# 多模態生成應用概論

# Outline

- 圖像生成（Stable Diffusion、DALL·E）
- 語音與音樂生成（MusicGen、ElevenLabs）
- 影片生成初探（Runway、Sora）

# 多模態生成是什麼

- **「多模態」= 模型能理解/生成多種類型的資料**
  - 文字（text）
  - 圖像（image）
  - 聲音 / 音樂（audio / music）
  - 影片（video）
- **多模態生成**
  - 文字 ➜ 生成圖片
  - 文字 ➜ 生成語音 / 音樂
  - 文字 ➜ 生成影片
  - 圖片 ➜ 生成文字描述（反過來）

「從只會打字的 AI，走向會畫圖、說話、配樂、做影片的『多才多藝 AI』。」

# 文字模型 vs 多模態模型

- **傳統（文字為主）的 LLM**
  - GPT、Claude、LLaMA…
  - 主要輸入 / 輸出：**文字**
  - 擅長：
    - 回答問題
    - 生成報告、Email、文件
    - 摘要 / 改寫文字
- **多模態生成模型**
  - 圖像生成：Stable Diffusion、DALL·E
  - 語音 / 音樂：ElevenLabs、MusicGen
  - 影片：Runway、Sora
  - 特色：
    - 輸入可能是文字、圖片、影片片段
    - 輸出可以是圖像、聲音、影片

# AI 如何「畫出」一張圖片

- 輸入：一隻狗
- AI 會從「一張雜訊圖」開始一步一步「去噪」，逐漸生成細節，最後還原成你想要的圖片，這個過程使用稱為 Diffusion（擴散模型）的技術

初始噪點 → 模糊 → 大致形狀 → 清晰細節

ChtGPT生成影像

# Stable Diffusion

prompt輸入

參數設定

過去生成的圖片



# DALL·E

- **DALL·E已整入ChatGPT，直接使用ChatGPT畫圖**

生成一張 PCB 上焊點氧化的不良圖片。

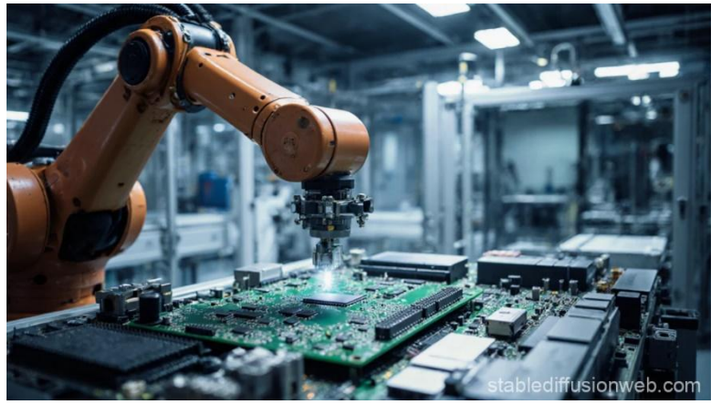
已建立圖像 ›

( 已為你生成一張 PCB 焊點氧化的不良示意圖。如果你需要更多角度、不同嚴重程度或 10 張成套圖片，也可以告訴我。 )

# Stable Diffusion vs DALL·E

| 項目 | Stable Diffusion | DALL·E |
| --- | --- | --- |
| 模型屬性 | 開源模型 | 商業模型（OpenAI） |
| 部署方式 | 本地、內網、雲端皆可 | 雲端 API / ChatGPT |
| 特點 | 可微調、可客製化、架構彈性 | 生成效果穩定、語意理解強 |
| 成本 | 低、可自架運行 | 使用 API / 訂閱費用 |
| 適合場景 | 企業內部資料、影像資料擴增、模型微調 | 快速生成高品質提案圖、行銷圖片 |
| 操作難度 | 需技術人員介入 | 幾乎零門檻 |

# Stable Diffusion vs DALL·E

| | 「在電子工廠中，一台自動化機械手臂正在組裝電路板，乾淨工業風，光線柔和。」 | 「工人在產線操作機械，正確穿戴護具，平面插畫風格。」 | 「未來風格的智慧製造設備，淡色科技感，乾淨背景。」 |
|---|---|---|---|
| Stable Diffusion |   |   |   |
| DALL·E |  |  |   |

