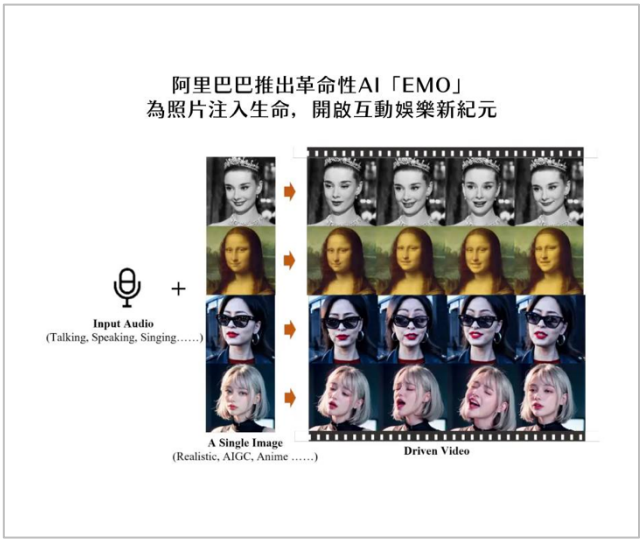
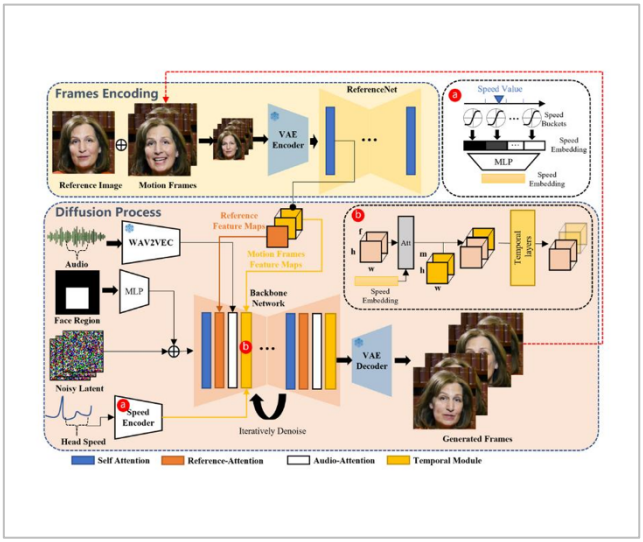
阿里巴巴推出革命性 AI「EMO」

為照片注入生命，開啟互動娛樂新紀元

人工智慧的創新應用持續為人們的生活帶來驚喜。近日中國網路巨擘阿里巴巴智慧運算研究所發表了一項顛覆性的技術成果——「EMO」（Emote Portrait Alive），一種全新的 AI 影片生成模型。這項技術的核心在於其能夠將一張靜態照片與相應的音訊結合，創造出能夠張嘴說話甚至唱歌的動態人像。

EMO 不依賴通常難以捕捉人類表達細微差別的傳統方法，而是直接將音訊波形轉換為視訊幀。這意味著創建動畫不需要中間 3D 模型或臉部標誌。相反，它專注於捕捉與自然語音相關的微妙面部動作和個人面部風格，這一進步無疑為多媒體互動和娛樂領域帶來了新的篇章。





EMO 模型的多語言能力極大地擴展了其應用範圍，使其不僅能理解英語，還能精通韓語等多種語言。阿里巴巴展示的一系列 EMO 生成的影片中，尤其是以 Sora 為例的東京街頭人物影片，展示了 EMO 在細節處理上的高超技術。這些影片讓人們能夠清楚地看到，不僅人物的嘴型與音訊同步，整體表情和頭部動作也與其說話或唱歌時的內容和情感密切相關，呈現出接近真實人類的自然表達。

對比於 NVIDIA 的「Audio2Face」工具，外媒《Mashable》指出 EMO 在表達相對應音訊的複雜情感方面顯得更加出色，而「Audio2Face」則更像是一個戴著面具的木偶。這一比較突出了 EMO 在捕捉和再現人類情感表達方面的先進性。

EMO 的運作機制基於 Audio2Video 擴散模型，這一模型經過超過 250 小時的談話影片（包括演講、電影、歌唱表演等）進行訓練，並採用了注意機制和音訊注意機制的配對，確保生成的面部動作與音訊內容保持一致性。EMO 的訓練過程分為編碼階段和擴散階段，其中編碼階段利用 ReferenceNet 從參考圖像和動態影格中提取特徵，而擴散階段則通過預訓練的音訊編碼器和面部區域掩模來精細控制面部圖像的生成。

儘管 EMO 目前尚未向大眾開放測試，但其所展示出的能力已經讓人們對未來充滿了期待.

參考資料｜https://www.inside.com.tw/article/34341-alibaba-emo-model

影片連結｜Alibaba presents EMO AI - All Demo Clips Upscaled to 4K