國立高雄科技大學

智慧商務系

碩士論文

結合大型視覺語言模型與多模態檢索技術於電競賽事影片精華偵測之研究：以《英雄聯盟》為例

An Applied Study on Esports Highlight Detection in Match Videos Using Large Vision-Language Models and Multimodal Retrieval Techniques: A Case Study of League of Legends

研究生：王聖維

指導教授：謝文川　博士

中華民國一一四年六月

結合大型視覺語言模型與多模態檢索技術於電競賽事影片精華偵測之研究：以《英雄聯盟》為例

An Applied Study on Esports Highlight Detection in Match Videos Using Large Vision-Language Models and Multimodal Retrieval Techniques: A Case Study of League of Legends

研究生：王聖維
指導教授：謝文川　博士

國立高雄科技大學

智慧商務系

碩士論文

A Thesis Presented to
Department of Intelligent Commerce
National Kaohsiung University of Science and Technology in Partial Fulfillment of the Requirements
For the Degree of Master of Science
In
Intelligent Commerce

June. 2025

Kaohsiung, Taiwan, Republic of China

中華民國一一四年六月

結合大型視覺語言模型與多模態檢索技術於電競賽事影片精華偵測之研究：以《英雄聯盟》為例

研究生：王聖維　　指導教授：謝文川　博士

國立高雄科技大學智慧商務系碩士班

摘 要

本研究旨在探討大型視覺語言模型（Large Vision-Language Models, LVLMs）於電競賽事影片中的應用，聚焦於兩項核心任務：片段檢索（Moment Retrieval, MR）與精彩片段偵測（Highlight Detection, HD）。考量到多模態學習在自然語言與視覺理解領域的高度發展，本研究導入 LVLM 作為弱監督語句生成工具，期望藉由其語意對齊與描述生成能力，減輕人工標註成本，並提升模型的跨模態學習效果。研究以《英雄聯盟》職業賽事影片為資料來源，透過人工標註建立基礎語句庫，並採用 Gemini 自動生成事件敘述語句，以建構弱監督預訓練資料。模型架構部分採用 VideoLights，融合 SlowFast 動作特徵與 CLIP、BLIP 的語言模態，實施多組變項操弄實驗，包括語句型態、是否進行弱監督預訓練、模態特徵組合，以分析各因素對 MR 與 HD 任務表現的影響。實驗結果指出，複雜敘事語句在多數 MR 指標上優於簡單語句，特別是在 R@0.3、R@0.5 與 mAP 表現上展現更佳語意對齊能力，而在 HD 任務中亦有助於提升對於高階顯著事件的辨識精度。進一步的預訓練比較則顯示，導入 LVLM 自動生成語句進行弱監督預訓練的模型，於 MR 任務之高門檻指標大多優於無預訓練模型，顯示語義引導對模型泛化能力有所助益。模態消融實驗亦揭示 CLIP 特徵在部分實驗設定中未必為必要模態，甚至移除後模型表現有所提升；而 BLIP 特徵則在語句語意對齊上展現穩定貢獻，為模型核心語言模態之一。綜合而言，本研究建立一套以 LVLM 為基礎之半自動化語料建構流程，並實證其在電競影片精華檢索任務中的可行性與效能貢獻。研究結果可為未來多模態檢索模型設計提供參考，亦具備推廣至其他高語意結構場域（如運動賽事、醫療影像等）的潛力與價值。

關鍵字：大型視覺語言模型、片段檢索、精彩片段偵測、弱監督學習

An Applied Study on Esports Highlight Detection in Match Videos Using Large Vision-Language Models and Multimodal Retrieval Techniques: A Case Study of League of Legends

Student：Sheng-Wei, Wang Advisors : Dr. WC, Hsieh

Graduate Institute of Information Management

National Kaohsiung University of Applied Sciences

ABSTRACT

This study investigates the application of Large Vision-Language Models (LVLMs) in esports video analysis, focusing on Moment Retrieval (MR) and Highlight Detection (HD) tasks. To reduce manual annotation costs and enhance cross-modal learning, we adopt LVLMs as a weak supervision tool by using Gemini to automatically generate narrative queries and BLIP to assess saliency scores. Using League of Legends tournament videos as data, we construct both manually annotated and LVLM-generated datasets. The base model, VideoLights, integrates SlowFast video features and CLIP/BLIP textual features. We conduct experiments across three dimensions: query type (simple vs. complex), pretraining presence, and modality ablation. Results show that complex queries improve semantic alignment in MR tasks and enhance performance in HD tasks under stricter evaluation thresholds. Weakly supervised pretraining further improves high-precision metrics such as R1@0.7 and mAP@0.75. Modality ablation reveals that CLIP features are not always necessary, while BLIP features consistently benefit text-video alignment. Overall, our semi-automated LVLM-based data construction pipeline proves effective in esports highlight detection and may generalize to other domains such as sports or medical video analysis.