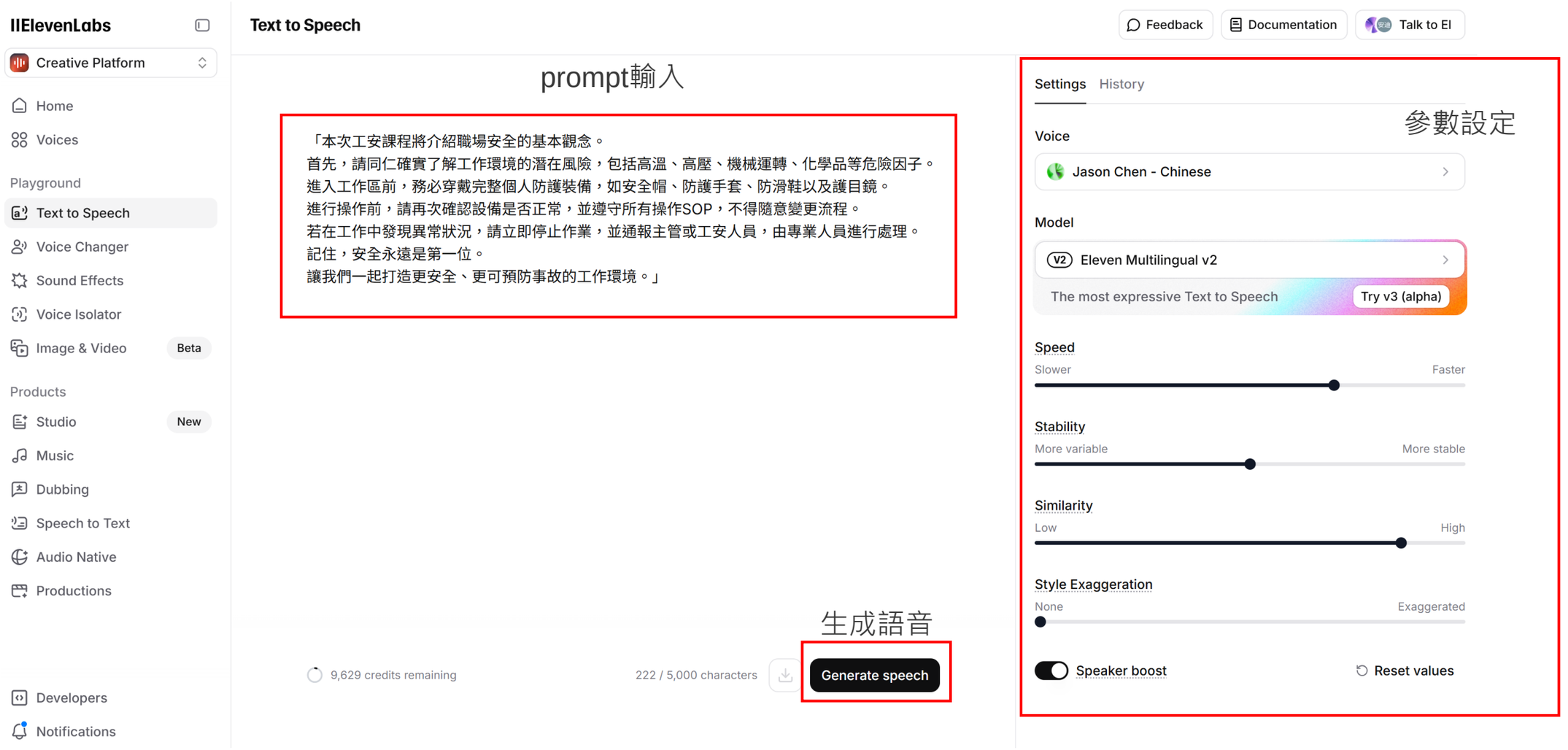
# 語音生成是什麼

- 將文字轉成語音（TTS：Text-to-Speech）
- 使用深度學習模型生成自然語調
- 已能做到：
  - 多種語氣（沉穩、熱情、專業、客服語調...）
  - 多語言切換
  - 自動調整斷句、停頓、強調

# 語音生成：ElevenLabs

- **能做什麼：**
  - 文字 → 語音（Text-to-Speech）
  - 複製聲音（Voice Cloning）
  - 多語言配音
- **應用案例：**
  - 教育訓練影片自動產生配音
  - SOP 語音版本
  - 客服語音 Bot
  - 製造業多語言員工訓練（中/英/泰/越南語）

