前瞻科技與管理　12 卷 2 期，1-11 頁（2024 年 5 月）
Journal of Advanced Technology and Management Vol. 12, No. 2, pp. 1-11 (May, 2024)
DOI:10.6193/JATM.202405_12(2).0001

從文字到動作的轉化：MotionGPT

蘇明祥 [1,*]　何書維 [2]　徐澍萬 [2]　林紘宇 [2]

[1] 東吳大學資料科學系助理教授
[2] 東吳大學資料科學系學生

摘 要

本次研究主要探討了利用大型語言模型（如基於轉換器的生成式預訓練模型〔Generative Pre-Trained Transformers, GPT〕系列）和其他人工智慧（Artificial Intelligence, AI）技術，如 Large Language Model Meta AI（LLaMA）和 T5 模型，進行文本到動作的轉換。文章詳細分析了這些模型的結構和功能，並比較了它們在生成動作影片方面的效能。研究使用了不同的技術進行實驗，如均方根正規化（Root Mean Square Normalization, RMSNorm）和絕對編碼，來探索最佳的文本到動作轉換方法。研究結果顯示，T5 模型在根據文本描述生成動作方面表現更為優異，特別是在呈現關鍵動作和避免不必要動作方面。這些發現為未來的動作生成技術發展提供了有價值的見解。

關鍵詞：動作生成、GPT、轉換器、動作擴散、文本轉動作

[*] 通訊作者：蘇明祥
電子郵件：huntfox.su@gmail.com
（收件日期：2024 年 1 月 15 日；修正日期：2024 年 3 月 5 日；接受日期：2024 年 3 月 6 日）

Journal of Advanced Technology and Management Vol. 12, No. 2, pp. 1-11 (May, 2024)
DOI:10.6193/JATM.202405_12(2).0001

Text-to-Motion Transformation: MotionGPT

Ming-Hsiang Su[1,*], Shu-Wei Ho[2], Shu-Yu Hsu[2], Hung-Yu Lin[2]

[1]Assistant Professor, Department of Data Science, Soochow University
[2]Student, Department of Data Science, Soochow University

Abstract

This study explores text conversion to motion using large language models like the generative pre-trained transformer (GPT) series and other artificial intelligence (AI) technologies like Large Language Model Meta AI (LLaMA) and the T5 model. It analyzes the structure and functions of these models in detail, comparing their effectiveness in generating motion videos. Various techniques like root mean square normalization (RMSNorm) and absolute encoding were employed to identify the best method for text-to-motion conversion. The findings indicate that the T5 model generates actions based on textual descriptions, especially in presenting critical motions and avoiding unnecessary movements, offering valuable insights for future advancements in motion generation technology.

Keywords: motion generation, GPT, transformer, motion diffuse, text-to-motion

* Corresponding Author: Ming-Hsiang Su
E-mail: huntfox.su@gmail.com

壹、緒論

近幾年文字轉圖像的生成式人工智慧（Artificial Intelligence, AI）已經逐漸成熟，不管在學術界或者產業界，人們可以使用相對比較口語化的方式與機器人互動來獲得想要的圖片，甚至是會動的圖片或者影片。有研究指出在生成動畫過程中，可以保留特定目標曲線扭曲和循環扭曲的方法，並運用名為 Recursive Specularity Factorization Network（RSFNet）的技術學習特徵相關性，以創造出具有循環視覺效果的影片。這樣的方法在地圖圖像的處理上，塑造出一種獨特的地圖藝術風格影片（黎氏玉幸，2023）。

基於轉換器的生成式預訓練模型（Generative Pre-Trained Transformer, GPT）系列是由 OpenAI 提出的預訓練語言模型，這一系列的模型可以執行非常複雜的自然語言處理（Natural Language Processing, NLP）任務，例如回覆問題、生成文章和程式碼，或者翻譯文章內容等。另外 GPT 採用轉換器（Transformer）作為解碼器（Decoder），轉換器由 Google Brain 所推出，主要是處理自然語言的順序輸入資料，用於翻譯、文字摘要等任務上。隨著 GPT 的發展越來越成熟，大型語言模型的運用也不僅侷限於文本對文本（Text-to-Text）的功能上，例如本次研究的 MotionGPT 就是文本轉動作（Text-to-Motion）的應用，輸入對動作或行為的文本及描述，模型會利用輸入的文字描述生成動作。

