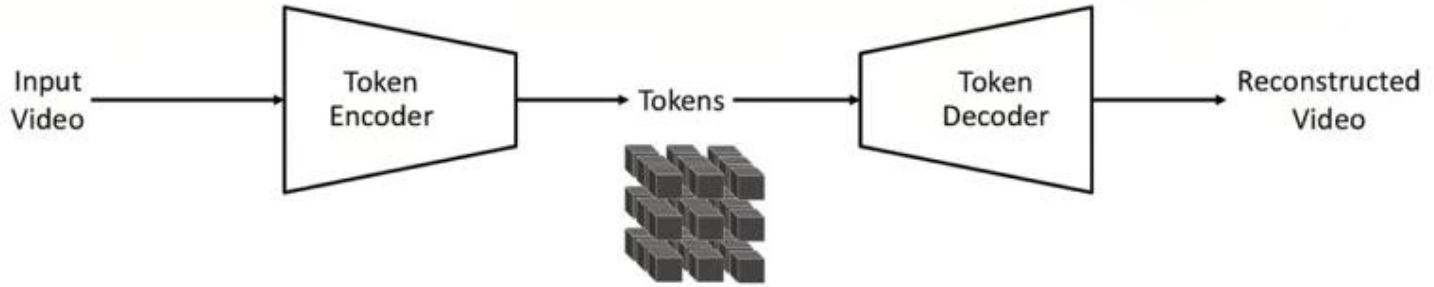
- 功能：將影片轉換成可被模型理解的「token」表示。
- 運作方式：影片 → 視覺編碼器 → token embedding → 量化。
- 角色：這是連接影片資料與語言式模型（如 transformer）的橋樑，使影片能以文字般的方式被處理。

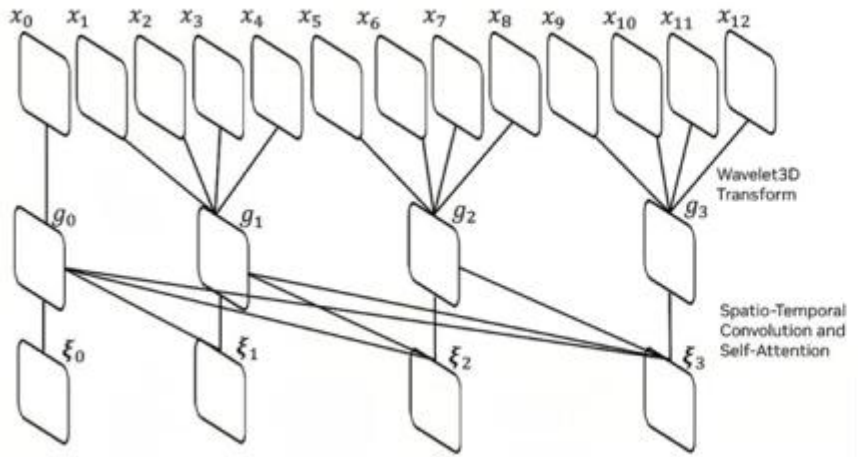


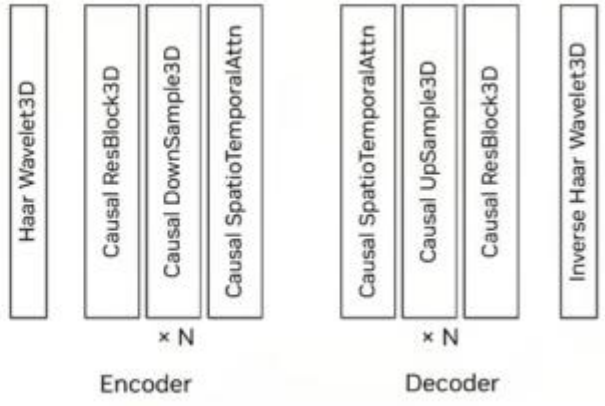
Video Tokenizer（影片分詞器）

Cosmos Tokenizer to handle both continuous and discrete tokenization for images and videos. It follows an **encoder-decoder design**.