# ElevenLabs




中文

英文

# 音樂生成：MusicGen

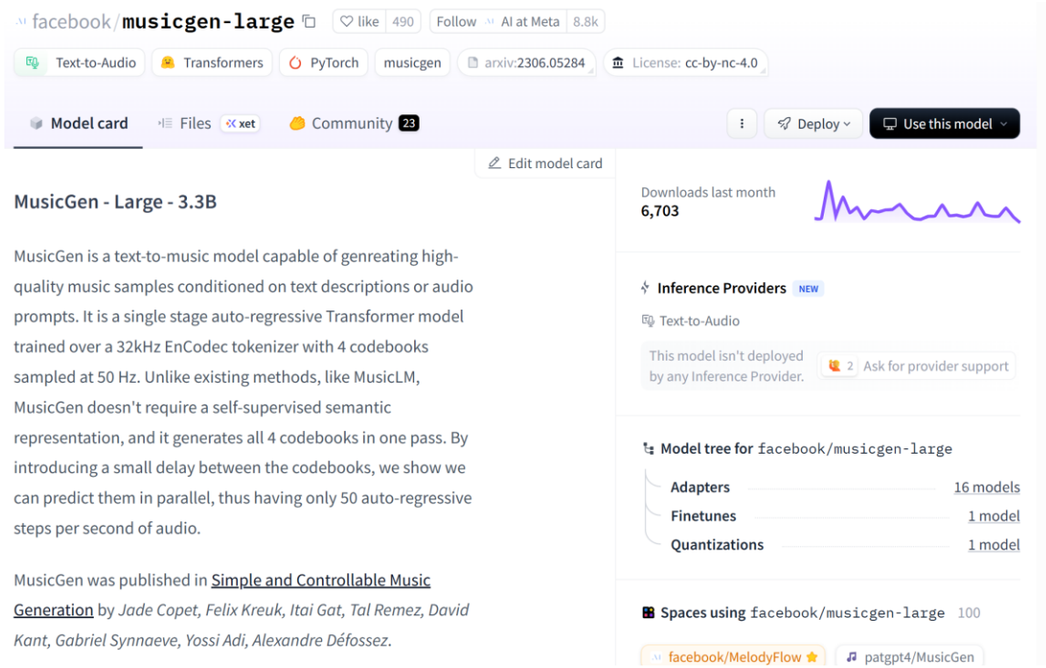
- **Meta 開源音樂生成模型**
- **可依照文字描述生成背景音樂，例如：**
  - 「科技感、節奏明快的背景音樂」
  - 「輕柔鋼琴、平靜、適合作為 BGM」
- **應用案例：**
  - 企業影片背景音樂
  - 產品展示影片
  - 教學教材配樂

# MusicGen

## Text-to-music generation

MusicGen is an audio generation model specifically tailored for music generation. Music tracks are more complex than environmental sounds, and generating coherent samples on the long-term structure is especially important when creating novel musical pieces. Our modeling approach naturally extends to stereophonic music generation.

| | | |
|---|---|---|
| A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a heroic battle. | MusicGen 3.3B ▶ 0:16 | MusicGen 3.3B stereo ▶ 0:00 |
| Classic reggae track with an electronic guitar solo | MusicGen 3.3B ▶ 0:08 | MusicGen 3.3B stereo ▶ 0:09 |
| drum and bass beat with intense percussions | MusicGen 3.3B ▶ 0:00 | MusicGen 3.3B stereo ▶ 0:00 |
| A dynamic blend of hip-hop and orchestral elements, with sweeping strings and brass, evoking the vibrant energy of the city. | MusicGen 3.3B ▶ 0:00 | MusicGen 3.3B stereo ▶ 0:00 |




https://audiocraft.metademolab.com/musicgen.html
https://huggingface.co/facebook/musicgen-large

# 語音 / 音樂生成的倫理與安全注意事項

- **個人聲音 = 個資**
  - 未經本人允許，不得克隆真實聲音
  - 若用在公司內部：須由聲音提供者授權建議存留「同意紀錄」
- **深偽語音（Deepfake）風險**
  - 可被濫用於詐騙、假指令、假電話
  - 公司內部建議：禁止用別人的聲音做敏感指令設定清楚使用範圍
- **音樂版權問題**
  - 使用需注意：是否使用受保護的素材、是否有商用授權