此外，在目前也有幾個開源 AI 產品在市面上推出，像是 Runway ML、DeepBrain、Pika Lab、Rememory 與 MotionDiffuse。前三者在文字轉動作的可用度上較為不成熟，主要以小動作、小動畫為生成目標，比如 DeepBrain 透過文本描述便可以產生虛擬主播，可是產出的影片較為單調，主要指可以生成出新聞播報的畫面。另外，Pika Lab 雖然可以藉由提供的圖片加上文本描述來製作大約三秒的短動畫，但是動畫主角在動作上的細緻度比較粗糙，而且動作的完整度也會因為時間的限制，造成動作不完整的狀況發生。MoMask（Guo, Mu, Javed, Cheng, and Wang, 2023）這個方法使用層次化的量化方案來表示人體運動，並利用 Masked Transformer 和 Residual Transformer 來生成動作，並且可以在不需要額外模型微調的情況下適應相關任務，例如文本引導的時序修補（Tmporal Inpainting）。

貳、文獻探討

在文本描述生成 3D 人型動作的模型當中，最一開始常見的方式是利用擴散模型作為文字與動作影像不同類別的，一種深度學習的生成模型，目的為模擬數據的擴散過程，也就是將結構化的數據逐步轉化為無結構的「噪音」，然後從「噪音」當中嘗試生成結構化的數據。此模型的核心想法是期望透過「噪音」來將文字與動作影像資料建立在同一個基礎上，而最終目標是獲得文本描述的動作影像。模型過程中就是透過將文本經過自注意力機制（Self-Attention）、交叉注意力機制（Cross-Attention）和前饋神經網路（Feedforward Neural Network, FNN）三層架構去計算出目標動作影像的「解除噪音」條件。接著再透過這項條件，將非結構化且經過加噪的動作影像解除噪音，最終即可獲得文本所描述的動作影像（Zhang et al., 2022）。在相同時期且也一樣使用擴散模型作為基礎來製作生成模型還有

Tevet, Raab, Gordon, Shafir, Cohen-Or, and Bermano（2022）的研究，在此研究當中作者特別使用幾何損失來改善運動的物理真實性。此外，此篇文獻作者也強調行動裝置管理（Mobile Device Management, MDM）使用的圖形處理器（Graphic Processing Unit, GPU）資源較少，在運算效率上優於其他同為擴散模型的運動生成模型。

接著來到 2023 年之後，大型語言模型逐漸發展成熟，像是 ChatGPT 為 2023 年帶來一股大型語言模型（Large Language Model, LLM）風潮。有研究提出使用 LLMs 加入 MotionGPT 當中，替動作生成模型帶來卓越性的發展，不僅提高動作生成結果的多樣性以及複雜性，更提升模型的靈活性（包含：生成序列長度、條件之下生成結果的控制程度）。T5 模型在計算過程中是透過分別將文本描述以及動作序列資料分別經過各自的編碼本（Codebook）去轉換成令牌（Token），接著將這些令牌再經過賦予權重的過程，告知 T5 模型不同特徵的重要性，最終即可獲得初始提示語所設定的動作序列（Jiang, Chen, Liu, Yu, Yu, and Chen, 2023）。不僅如此，也有研究嘗試使用 Large Language Model Meta AI（LLaMA）作為語言模型，在 MotionGPT 當中執行與 T5 相同任務的目標。本研究將文本描述與經過向量量化的動作序列資料，透過 LLaMA 預先訓練的方式去建立 MotionGPT（Zhang et.al., 2024）。與 T5 比較不同的地方是 LLaMA 所理解的文本描述並未經過轉換，但是 LLaMA 與 T5 同屬語言模型，在動作生成的彈性及豐富度都相較基於擴散模型的動作生成模型效果好。