- **連續型標記(a):** 顯示為具有高度(H/SHW)、寬度(W/S HW)、深度或嵌入維度(C) 的三維資料塊。
- **離散型標記(b):** 顯示為二維網格，其中每個單元格包含一個數字，代表一個離散的嵌入值。
- 這兩種表示方式都是為了加速開發預測與生成未來虛擬世界物理感知影片的神經網路。

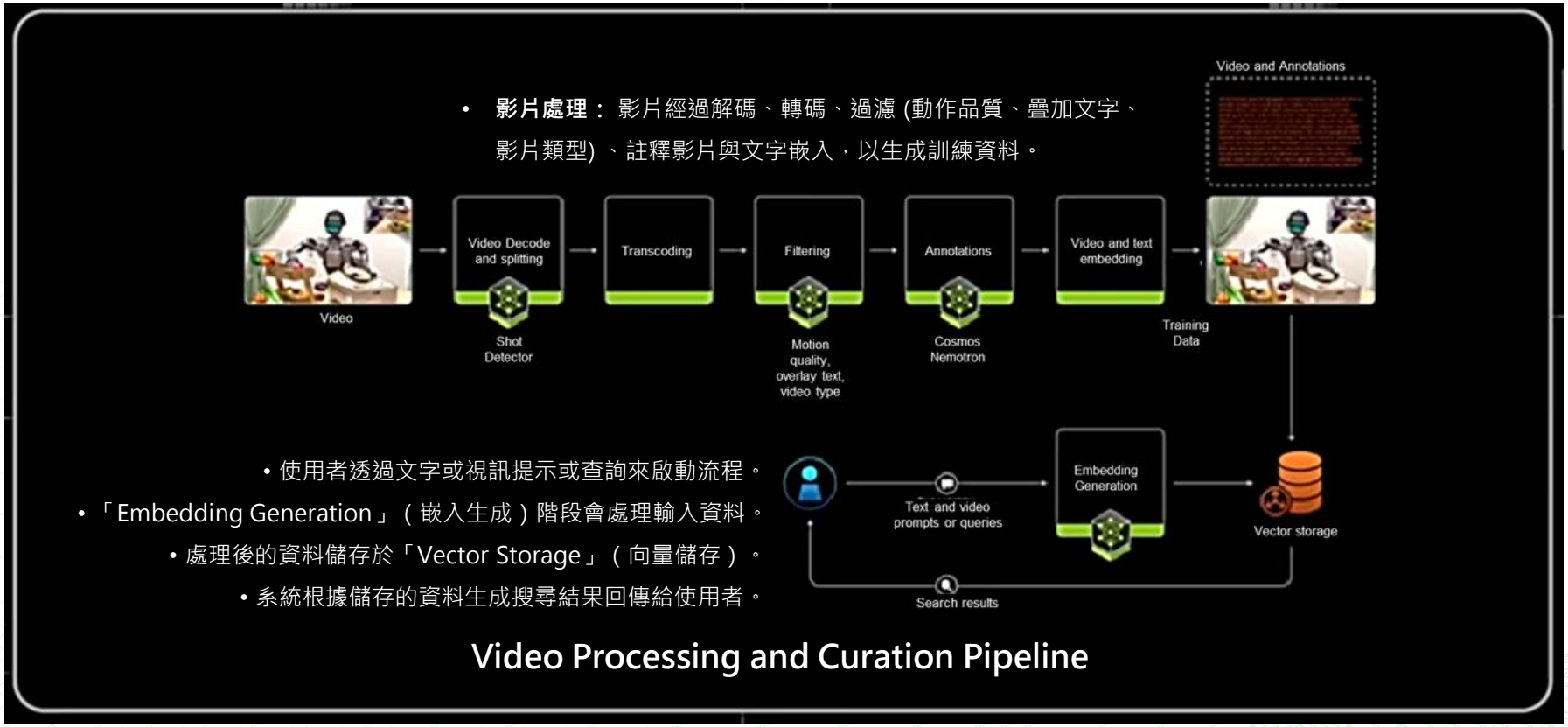
輸入 x_0、x_1,.....x_{12} 會經過分組的中間輸出 g_0、g_1,.....g_{12} 處理，並透過時空卷積與注意力運算進一步精煉。

NVidia Cosmos (WFM)