Keywords：Large Vision-Language Models、Moment Retrieval、Highlight Detection、Weakly Supervised Learning

誌　謝

本論文能順利完成，絕非我一人之力，而是眾多師長、家人、朋友在背後默默支持與鼓勵的結果。在此，謹向所有在我求學及研究歷程中給予幫助與指導的人致上最誠摯的感謝。

首先，我要衷心感謝我的指導教授－謝文川教授。教授在學術研究上始終秉持嚴謹與求真的態度，從研究主題的確立、方法的選擇，到實驗的執行與論文的撰寫，都給予我細心的指導與寶貴的建議。教授不僅是我學術上的引路人，更在遇到瓶頸與困難時，耐心開導、鼓勵我堅持下去。教授的專業知識、治學精神以及對研究的熱情，將成為我日後學習與工作的楷模。

同時，我也要感謝在論文口試中給予我寶貴意見與建議的口試委員－陳俊卿教授及張添香教授。委員們透過專業的提問與中肯的指導，幫助我檢視研究中可能的不足與改進方向，使本論文能更加完整與嚴謹。他們的建議不僅提升了研究的品質，也拓展了我對研究領域的更深理解。

其次，感謝我的家人。他們是我最堅強的後盾，無論是生活上的照顧，還是心靈上的支持，都讓我能夠無後顧之憂地專注於研究。當我因研究進度感到焦慮或徬徨時，家人總是給予我安慰與鼓勵，提醒我要保持信心與毅力。他們長久以來的付出與包容，讓我能一步步走到今日。

我也要特別感謝我的女朋友。在這段研究與寫作的過程中，她不僅是陪伴我、鼓勵我的對象，更是在我最感疲憊與迷惘時，給予我力量與溫暖的人。無論是熬夜寫作的孤獨時刻，或是面對挫折時的低潮，她都始終在旁守護，讓我重新振作、繼續前行。能夠擁有她的支持與陪伴，是我最珍貴的幸運。

此外，我要感謝在研究過程中曾經幫助過我的洪立穎學長與陳政宏同學。無論是課題討論上的啟發、技術操作上的指導，或是日常交流中所帶來的靈感與建議，這些互助的時光都使我獲益匪淺。在他們的幫助下，我不僅提升了專業能力，更深刻體會到合作與分享的重要性。這段共同奮鬥的經歷，將會是我研究生涯中最寶貴的回憶之一。

最後，感謝所有在這段求學旅程中曾經給予我協助的人。雖無法一一列名，但每一份鼓勵、每一個幫助，我都銘記在心。本論文的完成，凝聚了許多人的心力與支持。

謹以此誌，向所有幫助、支持與陪伴過我的人致上最深的謝意。

目錄

表目錄

圖目錄

壹、緒論

一、研究背景

在人類的歷史中，體育競技一直是我們生活的重要組成部分。傳統體育競技一直是人們追求激情和競爭的重要途徑，然而，近年來，隨著數位媒體與即時串流技術的進步，電子競技（Esports）迅速崛起，成為新興主流娛樂產業之一。電子競技結合了遊戲操作技巧、戰術理解與團隊協作，不僅吸引大量年輕觀眾，也逐漸發展出職業聯賽與全球化的賽事體系。

《英雄聯盟》（英語：League of Legends，簡稱 LoL）是由 Riot Games 開發及發行的一款 5v5 多人線上戰鬥技術型（MOBA）遊戲，英雄聯盟在電競比賽上創下了許多其他遊戲難以超越的成就，例如，根據 Riot Games 公布的數據《英雄聯盟》在 2024 世界大賽總決賽中刷新了觀看紀錄，同時觀看人數高達 694 萬人，成為電競史上最受歡迎的賽事之一。

圖 1　2024 年世界大賽總決賽觀賽人數

（資料來源：**https://game.ettoday.net/article/2847638.htm**）

隨著賽事熱度的提升，衍生出的「賽事精華剪輯影片」成為觀眾回顧比賽與快速掌握重點的主要媒介。無論是官方頻道（如 LCK Global、LCS），或是經授權的內容創作者（如最強聯盟、Onivia），都會在賽後上傳經剪輯的精華片段，濃縮團戰、擊殺、目標物爭奪等精彩畫面，滿足觀眾碎片化時間下的收視需求。這樣的影片不僅讓錯過直播的觀眾能夠追趕比賽的精華，同時也為那些已經觀看了直播的

觀眾提供了一個回顧和分享的機會。這些精華剪輯影片的上傳，進一步擴大了電子競技比賽的觀眾群體，使更多人有機會欣賞和分享這些精彩的比賽時刻。

同時在當今的媒體消費習慣中，觀眾對於觀看影片的方式有了顯著的改變。相比於以往，短影音對於人們往往擁有更強的吸引力（Liu et al., 2021），人們現在更加傾向於觀看幾分鐘長度的直播精華片段，而不是完整的 30 分鐘或更長時間的直播影片。這種轉變主要源於短片更加迎合了人們快節奏的生活方式（Wang, 2020），與傳統完整轉播相比，3 至 10 分鐘內的賽事精華能更迅速傳達比賽亮點，符合現代觀眾追求「高資訊密度」與「時間效率」的內容偏好。因此，短影片成為了一種高效的娛樂方式，使觀眾能夠在短時間內獲得精彩的內容精華（Wright, 2017）。

這對於如英雄聯盟這樣的遊戲而言，提出了一個重要的挑戰。一場英雄聯盟的比賽通常持續 20 至 50 分鐘，而精華影片的長度則通常在 5 至 15 分鐘之間。在短短的幾分鐘內提煉出一場比賽的精華，並在不損失比賽完整性的前提下，讓觀眾將注意力集中在重要事件上，成為了一項重要的課題。

然而，現階段這類精華影片的生成仍高度依賴人工剪輯，從選擇片段、配音到節奏編排皆需耗費大量時間與人力。面對內容產製規模與速度的雙重壓力，如何導入人工智慧技術，自動化識別與剪輯精華片段，成為亟待解決的挑戰。