因此，本研究期望將 T5 以及 LLaMA 兩種不同的語言模型去進行成效上的差異比較。在同為轉換器架構的語言模型為前提，比較以 LLaMA 作為語言模型的 MotionGPT（Zhang et.al., 2024）以及以 T5 作為語言模型的 MotionGPT（Jiang et al., 2023），通過三種不同任務難度的提示語輸入，同時提供給兩種 MotionGPT 進行動作生成並且進行結果比較。

參、實驗方法與結果

一、實驗流程

本次研究通過將兩個已存在的 MotionGPT 模型進行比較生成實驗，分別是 LLaMA 跟 T5，旨在深入分析和評估它們在動作影片生成方面的性能。實驗過程首先基於網路上對這兩個模型的共同評估分數（弗雷謝特起始距離〔Fréchet Inception Distance, FID〕、R-Precision 與 Diversity）進行起點定位，這些分數反映了模型的基礎能力和潛在的效果。

接著，經過全面訓練的這兩個模型被用來生成動作影片，這一步驟是實際應用和觀察的關鍵。生成的影片接受肉眼觀察，以評定模型的實際運作情況，尤其是在動作的自然度、流暢性以及對於不同類型動作的適應能力。

通過本研究可以更精確地理解各模型在特定條件下的表現，以及它們在處理複雜動作時的效能差異。此外，肉眼觀察的結果將提供直觀的反饋，進一步指導模型的優化和調整。透過這些步驟，本研究不僅提供了一種對現有技術的評估方法，同時也為未來動作生成技術的發展方向提供了寶貴的見解。

二、模型架構說明

（一）LLaMA

在 MotionGPT 中的內部包含多個重要的組件，這些組件共同作用，使得模型能夠將經過向量量化變分自動編碼器（Vector-Quantized Variational AutoEncoder, VQVAE）轉換來的條件任務序列有效地處理並且能夠更好的理解動作的數據。以下將深入探討 LLaMA 模型的各個組成部分，包括線性層、嵌入層、均方根正規化（Root Mean Square Normalization, RMSNorm）、合併線性層以及多層感知器（Multilayer Perceptron, MLP）（如圖 1），並解釋它們在處理語言數據時的功能和重要性。

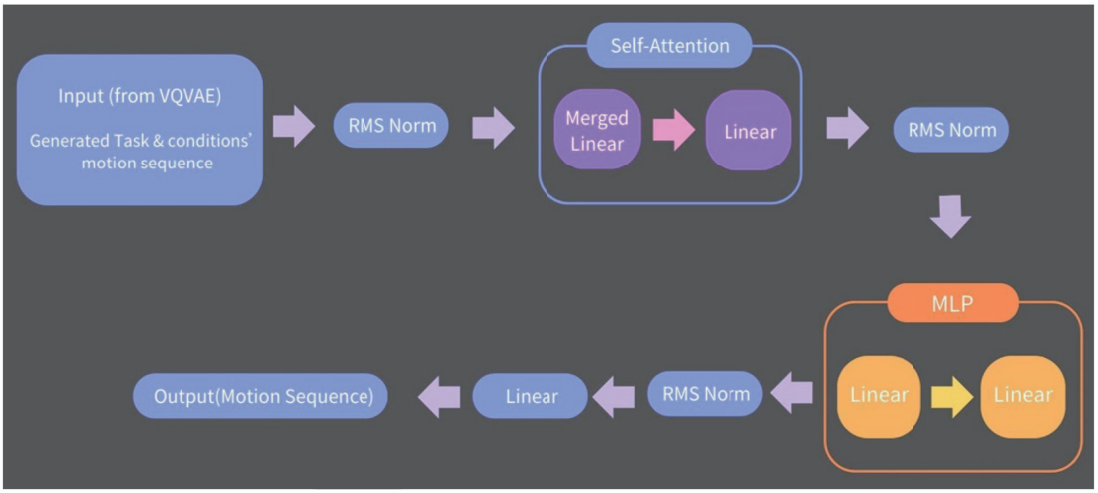


圖 1　LLaMA 模型生成架構圖

資料來源：本研究繪製。

1. 線性層（Linear）

線性層在 LLaMA 模型中扮演著關鍵的角色。線性層主要用於將輸入的特徵映射到另一個維度進行運算，並在最後將輸出的特徵還原至原本輸入的維度。這種映射和還原的過程使得模型能夠在不同的維度上理解和處理數據，從而捕捉到更多的特徵信息，增強模型的學習和預測能力。