➢ Video Processing and Curation Pipeline（影片處理與整理流程）- 詳如下頁說明

- 功能：負責前處理訓練用影片資料。
- 運作方式：包含畫面擷取、資料清理、特徵抽取與分段。
- 角色：確保輸入資料乾淨一致，讓模型學習更有效率。

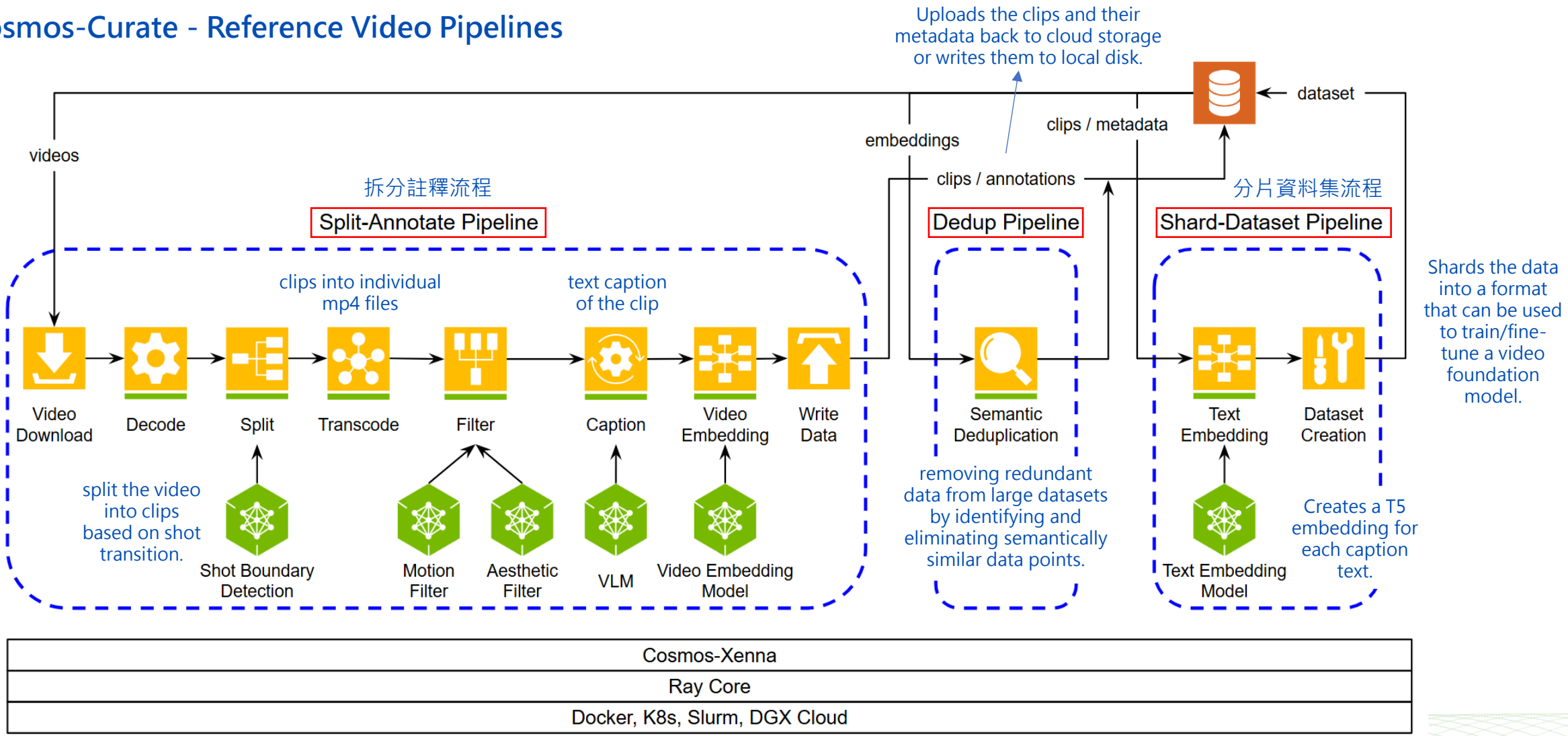
Cosmos Curator 是一個框架，可讓開發人員針對物理 AI 開發所需的大量感測器資料進行快速篩選、註記及去除重複項目，建立量身打造的資料集來滿足模型需要。接著，開發人員可以立即透過 NVIDIA Cosmos Dataset Search 來查詢這些資料集，並針對目標的後期訓練檢索場景。加速高效的資料集處理與生成。



NVidia CUDA (Compute Unified Devices Architecture，統一計算架構)