二、研究動機

隨著電子競技的蓬勃發展，特別是在《英雄聯盟》（League of Legends）這類大型多人線上戰略遊戲中，人們對於比賽的興趣和關注度持續上升。一場典型的《英雄聯盟》比賽通常持續 25 至 50 分鐘，而一天內的比賽數量可從 4 場到達 9 場不等。對於普通觀眾來說，由於時間有限，很難完整觀看每場比賽的直播或錄影。因此，比賽的精華剪輯成為了觀眾了解賽事重點和精彩瞬間的主要方式。這些精華片段不僅展現了選手的高超操作和在極限狀態下的反應能力，也節省了觀眾的時間，讓他們能夠快速獲得賽事的重點。

然而，傳統上這些比賽精華的製作需仰賴大量人力，從反覆觀看到手動選取精彩片段，不僅耗時也不易規模化。因此，如何透過深度學習與大型視覺語言模型（Large Vision-Language Models, LVLMs）技術，自動識別比賽中的關鍵事件，成為一個值得深入探討的方向。

三、研究目的

隨著深度學習技術與多模態模型的快速發展，越來越多研究致力於從長影片

中自動擷取重點事件。其中，精彩片段偵測（Highlight Detection, **HD**，自動識別影片中具高觀賞價值的精華時刻）與片段檢索（Moment Retrieval, **MR**，依據文字描述從長影片中準確定位相符的時間片段）是目前主流的影片精華檢測任務設定，已在一般生活影片、體育競技等資料集上取得良好成果。然而，這些方法在應用至語意密度更高、事件邊界複雜的《英雄聯盟》（League of Legends）電競賽事中，其表現與可行性仍有待驗證。

本研究旨在探討：現有的 HD/MR 模型架構，是否能有效應用於《英雄聯盟》比賽影片中自動辨識精華片段；並進一步探討：是否能透過大型視覺語言模型（Large Vision-Language Models, LVLMs，指能同時理解與生成影像與文字內容的多模態預訓練模型）所生成的圖文對應敘述，來取代部分人工標註資料的負擔，實現弱監督訓練的可行性。

為此，本研究選用在多模態影片檢索任務上表現優異的 VideoLights 模型（多模態影片檢索模型，結合影片與文字特徵進行跨模態匹配）作為檢測框架，並設計雙階段訓練流程。研究共收集 46 場《英雄聯盟》職業賽事影片，其中預訓練階段使用 Gemini 2.5 Flash Preview 模型（Google 開發的大型多模態模型，可同時處理影像、文字與其他輸入）為 26 場比賽影片生成對應敘述，標註超過 5000 筆事件構成弱標資料集，微調階段則使用人工標註之精華語意片段資料，涵蓋 20 場比賽影片並標註超過 600 筆事件，進行模型調整與效能測試。此資料規模兼顧了訓練多樣性與標註品質，並有助於驗證多模態模型在高語意密度任務下的適應性。透過上述設計，本研究目標在於：

(一).評估 HD/MR 任務模型在語意密度高、專業術語豐富的《英雄聯盟》比賽影片上的表現

(二).檢驗由 LVLM 所生成的合成敘述資料，是否能有效支援弱監督訓練並減少標註負擔。

(三).分析上述方法於高結構語境下的適應能力與挑戰，為未來多模態模型於領域特化任務中的應用提供經驗與反思依據。

四、論文架構

第一章將聚焦於電子競技與直播精華的關聯性，探討英雄聯盟這款線上遊戲以及觀眾的觀看傾向，同時闡述研究動機目的。第二章將專注於直播精華檢測領域的文獻探討，特別是介紹現今在這一領域的研究，並對 LVLM 技術進行概述，說明其在直播精華檢測中的應用情況。接著，第三章將呈現研究方法，包括資料集的

收集、來源、預處理方法以及所選用的模型架構。第四章將展示實驗結果與分析，根據第三章中提出的研究方法，進行模型的訓練並對實驗結果進行詳細分析。最後，第五章將提出結論與未來展望，對分析結果進行總結，並探討未來可持續的研究方向以及可能性。

貳、文獻探討

本章將進行文獻探討，並分為四個主要部分。首先，在 2.1 節中，我們將回顧現有研究，探討精華檢測中使用的特徵及相關技術。接著，在 2.2 節中，我們將介紹現有的直播遊戲精華模型如何對影像資料進行標註。隨後，在 2.3 節中，我們將探討 Large Vision-Language Models（大規模視覺-語言模型）及其在精華檢測中的應用，並評估其潛在優勢與挑戰。最後，在 2.4 節中，我們將對本章內容進行總結。

一、片段檢索（MR）與精彩片段偵測（HD）

(一).片段檢索（MR）

近年來，隨著多模態學習與影片理解技術的快速發展，片段檢索逐漸成為視覺語言研究領域中的重要子任務。此任務的核心目標是基於自然語言描述，從一段未剪輯的影片中準確地定位與語意相符的時間區段，亦即給定影片及一段文字描述後，模型需輸出對應的時間範圍$[t_{start}, t_{end}]$，以精確反映影片中符合該描述的片段。

片段檢索的概念最早源於 Regneri et al.（2013），該研究首次探討將自然語句與影片中的動作進行語義對齊。然而，此任務正式獲得命名並引發廣泛研究關注，則始於 Gao et al.（2017）的研究。該研究首次明確定義了片段檢索任務，並建立了專門用於任務評估的基準資料集 Charades-STA，為後續的研究提供重要且系統性的基礎。

隨著相關技術的發展與進步，片段檢索已逐漸成為衡量影片理解能力及跨模態對齊能力的重要指標，並廣泛應用於各種多媒體情境當中。具體而言，其主要的應用包括：

1.精華片段檢索：例如體育賽事中，快速定位關鍵事件。

2.影片內容檢索系統：讓使用者可透過自然語言查詢特定影片片段，如電影中的特殊場景或重要事件。

3.教育影片之知識索引與檢索：協助學習者迅速找到與特定主題相關的教學內容。

4.智慧監控與行為分析，例如透過文字描述快速定位特定行為發生的時段。

然而，儘管片段檢索 已於一般生活情境中取得顯著成果，其於高語意密度且結構複雜的應用場域，如電子競技（Esports）的研究與實作仍相對有限。以《英雄聯盟》為例，其比賽影片具有極高的事件頻率與角色互動密度，且事件常含多層語