2. 嵌入層（Embedding）

這一部分對於理解詞彙之間的關聯性至關重要。嵌入層的工作是將詞彙索引轉換為嵌入維度的表示形式，通過這種方式，模型能夠將單詞轉化為更豐富的數字表示，使得模型能夠捕捉到詞彙之間的細微差異和深層連繫。

3. RMSNorm

RMSNorm 是 LLaMA 模型中不可或缺的部分，它在模型的多個階段中發揮作用。具體來說，RMSNorm 會於潛在嵌入（Latent Embedding）之後、進入 MLP 之前（自注意力機制）以及最後輸出前進行。這三個階段的 RMSNorm 處理確保了生成的分布圖不會出現巨大的變化，從而使分布更加穩定。這有助於避免訓練過程中出現梯度消失或爆炸的問題，進而保證模型的穩定訓練和有效學習。

4. 合併線性層（Merged Linear）

它主要用於生成相對應的「鍵」和「值」來進行查詢的動作。在這裡，主要是針對重要的動作特徵進行強化訓練的部分。讓模型在訓練完之後，可以更明確回應本研究想要做的動作，從而提高其處理和預測的準確性。

5. MLP

在 LLaMA 裡，MLP 是通過數個線性層進行更深入的特徵轉換，進而幫助模型理解更複雜、更高階的特徵。這種深層的特徵學習使得 LLaMA 模型不僅能夠理解文本數據的表面信息，還能夠挖掘到更深層次的語義和語境信息，從而大大提高了模型對場景的轉換。

（二）T5

MotionGPT-T5 模型的工作開始於將動作令牌與文本單詞融為一體，形成混合序列輸入。這些序列經過語言編碼器（Language Encoders）的映射，將動作和文字特徵聚集於同一維度空間，形成統一的語言表示（如圖 2）。這種表示不僅捕捉了文字的語義信息，也蘊含了動作的細節，為模型提供了豐富的情境線索。

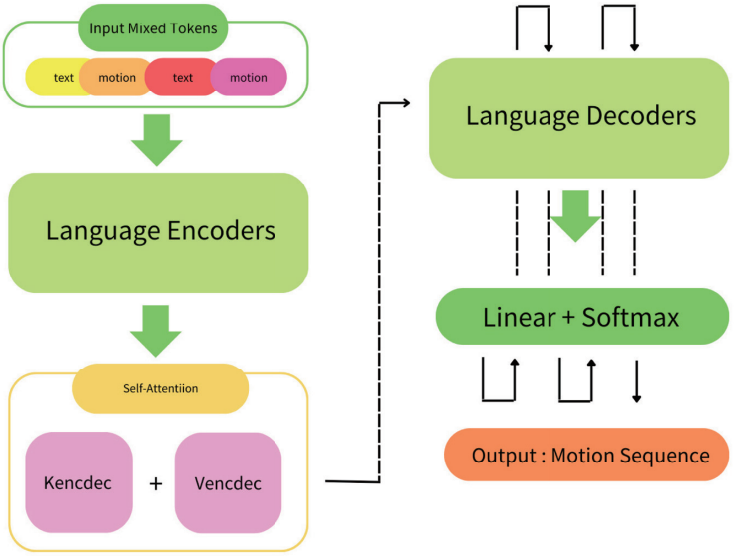


圖 2　T5 生成架構圖

資料來源：本研究繪製。

自注意力機制在 Kencdec 和 Vencdec 模組中發揮作用，如同 LLaMA 的合併線性層及 MLP 一樣，指導模型辨識哪些動作特徵至關重要，進而決定賦予這些特徵更大的權重。這種權重分配機制確保了模型在預測階段能夠專注於最具影響力的資訊，從而提高生成動作序列的準確性。