# ElevenLabs vs MusicGen

| 項目 | MusicGen | ElevenLabs |
| --- | --- | --- |
| 類型 &用途 | 文字或音樂提示 → 生成**音樂 (Music)** | 文字提示 → 生成**語音 / 複製人聲 (Voice TTS / Voice Cloning)** |
| 核心功能 | ● 提供從簡單文字描述生成完整音樂<br>● 支援音樂特徵控制（旋律 / 風格） | ● 高品質文字→語音<br>● 可複製特定人聲 / 音色（提供 Instant Voice Cloning、Professional Voice Cloning） |
| 部署特性 | 開源模型（Meta AudioCraft）可供研究 / 自架 | 商用平台為主，提供 API 與雲端服務，亦有商用授權條款 |
| 適合場景 | ● 影片背景音樂、產品展示影片配<br>● 教學影音配樂、自主品牌音樂生成 | ● 教材配音、語音導覽、客服語音機器<br>● 品牌專屬聲音、語音克隆、語音播報 |
| 風險 / 注意事項 | 音樂版權、音樂生成品質、風格控制 | 聲音克隆涉及個人聲音權利、假冒風險、合規與授權問題 |

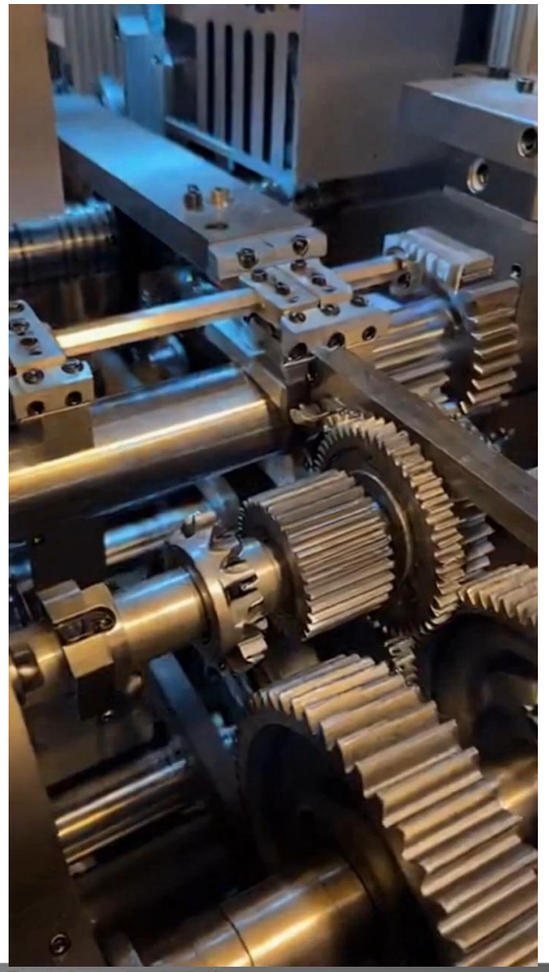
# 影片生成是什麼

- **輸入文字描述（Prompt）→ AI 自動生成影片**
- **影片可包含：**
  - 人物、物體
  - 運動光線、攝影風格
  - 連續動作、鏡頭角度
- **影片生成的關鍵技術：**
  - Diffusion（和圖像生成類似，但更複雜）
  - 時序一致性（Temporal Consistency）
  - 物理與場景理解（Physics-aware）

# 影片生成：Sora

- **OpenAI 新一代影片生成模型**
- **特色**
  - 高度理解「物理運動」
  - 影片長度可達 1 分鐘（未來更長）
  - 畫質細緻
  - 支持鏡頭語言（tracking shot / dolly / close-up）
  - 能夠生成複雜場景（人物 + 動作 + 互動）
- **適合場景**
  - 高品質技術影片
  - 工業場景模擬
  - 虛擬產品展示
  - 未具備拍攝條件時的「替代影片素材」