意，例如「選手職位」、「技能搭配」、「作戰地點」與「戰術轉折」等。此類高度結構化語境對 MR 模型的語意理解與跨模態對齊能力帶來極大挑戰。

本研究即以此為切入點，採用 VideoLights 為 MR 架構基礎，嘗試將其應用於《英雄聯盟》精華片段的自動擷取任務中，並進行多層次實驗分析。具體而言，本研究以 Gemini LVLM 生成的文字敘述作為弱監督資料來源，並與人工撰寫敘事查詢與標記片段構成的強監督資料進行比較。此外，特別設計了簡單與複雜兩種敘事查詢策略，觀察不同語意層級對擷取效能的影響。透過這樣的實驗設計，不僅能驗證既有 MR 模型於電競語境的遷移能力，也能探索高語意密度任務下的資料設計準則與模型應用邊界。

(二).精彩片段偵測 (HD)

近年來，隨著數位媒體與視訊平台的迅速普及，精彩片段偵測已成為影片分析與理解領域中的核心任務之一。其目標在於自動識別影片中的精彩片段，例如體育賽事中的進球瞬間或電競比賽中的關鍵擊殺場面，從大量未剪輯的影片中有效擷取高顯著性事件，協助觀眾快速瀏覽或提升後續的影片再製效率。

HD 任務的起源可追溯至傳統的影片摘要技術，早期主要依賴視覺或聲音訊號的變化來判斷片段的重要性（Sun et al., 2014）。隨著深度學習的快速進展，越來越多研究開始運用卷積神經網路（CNN）、時間序列網路（LSTM）以及視覺注意力機制（attention）等架構進行學習式片段選擇（Mahasseni et al., 2017），使得模型能從大規模資料中自動學習與「精彩片段」相關的時序與語意特徵，顯著提升偵測的準確性與泛化能力。

HD 技術可依據標註需求區分為三類：監督式、弱監督式與非監督式。監督式 HD 雖然精度較高，但仰賴大量具細節標註的資料，標註成本與時間成本皆相當高昂。為此，近年有研究提出弱監督學習方法，藉由僅包含精華片段的正樣本資料與未標記資料，結合半監督訓練機制以降低人工成本。例如，Lei et al.（2021）便透過影片中自動語音辨識（ASR）標題進行弱監督訓練，輔以多步驟方法來提升偵測表現。更進一步地，近期研究也開始導入大型視覺-語言模型（Large Vision-Language Model, LVLM），如 Paul et al.（2023）嘗試利用 LVLM 進行自動語意標註，以取代人工標記，不僅有效提升標註一致性，亦可快速擴展至大規模資料集。HD 任務與其他影片理解任務（如事件偵測、影片摘要、片段檢索等）有密切關聯。其中，以 VideoLights 模型為代表的最新研究，採用多任務學習策略將 MR 與 HD 結合，並透過 FRA 模組、Bi-CMF 注意力機制與 Uni-JFM 任務融合設計，促進跨任務語意互補，提升精華片段的偵測品質（Paul et al., 2023）。

在應用層面，HD 技術在短影音盛行的當代媒體環境中具有極高實用價值，尤

其在電競領域，能有效降低人工剪輯成本、提升觀賞體驗與內容傳播效率。然而，由於電競比賽片段高度結構化、事件密度極高，且視覺資訊複雜（如角色分工、地圖位置、戰術目標等），傳統 HD 模型常難以充分理解其中語意結構與觀眾關注焦點。因此，本研究針對《英雄聯盟》賽事影片，探討現有 HD 模型於高結構語境中的適應性，並進一步分析是否可透過語意強化的敘事標註或 LVLM 所產生的弱監督資料，有效提升模型於真實電競語境下的表現。

二、電競影片精華檢測與事件檢索的標註方法回顧

在電競賽事影片中精彩片段偵測（HD）與片段檢索（MR）的高品質資料標註是關鍵挑戰。早期方法多依賴人工將長達數十分鐘的比賽影片中精華時刻標記出來，或定義哪些瞬間屬於「精彩」；然而全人工標註成本極高且耗時。為降低人力負擔，研究者們探索了各種標註策略，包括人工標註、人機協作、群眾智慧（眾包）以及弱監督學習等，並在資料效率、成本投入與標註精確度上各有權衡。

遊戲精華（Game Highlights）是指在電子遊戲比賽或遊戲錄影中，那些最引人入勝、最具戲劇性、最令人驚嘆的時刻或片段。這些時刻通常包括了玩家的高超操作、關鍵時刻的轉折、戰術的精彩展示以及遊戲內的意外事件等。精華片段的目的是將整場遊戲或比賽中最重要、最吸引人的部分濃縮成短小而精煉的片段，讓觀眾在短時間內獲得最大的觀賞樂趣和感受。

(一).傳統人工標註與人機協作

1.人工標註：

人工逐段標記精華片段能提供精確且具專業判斷的標籤，但需要耗費大量時間和人力。在電競領域，人工標註往往由專家或有經驗的玩家完成，其優點是準確度高、標註的一致性較易控管，缺點則是資料量受限，成本昂貴且不易大規模擴充。早期一些體育/電競影片摘要研究多採人工標記每個可能精華鏡頭是否屬於精華，作為模型的訓練範本，Chen et al.（2022）使用人工標註的方式，由多位專家對直播影片進行精華片段標註，每個影片的標註耗時約為影片長度的 5 倍，並透過多重驗證確保資料品質。此方式資料質量高，但「資料效率」（單位人工能標記的有效資料量）偏低，不利涵蓋電競影片中繁複多樣的情境。