當關鍵動作特徵被標定後，語言解碼器（Language Decoders）連同線性加 Softmax 模組進行動作預測，將先前階段的資訊轉化為完整的動作序列。這些序列最終可轉換為動畫格式，應用於各種需要動作生成的領域。

三、兩個模型之間的比較

在這兩中不同的語言模型中，訓練和運作方式上具有顯著差異。首先，LLaMA 在訓練過程中採用了 RMSNorm 方法。這種方法通過計算機率分布的均方根來保持模型在學習過程中的穩定性。相對而言，T5 模型則使用了層歸一化（Layer Normalization, LayerNorm），這是一種基於常態分布的正規化方法，廣泛用於各種深度學習模型中，以幫助模型有效學習並適應不同的數據特性。在處理輸入數據方面，T5 的方法較為獨特，它會將文字和動作序列樣式通過一個編碼本轉換成統一的令牌格式。在模型內部，根據令牌出現的機率來決定接續的令牌應該是什麼，這種基於機率的方法使得 T5 在預測下一步時更精確。與此相反，LLaMA 在處理輸入時似乎是提供動作文字描述，然後根據預先訓練的結果來計算下一步的動作序列樣式。這表示 LLaMA 在預測下一步時，更多依賴於訓練過程中累積的數據和經驗，而非僅僅基於令牌格式的相似性。這種方法在處理複雜序列和模式時可能更有效，尤其是在需要深層次理解語境和細節的情況下。此外，LLaMA 和 T5 在編碼方式上也呈現出明顯差異。LLaMA 採用的是旋轉式編碼，這種編碼方式具有較大的彈性，使模型在判斷不同語境下的詞義時更為靈活和準確。旋轉式編碼通過改變詞向量的角度來表示不同的語義和語境，對於處理長距離依賴關係和複雜語義結構尤其重要。相比之下，T5 則採用了一般的絕對編碼，絕對編碼在處理序列數據時非常有效，它通過為每個位置分配一個固定的編碼來保持詞序的一致性。這種方法對於保持序列中詞彙的相對位置和關係非常有幫助，特別是在翻譯或文本生成任務中。然而，它可能不如旋轉式編碼那樣在處理長距離依賴和複雜語境時有效。

綜上所述，LLaMA 和 T5 在訓練方法、數據處理策略和編碼方式上各有特色，這些差異使它們在處理不同類型的語言任務時各有優勢。LLaMA 的旋轉式編碼和 RMSNorm 方法使其在理解複雜語境和長距離依賴關係方面更為出色，適合需要深度語境理解的任務。而 T5 的統一令牌格式和絕對編碼策略則在簡化輸入數據和提高模型泛化能力方面表現出色，使其在廣泛的 NLP 任務中，如文本翻譯、摘要生成或資訊檢索等方面非常有效。此外，T5 的這些特點也使其更容易與其他模型整合，提供了更大的靈活性，特別是在需要與不同類型模型或系統結合的場景中。因此，在選擇使用哪種模型時，應根據具體的應用場景和需求來做出決策。如果任務需要深度語境理解和處理複雜的語義結構，LLaMA 可能是更合適的選擇；而對於需要高度泛化能力和與其他系統整合的場景，T5 則可能更為適合。瞭解這些模型的特點和優勢，可以幫助本研究更有效地選擇和應用適合的技術，以解決特定的語言處理任務。