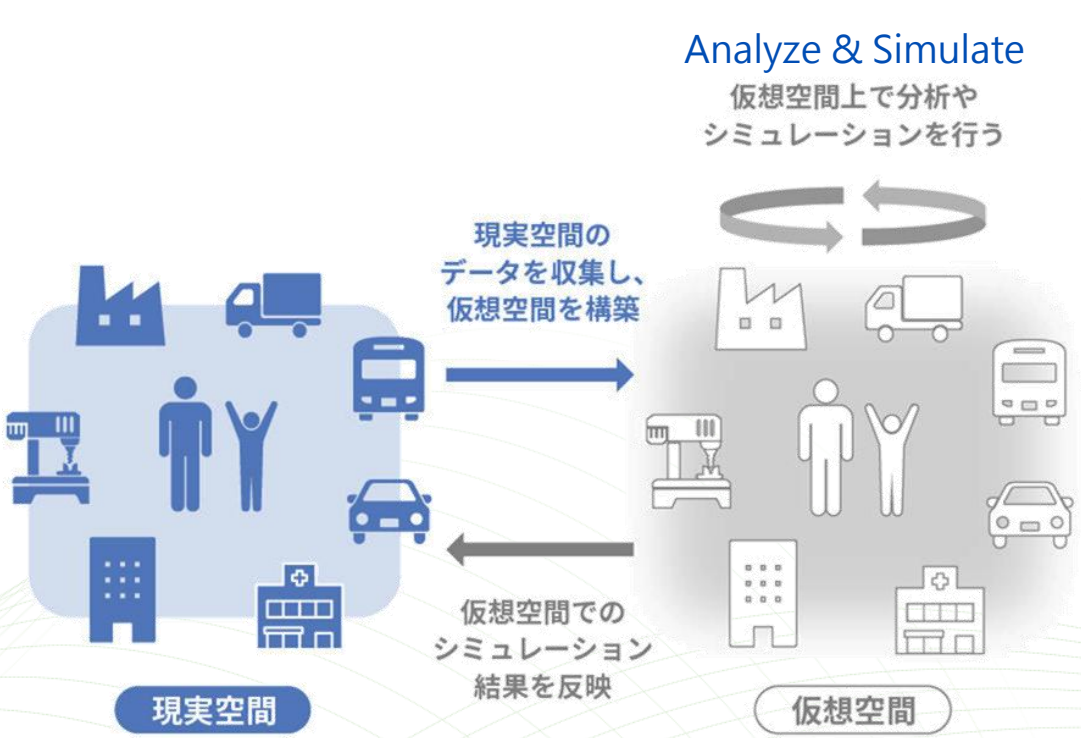
NVidia Cosmos (WFM)

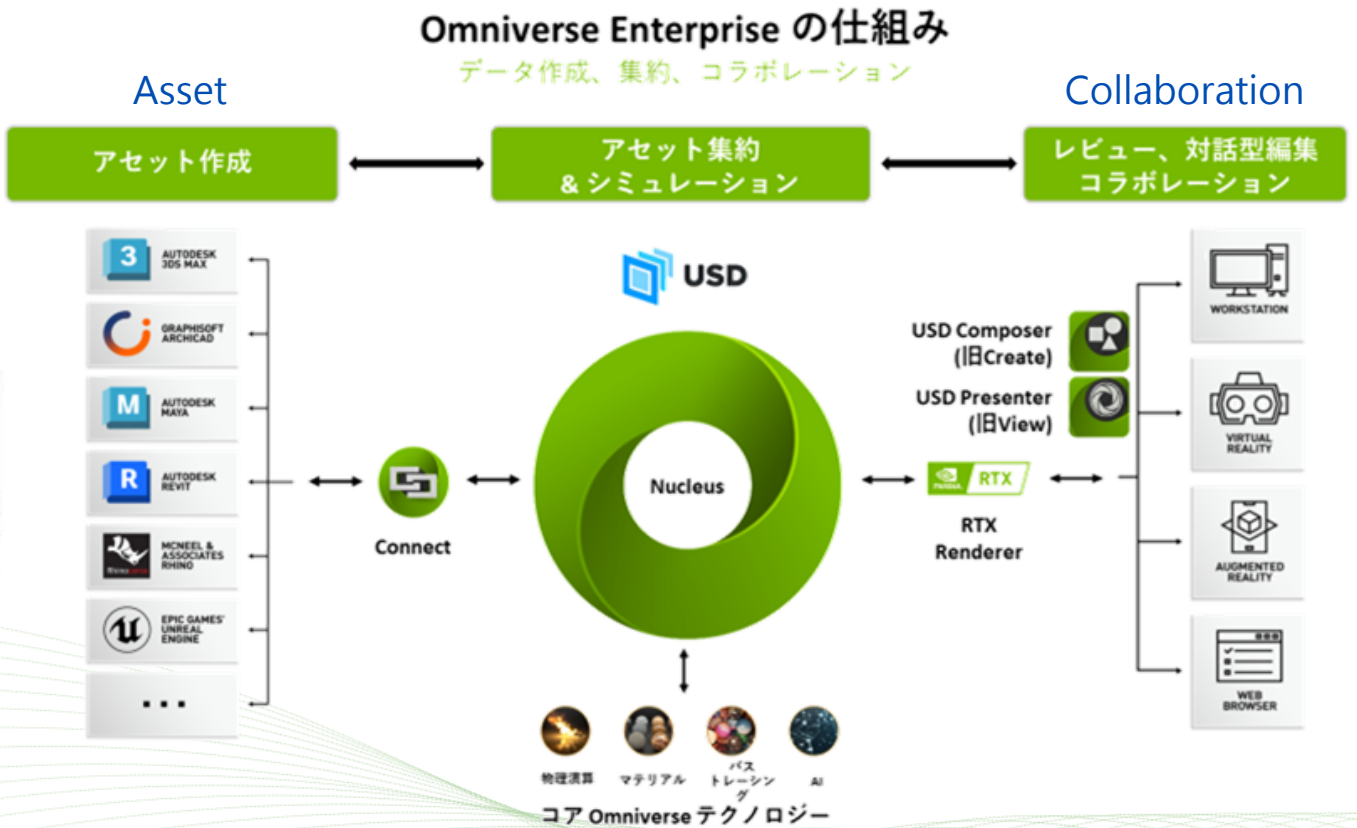
Cosmos-Curate - Reference Video Pipelines