2.人機協作：

為了在維持標註品質的同時降低人力，部分方法嘗試半自動化流程：例如先由演算法預選疑似精華的片段，再由人工作最後確認與修正。Song（2016）將資料標註通常被分為兩個階段進行，以提高標註效率與標籤品質，透過人機協作的場景分

類與群眾外包的精華片段標註方法以從電競直播中檢測精華。而 Chu & Chou（2017）提出的方法針對直播檢測精華檢測分為 MR 和 HD 兩部分。首先是 MR，透過辨識螢幕上顯示的指定文字訊息來偵測遊戲事件，並透過研究人員手動標記事件及其對應的時間戳記，以評估事件偵測的效能。其次是 HD，由研究人員檢查每個遊戲影片，並手動標記事件及其對應的時間戳。這種人機協同標註方式能顯著減少人工檢視的總時長，提高標註效率。這種做法可以結合模型的初步過濾與專家的精細判斷。優點是節省時間且仍保持不錯準確度；缺點是仍需一定人工參與，且模型預選階段若錯失關鍵時刻可能導致遺漏。

(二).群眾智慧標註與弱監督訊號

1.群眾外包：

隨著直播與影片分享平台的興起，研究者發現可以利用玩家社群自發產生的資料作為標註來源。例如，Song（2016）將資料標註通常被分為兩個階段進行，以提高標註效率與標籤品質，透過人機協作的場景分類與群眾外包的精華片段標註方法以從電競直播中檢測精華。Ringer et al.（2022）採用了弱監督學習的方法，並使用了兩種資料集進行精華片段的標註，正樣本標記是從多個影片分享平台（例如 YouTube）上由人工編輯並發布的英雄聯盟精華影片中收集而來，未標記樣本的資料集則是包含來自各種比賽的英雄聯盟專業廣播的回放影片。同樣的，Wang et al.（2019）提出了一種無監督的方法來自動生成遊戲"Honor of Kings"的精華片段，他們從"Penguin E-sports"平台收集了長直播影片和精華影片，並將剪輯過的精華影片作為正樣本，長影片作為負樣本以此來避免了手動標註的需要。Liaw & Dai（2020）使用群眾外包（crowdsourced）的方式來標記直播影片中的精華片段，這些片段是由粉絲在他們覺得有趣的時刻所產生的，並透過設定時間和觀看次數的門檻來過濾雜訊，以此建立資料集並評估模型的效能。這類群眾智慧策略在資料效率和成本上具有巨大優勢，能動員大量現有資料，零成本取得標註，但其精確度取決於群眾提供內容的可靠性。例如精華剪輯可能因製作者偏好有所側重，或長影片中也存在未被剪進精華影片的精彩時刻（屬於訓練負樣本的噪聲）。

2.弱監督訊號：

除了直接利用社群剪輯結果，許多方法嘗試發掘間接訊號作為精華標記依據，這可視為弱監督學習的一種。Chu & Chou（2017）的早期研究即屬於此類：他們並未大規模人工標記精華片段，而是結合遊戲內事件（如擊殺提示文字）、視覺效果（畫面中角色數量、畫面動態等手工特徵），由於遊戲事件（如畫面出現「擊殺」字樣）可自動偵測，他們實際上迴避了逐幀人工標記，而是把這些事件當作精華線索來弱標註影片。同時，他們的模型仍需一些資料調校，Chu & Chou（2017）在實

驗中僅使用了一個賽事（2014 世界賽）24 場比賽的影片來訓練與測試資料規模相對小，說明此類基於弱訊號的模型對資料量依賴降低，但也可能因樣本有限而影響泛化能力。

另一類常用的弱監督訊號是觀眾反應。許多電競直播平台允許觀眾即時發送聊天訊息、彈幕等，這形成了影片的一條文字流，可用來推測哪些時刻引發大量討論或激動反應。Han et al.（2019）率先將純文字聊天訊息用作精華預測依據，採用雙向 GRU 神經網路處理聊天串流，來學習何時是精華。他們在有監督設定下，用帶標註的精華時刻去訓練 GRU，使模型從觀眾留言頻率和內容中捕捉精華片段的模式，結果顯示單靠文字訊號即可取得優於傳統視覺模型的精華預測效果。Liaw & Dai（2020）則進一步提出長短期注意模型（LSTA）來分析聊天訊息。他們利用觀眾聊天串流作為線索，透過注意力機制在時間序列上找出關鍵訊息區段，幫助定位比賽中的精彩時刻。這兩項研究充分運用了弱標記訊號：聊天訊息是觀眾集體智慧的體現，不需要人工逐段註解影片內容，只需記錄每時刻的聊天流量或關鍵詞。然而，聊天訊號作為標註也有不足，並非所有精彩畫面都會立即引發大量訊息（有些精彩操作可能觀眾來不及反應），且聊天內容可能包含噪聲（例如與比賽無關的對話）。因此模型需要能抗噪並識別哪些聊天高峰真的是比賽高潮。總的來說，透過弱監督整合這類群眾反應數據，大幅提升了資料利用效率和降低成本，但在精確度上通常需結合其他線索或更複雜模型來彌補純文字訊號的局限。