四、資料集說明

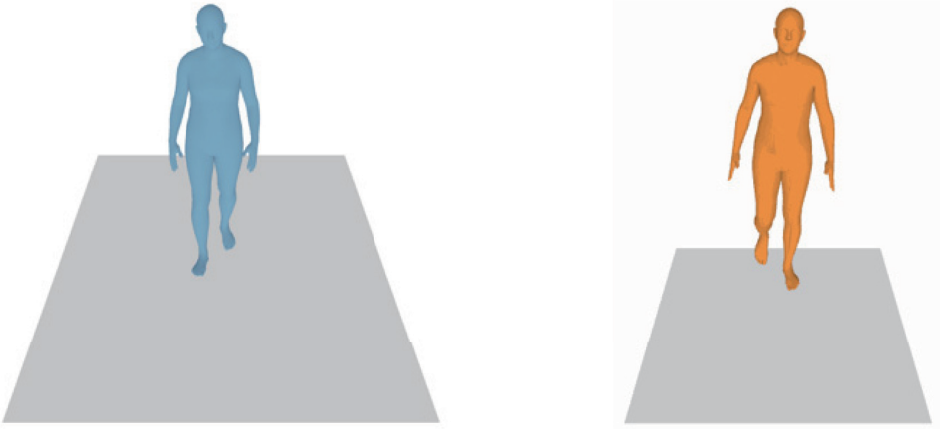
本研究室依據現今最大的動作序列資料集——HumanML3D。此資料集是依據 HumanAct12 跟 Amass 這兩個動作數據資料集所組成的大型數據集。它裡面內部包含 14,616 個動作和 44,970 個動作描述，這些都是由 5,371 個不同的單字組成，透過不同單字之間的排列組合，產生不同的描述及動作，透過這些描述及動作整理出一個 NumPy 格式的動作片段數據集。每一個資料都是一個動作附帶 3 ~ 4 個單句描述的 NumPy 檔案，每一個資料集都被降採樣至 20 fps，動作的持續時間為 2 ~ 10 秒。

五、生成結果及評估分數

本次研究給定三個不同的條件任務，其複雜程度由簡入深，分別是 “A person walks forward.”（如圖 3）、“A man kicks an object and walks forward.”（如圖 4）以及 “A man walks forward and picks ball with his left hand and runs back.”（如圖 5）。將這三個條件依序丟入已訓練好的模型，並生出動作影片。此外，本次的評估分數是以訓練後的模型所測試下來的分數（如圖 6），以下是生成出來的結果。

如表 1，從分數上可以觀察到 T5 模型只有在 R-Precision 指標的分數高於 LLaMA，且 LLaMA 產生的信賴區間上也大多小於 T5 模型所產生的。因此可以推斷在 LLaMA 模型生成的動作當中同質性較高，反而在 T5 模型生成的結果比較不集中。

此外，在眾多評估指標當中最直觀的 R-Precision 指標分數 T5 模型也是比 LLaMA 來的高，也就是說 T5 模型在針對動作關鍵字的理解會比 LLaMA 來的高，所以在範例當中，可以看到 T5 模型生成的結果會比較接近於參考文件描述的動作。換句話說，LLaMA 比較難以控制，會自行產生一些無關緊要的動作，而 T5 卻比較受控，可以根據輸入的文本和描述來產生動作。