**Prompt:**
鏡頭緩慢地掃過自動化設備的內部機械結構，逼真的光影效果突顯了金屬齒輪和零件。鏡頭從左到右緩緩移動，展現了精密工程的複雜細節。



https://openai.com/zh-Hant/index/sora/

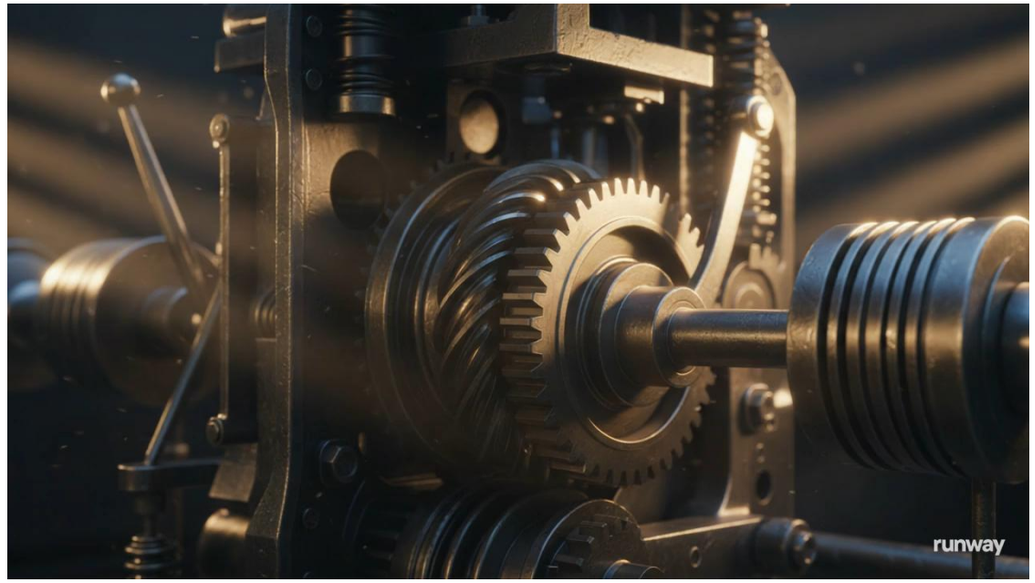
# Sora 影片生成 Prompt 技巧

| 構成元素 | 說明 | 範例 |
| --- | --- | --- |
| **主體（Subject）** | 明確描述人物 / 物品的外觀、年齡、穿著、特徵 | 「20 歲短髮女生，穿白色襯衫與牛仔裙」 |
| **動作（Action）** | 主體正在做什麼，越具體越好 | 「一邊走路一邊滑手機」 |
| **場景（Scene）** | 地點、時間、環境狀態、背景元素 | 「在傍晚台北信義區街頭，人潮穿梭」 |
| **光影效果（Lighting）** | 決定影片質感：夕陽、霓虹燈、柔光等 | 「金色夕陽照在臉上，柔光、暖色調」 |
| **鏡頭語言（Camera Work）** | 專業攝影指令讓影片更電影感 | 「低角度拍攝，鏡頭平滑向前推進」 |
| **風格（Style）** | 影片整體美術風格，如寫實、動畫、皮克斯等 | 「皮克斯動畫風，角色可愛誇張」 |
| **技術參數（Technical）** | 解析度、比例、角色一致性等 | 「4K，16:9，角色外觀保持一致」 |

# 影片生成：Runway

- **Runway：快速、易用、適合實際案子的影片生成工具**
- **特色：**
  - 文生影片（Text-to-Video）
  - 影片延伸（Video-to-Video / Extend Video）
  - 影片風格控制（cinematic / anime / 3D style）
  - 影片編輯（背景移除、遮罩、替換背景）
- **適合場景：**
  - 教學影片素材
  - 工廠示意影片
  - 行銷短影片
  - 產品展示動畫

**Prompt:**
鏡頭緩慢地掃過自動化設備的內部機械結構，逼真的光影效果突顯了金屬齒輪和零件。鏡頭從左到右緩緩移動，展現了精密工程的複雜細節。



# Runway

生成之影片

上傳一張圖片


輸入Prompt

生成影片

# Runway vs Sora

| 項目 | Runway | Sora（OpenAI） |
|---|---|---|
| 定位 | 已商用的 **影片生成 & 編輯平台** | 新一代 **高品質文生影片模型** |
| 主要功能 | ● 文字 → 影片（Text-to-Video）<br>● 影片 → 影片風格轉換（Video-to-Video）<br>● 背景移除、遮罩、合成<br>● 影片延伸、補幀、上色等剪輯輔助 | ● 文字 → 高畫質長片段影片<br>● 支援複雜場景、多角色、鏡頭運動<br>● 較佳的物理理解與連續動作一致性 |
| 操作方式 | ● 網頁介面操作<br>● 適合設計師、企劃、行銷直接上手 | ● 目前主要透過 API / 平台整合（仍在逐步開放）<br>● 多由開發者或合作夥伴串接到應用中 |
| 影片長度 & 品質 | ● 支援短秒數影片生成（適合社群短片）<br>● 畫質佳，適合一般教學 / 行銷用途 | ● 可生成較長、畫質更高的影片<br>● 場景細節與物理表現更逼真，適合高品質展示 |
| 優點 | ● 已成熟商用、功能完整<br>● 同時具備「生成 + 剪輯 + 特效」一站式工具<br>● 不需寫程式即可使用 | ● 影像品質與連貫性非常強<br>● 對文字描述、場景與動作理解能力佳<br>● 適合做高質感、複雜場景的 Demo 或概念影片 |