(三).LVLM 自動標註的創新及優勢

面對上述傳統標註方式的種種權衡，本研究引入大型視覺語言模型（LVLM）來自動生成描述句與提取語意特徵，提供了一條減少人工成本的新路徑。具體而言，我們讓預訓練的視覺-語言模型充當“自動解說員”，為影片片段產生語句描述，從中提取豐富的語意表示作為精華檢測線索。這種方式相當於用模型替代人工撰寫精華摘要或標籤，作為該片段的語意標註。相比傳統逐幀標記或僅用簡單信號，我們的方法具有明顯效率與效果優勢。

三、大型視覺語言模型（Large Vision-Language Models, LVLM）

(一).LVLM 的發展歷史與代表性模型概述

大型視覺語言模型（Large Vision-Language Models, LVLM）是結合計算機視覺與自然語言處理的大型預訓練模型，它們能夠同時理解影像與文本資訊，在多種下游跨模態任務中展現出色表現。LVLM 的發展歷程中出現了多個里程碑式的代表

性模型，包括 OpenAI 提出的 CLIP、Salesforce 提出的 BLIP 系列（BLIP、BLIP-2），以及 Google DeepMind 開發的 Gemini 等。這些模型的架構設計、訓練方法、跨模態對齊技術各具特色，並被廣泛應用於圖像字幕生成、跨模態檢索、視覺問答等任務。下文將依時序介紹這些關鍵模型及其技術脈絡。

OpenAI 的 CLIP（Contrastive Language-ImagePre-training）是大型視覺語言模型的開創者之一。CLIP 首次證明了利用大規模網路圖像-文本對進行對比學習可以學得通用且可遷移的視覺語言表徵（Radford et al., 2021）。CLIP 的模型架構由一個圖像編碼器和一個文本編碼器組成，透過對 400 億組（image, text）圖文配對資料的對比預訓練，將圖像和文字映射到共同的嵌入空間中。訓練目標是讓正確對應的圖文對在該空間中距離更接近，從而實現跨模態對齊。這種對齊機制使 CLIP 在無需針對特定資料集微調的情況下，就能直接以自然語言提示進行圖像分類和檢索，被稱為零樣本學習能力[1]。例如，只需提供目標類別的文字描述作為提示，CLIP 即可在 ImageNet 等資料集上達到與有標註訓練的模型相當的辨識準確率。CLIP 的問世為視覺語言模型奠定了重要基石：後續許多多模態模型都以 CLIP 作為預訓練基礎或參考範式，透過其學到的圖文對比對齊特徵來提升跨模態任務表現。

儘管 CLIP 展現了強大的零樣本遷移能力，但早期視覺語言模型在「理解」與「生成」任務上往往各擅其長，難以兼顧兩種功能。Li et al.（2022）提出的 BLIP（Bootstrapping Language-ImagePre-training）旨在建立統一的視覺語言預訓練框架，同時勝任視覺語言理解與生成任務。BLIP 的模型架構採用了多模態編碼器-解碼器混合設計：包含單模態的圖像編碼器與文本編碼器、注入圖像訊息的跨模態文本編碼器，以及能生成文字的圖像引導文本解碼器。在預訓練過程中，BLIP 結合了圖文對比損失（ITC）、圖文匹配損失（ITM）和語言建模損失（LM）三種目標，同時進行跨模態對齊和文本生成學習。此外，針對網路爬取圖像-文字資料常含有語義不符的雜訊問題，BLIP 創新性地引入了 captioner 與 filter 模組：先用 captioner 為圖像產生合成描述，再用 filter 剔除不符合圖像內容的劣質文字描述。透過這種 bootstrapping 策略，BLIP 有效擴充了預訓練語句且降低雜訊干擾，最終在圖文檢索、圖像描述、視覺問答等七項視覺語言任務上取得當時的最佳成果。值得注意的是，BLIP 在零樣本轉移到影片領域也展現出強大的泛化能力，顯示其學習到的跨模態知識具有一定的通用性。

[1] https://openai.com/index/CLIP/

Gemini[2]由 Google DeepMind 開發的 Gemini 系列模型是近年來 LVLM 發展的另一高峰。Gemini 是一組原生多模態的大型模型，旨在統一處理文字、圖像、音訊、影片甚至程式碼等多種輸入類型。與以往將不同模態子模型後期拼接的方式不同，Gemini 從架構設計上即考慮多模態融合，在預訓練階段就以大規模跨模態資料進行聯合訓練，讓模型自底向上學會同時理解多種訊息。隨後透過額外的多模態微調，進一步提高模型對各類複雜任務的處理效果。Gemini 1.0 發布於 2023 年底，包含不同規模的子模型（Ultra、Pro、Flash、Nano），其中最大規模的 Ultra 模型在圖像理解、影音分析、數學推理等領域的 32 項基準測試中，有 30 項超越先前最先進模型的表現。尤其在多模態綜合推理基準 MMMU 上取得 59.4%的新紀錄，展示了強大的跨領域推理能力。值得注意的是，Gemini 從一開始就被設計為多模態模型，而非事後將視覺模組與語言模型簡單結合。這種原生融合使其能無縫理解並推理混合了文字與圖像等形式的複雜資訊。作為 Google 繼 PaLM 2 之後的下一代基礎模型，Gemini 被視為 OpenAI GPT-4 的有力競爭者，在多模態對話助理、代碼生成功能等方面均展現出領先性能[3]。總之，Gemini 代表了 LVLM 領域最新的技術集大成者，朝向真正的通用多模態人工智慧邁出重要一步。

(二).本研究中的 LVLM 應用方式

如上所述，LVLM 提供了強大的跨模態對齊與語言生成能力。本研究充分利用了現有 LVLM 模型的這些優勢，將其應用於資料建構與特徵提取兩大方面，以輔助下游任務的進行。