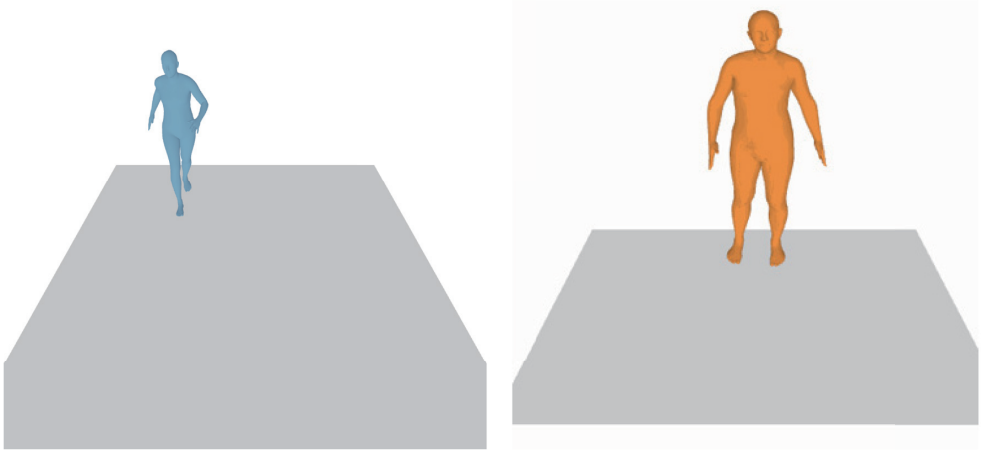


（a）LLaMA 模型生成結果　　（b）T5 模型生成結果

圖 3 “A person walks forward.” 的動作影片為簡單任務動作生成結果

資料來源：本研究整理。

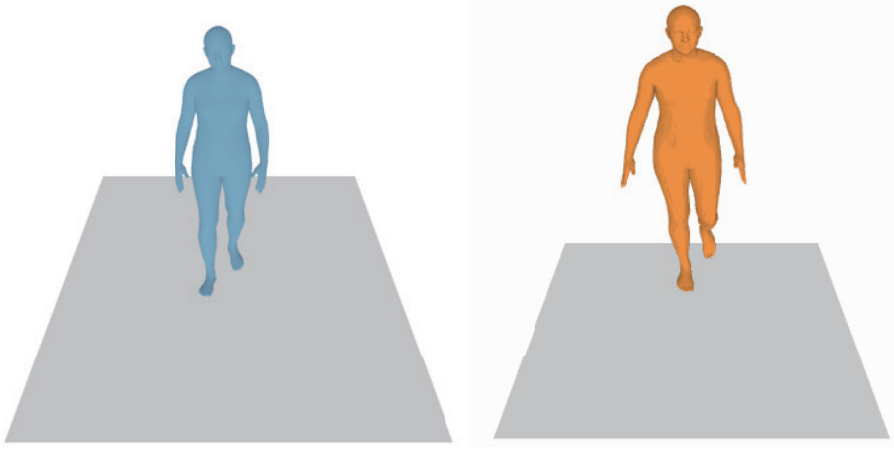
最後，在 FID 分數上雖然 T5 比 LLaMA 還要高，按照 FID 指標的定義來看 T5 的影像生成品質會比較差，但此分數只能代表生成結果與實際狀況（訓練資料集）的特徵相似度，因此如果將這分數以另外一個角度來看，可以發現 LLaMA 的 FID 分數比較低可能就是因為動作流暢度比較好，整個動作過程相當緩和順暢，T5 的生成結果當中則可以看到動作有可能會忽快忽慢，較不像是實際狀況會發生的。



（a）LLaMA 模型生成結果　　（b）T5 模型生成結果

圖 4　“A man kicks an object and walks forward.” 的動作影片為中等難度任務生成結果

資料來源：本研究整理。



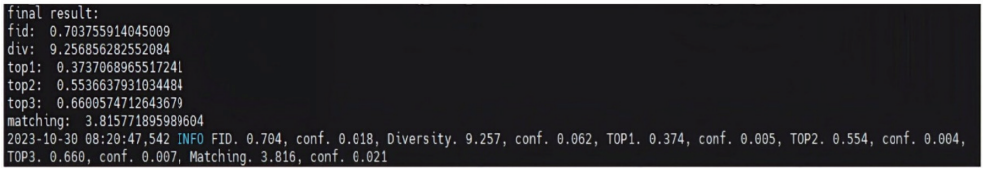
（a）LLaMA 模型生成結果　　（b）T5 模型生成結果

圖 5　“A man walks forward and picks ball with his left hand and runs back.” 的動作影片為較高難度任務生成結果