NVidia Omniverse

A digital twin is a technology that builds a virtual space that resembles the real space on a computer based on various data collected from the real space. The name was given because it is exactly like a twin.



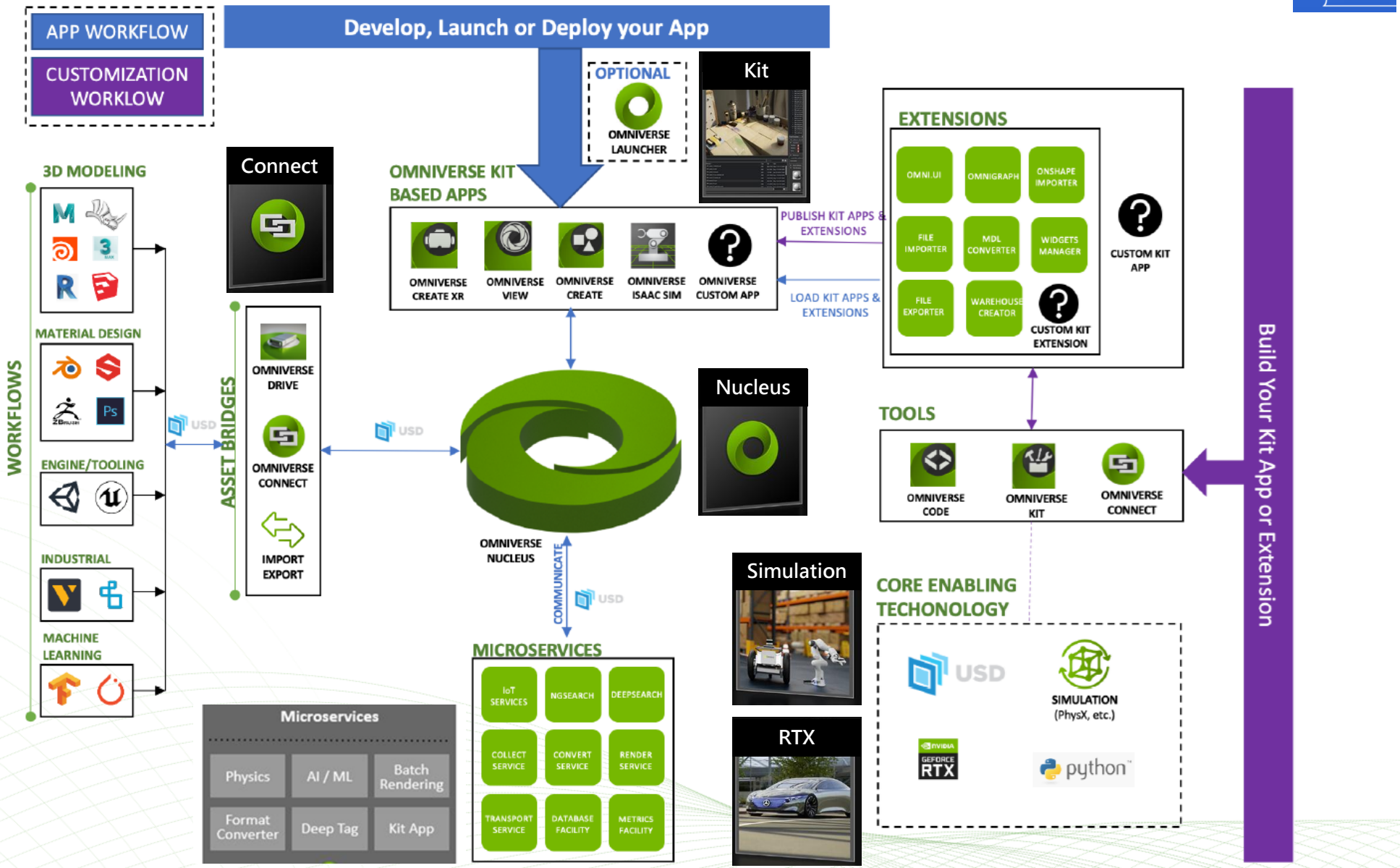


Core Omniverse technology

Jonathan Chen

NVidia Omniverse

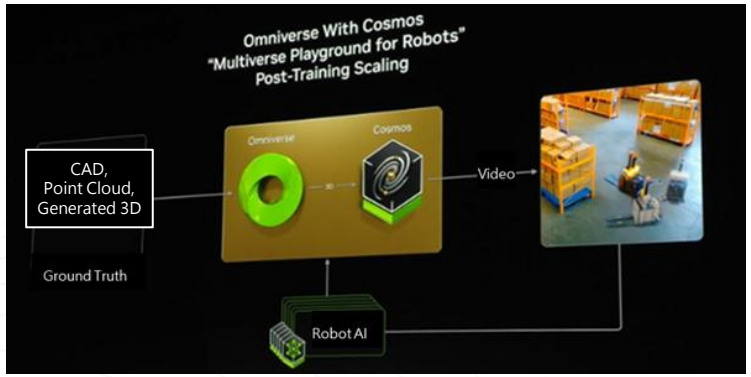
Omniverse由5個主要元件組成，包括用來連結各種客戶端應用與Nucleus DB的Omniverse **Connect**，擔任資料庫及協作引擎角色的Omniverse **Nucleus**，用來打造原生Omniverse應用及微服務的工具組 **Kit**，提供先進模擬能力的**Simulation**，以及能夠模擬實體世界光線環境的渲染平臺 **RTX**，上述元件再加上所連結的第三方數位內容與微服務，構成了Omniverse生態系統。

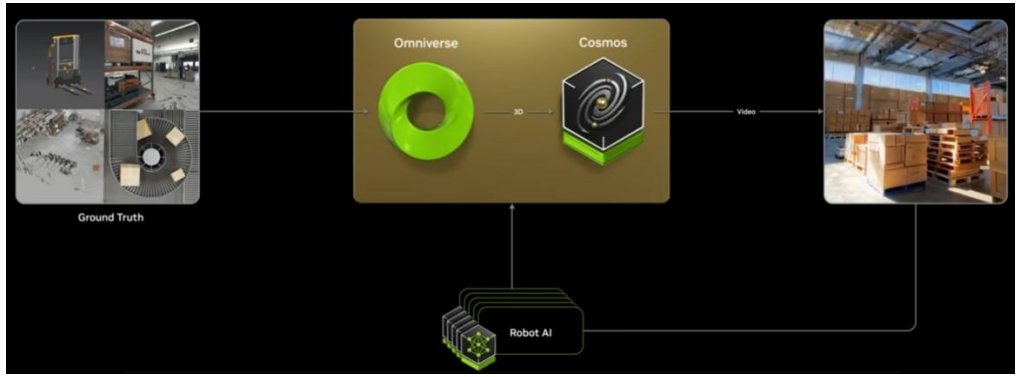


Jonathan Chen

NVidia Omniverse、Cosmos & OpenUSD

When paired with Omniverse, Cosmos creates a powerful synthetic data multiplication engine. Developers can use Omniverse to create 3D scenarios, then feed the outputs into Cosmos to generate controlled videos and variations. This can drastically accelerate the development of physical AI systems such as autonomous vehicles and robots by rapidly generating exponentially more training data covering a variety of environments and interactions. OpenUSD ensures the data in these scenarios is seamlessly integrated and consistently represented, enhancing the realism and effectiveness of the simulations.





Omniverse Cosmos Omniverse Cosmos Jonathan Chen