1.賽事影片敘事語句生成與預訓練資料建構：

本研究針對電競體育賽事影片資料集，面臨著缺乏文本標註的問題。因此，我們引入 LVLM 來自動生成影片敘事語句，作為預訓練階段的輔助語句。具體而言，我們使用了 Gemini 模型對賽事影片進行描述生成。操作上，將影片關鍵禎或片段截取輸入 Gemini，讓模型產生對應的自然語言敘事句子。Gemini 則作為多模態模型，可理解影像內容並給出語意豐富的文字敘述。透過 LVLM 模型，我們為每段賽事影片自動產生了敘事描述，例如比賽中的關鍵事件解說、選手操作與比賽狀態等語句。這些自動生成的描述相當於額外的弱標註資料，極大擴充了預訓練語句庫的規模與多樣性。有了這批由 LVLM 提供的合成敘事語句，我們得以在預訓練階

[2] https://blog.google/technology/ai/google-gemini-ai/#performance

[3] https://en.wikipedia.org/wiki/Gemini_(language_model)#:~:text=Gemini%20is%20a%20family%20of,was%20rated%20as%20highly%20competitive

段以「影片-語句」配對的形式對模型進行多模態學習，儘早讓模型對影像和語言訊息進行關聯對齊。這不僅緩解了資料不足問題，也提高了模型對賽事情境的理解能力。總之，透過 Gemini 的自動敘事生成，我們構建了大規模且內容豐富的預訓練資料，有助於下游任務的學習。

2.語言特徵擷取與多模態任務輸入融合：

在模型的下游任務中，精彩片段偵測（HD）與片段檢索（MR），我們同樣引入 LVLM 的文本編碼能力來提取語言模態特徵，作為模型的重要輸入之一。具體做法是：對於任務中涉及的文字訊息（如賽事影片對應的描述語句、用戶查詢的文字敘述等），我們分別採用 BLIP 的文本編碼器與 CLIP 的文本編碼器將其編碼為向量表示。BLIP 的文本編碼器本質上與 BERT 類似，已在大規模圖文語句上訓練，可產生包含豐富語意的文字特徵 CLIP 的文本編碼器則與圖像編碼器共享對齊的特徵空間，能將文字映射到與影像對齊的嵌入表示。透過這種方式，預訓練良好的語言特徵被融入下游模型，使其在判斷精彩片段或檢索關鍵時刻時，能同時利用文字描述所蘊含的語意資訊。這有效加強了多模態輸入的融合，提升模型對任務的理解與預測表現。

3.LVLM 語言特徵對模型訓練效果的影響評估：

應用上述 LVLM 產生的語言特徵後，我們關注其對下游模型訓練效果的影響。實驗結果顯示，引入 LVLM 語言特徵的模型在各項指標上均優於只使用原始視覺特徵的基線模型。這種性能提升一方面歸功於 LVLM 提供的語言表徵含有豐富的語意與知識，使模型在學習時能參考更高層次的語義線索；另一方面，LVLM 本身在巨量語句上預訓練所學得的泛化能力，有助於緩解下游任務中訓練資料不足的問題。例如，CLIP 模型透過預訓練學到通用的圖文對齊表示，因而即使不使用目標任務的標註資料，其提取的特徵也能在多種視覺任務中取得與有監督模型相當的表現。同樣地，在本研究中，BLIP-2 與 Gemini 生成的敘事語句以及 BLIP/CLIP 編碼的文本特徵，相當於將這些模型背後所學習的龐大語句知識轉借給我們的任務模型使用。這對於資料稀少的場景非常有利：模型不再完全依賴少量訓練樣本來學習，因為預訓練語句中的常識知識與語意模式已經隱含於 LVLM 特徵中。也就是說，LVLM 提供的優勢主要在於大規模語句學習所涵蓋的廣度與多樣性，解決了資料不足下模型難以學到泛化表示的痛點。

參、研究方法

一、研究流程

本研究旨在探討現有視覺語言模型（如 VideoLights）於《英雄聯盟》電競賽事影片中執行片段檢索（MR）與精彩片段偵測（HD）任務的適用性，並進一步驗證自動生成敘事語句資料是否能部分取代傳統人工標註流程。整體研究流程如圖 2 所示，主要分為以下五個階段：

(一).資料收集與影片片段切割：

本研究從多場《英雄聯盟》職業賽事中蒐集完整賽局影片，並依不同用途進行切分：(1) 預訓練資料以每 10 秒擷取一幀畫面作為代表影格，供自動生成語句使用；(2) 微調資料以每 150 秒切分為一段片段，供人工撰寫完整敘事與標註事件。上述視覺輸入將用於後續的語句構建與特徵抽取。

(二).敘事語句（Query）建構：

本研究的語句來源包括：(1) 由大型視覺語言模型（Gemini）針對代表影格自動生成的簡潔語句，形成弱監督資料；(2) 由人工撰寫、包含角色位置、事件地點及支援角色等細節的高語意敘事描述。人工語句再依語意細緻度分為簡單與複雜兩類，作為後續分析的依據。

(三).資料標註與格式轉換：

將每則 query 搭配其對應片段，進行 saliency 標註與時間區段定位，形成符合 VideoLights 模型輸入規範之資料格式，包括：qid, query, vid, duration, relevant_clip_ids, relevant_windows, saliency_scores。

(四).訓練與測試資料切分：

標註完成的資料集依照訓練用途進行分割，訓練集、驗證集與測試集的比例分別為 70%、15%、15%。切分時考量事件分布的均衡性，確保每類事件（如擊殺、推塔、搶龍等）在三組資料中具有基本代表性，以維持模型學習與評估結果的穩定性與一致性。