資料來源：本研究整理。

Metrics	Value
Metrics/Matching_score/mean	5.629596090316772
Metrics/Matching_score/conf_interval	0.018782616109086245
Metrics/gt_Matching_score/mean	2.966888725757599
Metrics/gt_Matching_score/conf_interval	0.007634163085416427
Metrics/R_precision_top_1/mean	0.22839439809322357
Metrics/R_precision_top_1/conf_interval	0.002369827349219169
Metrics/R_precision_top_2/mean	0.3481681004166603
Metrics/R_precision_top_2/conf_interval	0.0027248951591724553
Metrics/R_precision_top_3/mean	0.4327801734209061
Metrics/R_precision_top_3/conf_interval	0.0022465314859864984
Metrics/gt_R_precision_top_1/mean	0.5108836174011231
Metrics/gt_R_precision_top_1/conf_interval	0.002770680548583256
Metrics/gt_R_precision_top_2/mean	0.7023922443389893
Metrics/gt_R_precision_top_2/conf_interval	0.002737050585306169
Metrics/gt_R_precision_top_3/mean	0.7982219815254211
Metrics/gt_R_precision_top_3/conf_interval	0.0019250862270249476
Metrics/FID/mean	1.8739854454994203
Metrics/FID/conf_interval	0.05170700140008517
Metrics/Diversity/mean	9.093709897994994
Metrics/Diversity/conf_interval	0.11440127472780136
Metrics/gt_Diversity/mean	9.422248220443725
Metrics/gt_Diversity/conf_interval	0.09050584855098658
Metrics/MultiModality/mean	5.551373720169067
Metrics/MultiModality/conf_interval	0.20367407273042049

（a）T5



（b）LLaMA

圖 6　訓練模型後的評估

資料來源：本研究整理。

表 1　兩個模型統整分數的比較

評估分數	LLaMA	T5
FID	0.704	1.874
Diversity	9.257	9.090
R-Precision Top 1	0.374	0.228
R-Precision Top 2	0.554	0.702
R-Precision Top 3	0.660	0.798

註：FID：弗雷謝特起始距離（Fréchet Inception Distance）。
資料來源：本研究整理。

肆、結論

綜合以上的分析，T5 模型在根據文字描述產生動作方面，相比 LLaMA 模型展現出更高的準確性。這一優勢主要體現在 T5 生成的結果與原文本描述的一致性上。相對於 LLaMA、T5 在關鍵動作的呈現上不會引入過多不必要的串接動作，使得產出的內容更加精煉且直接。例如，在觀察前三個案例中 T5 模型生成的圖像互換格式（Graphics Interchange Format, GIF）檔案時，可以清楚地看出其與文本描述的高度吻合，這些案例證明了 T5 在捕捉和實現文本描述的動作細節方面的優越性。與 LLaMA 相比，T5 模型展示了更佳的文本到動作轉換能力，這一點在直觀的視覺呈現中得到了充分的體現，因此，T5 模型在這方面的表現值得肯定。

參考文獻

黎氏玉幸，2023，《靜止影像生成動畫、地圖藝術風格影片產生與影片重新排序而產生新動畫》，國立成功大學資訊工程學系博士論文。

Guo C., Mu Y., Javed M. G., Wang S., and Cheng L., 2023, “MoMask: Generative Masked Modeling of 3D Human Motions,” https://doi.org/10.48550/arXiv.2312.00063 (accessed January 10, 2024).

Jiang B., Chen X., Liu W., Yu J., Yu G., and Chen T., 2023, “MotionGPT: Human Motion as a Foreign Language,” https://doi.org/10.48550/arXiv.2306.14795 (accessed January 10, 2024).

Tevet G., Raab S., Gordon B., Shafir Y., Cohen-Or D., and Bermano A. H., 2022, “Human Motion Diffusion Model,” https://doi.org/10.48550/arXiv.2209.14916 (accessed January 10, 2024).

Zhang M., Cai Z., Pan L., Hong F., Guo X., Yang L., and Liu Z., 2022, “MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model,” https://doi.org/10.48550/arXiv.2208.15001 (accessed January 10, 2024).

Zhang Y., Huang D., Liu B., Tang S., Lu Y., Chen L., Bai L., Chu Q., Yu N., and Ouyang W., 2024, “MotionGPT: Finetuned LLMs are General-Purpose Motion Generators,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7), 7368-7376. doi:10.1609/aaai.v38i7.28567