(五).模型訓練與推論：

使用 VideoLights 模型進行雙階段訓練與推論，第一階段使用 LVLM 生成之弱監督資料進行預訓練，第二階段則使用人工標註資料進行微調（fine-tuning）。模型

同時處理 MR 與 HD 任務，並輸出事件片段預測結果。

(六).結果評估與比較分析：

透過多項指標評估模型於不同資料類型下之表現，包括 MR 任務的 R1@0.5（Recall at 1 with IoU threshold 0.5，首個預測片段與標註片段交併比 ≥ 0.5 的比例）、mAP（mean Average Precision，所有查詢的平均精確度均值）等指標，以及 HD 任務的 HIT@1 指標（首個預測結果與標註事件完全匹配的比例），並針對敘事語句複雜度進行比較分析，探討語意強度與模型效能之關聯性。

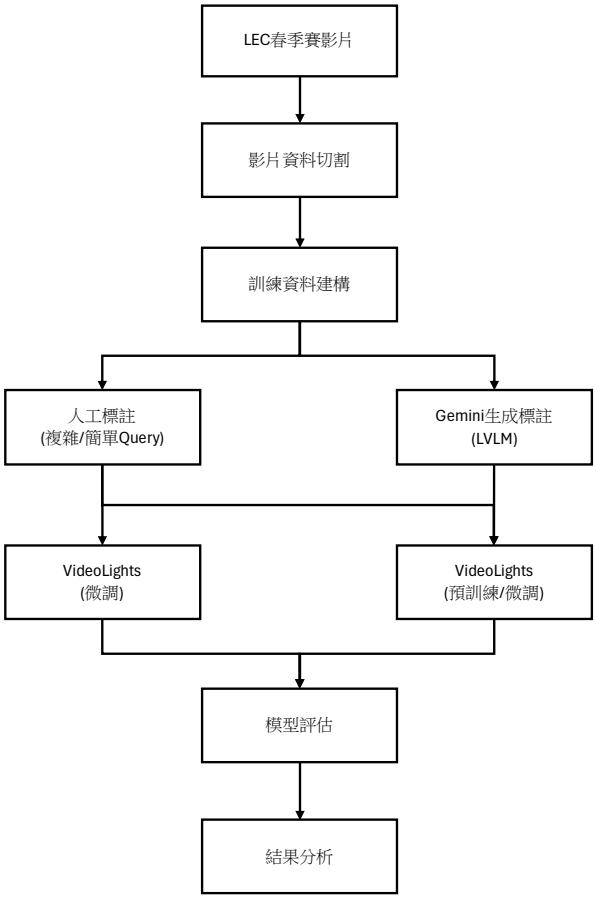

圖 2　研究流程圖

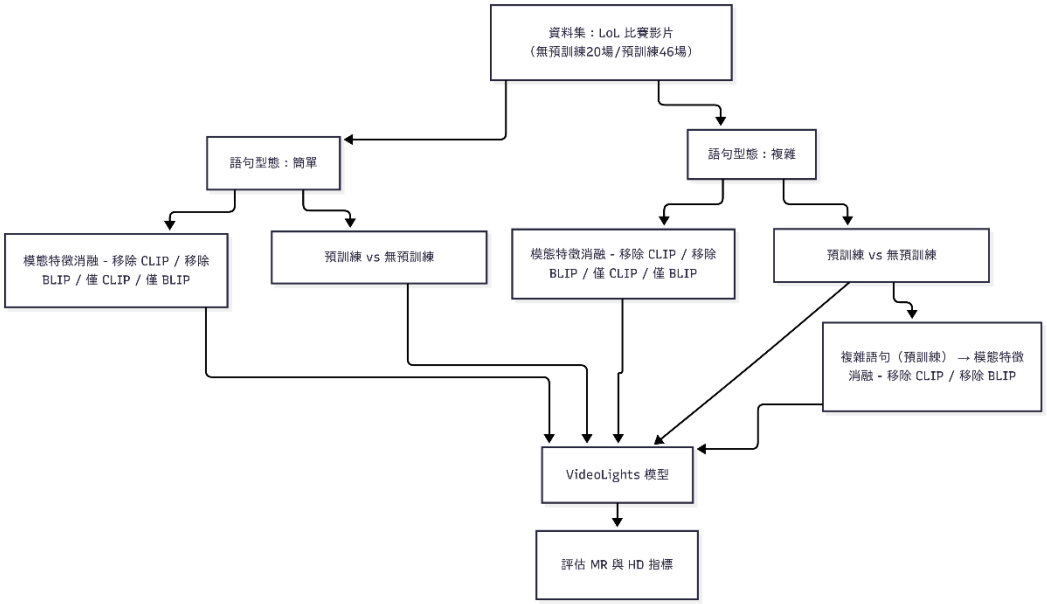

圖 3　研究架構圖

本研究架構以圖 3 所示，《英雄聯盟》比賽影片（無預訓練 20 場/預訓練 46 場）為資料來源，首先依據語句型態分為簡單及複雜並在各組下進行模型消融實驗（移除 CLIP /移除 BLIP /僅 CLIP / 僅 BLIP）。接著針對兩種語句型態分別比較有無弱監督預訓練之影響，進一步檢驗預訓練對不同語句型態的適配性。由於結果顯示複雜語句在預訓練下效果較佳，因此再以此組合進行最後的模態消融，藉此確定最優特徵組合。所有設定最終導入 VideoLights 模型，並以 MR 與 HD 指標進行效能評估。

二、資料來源

本研究旨在探討多模態模型於電競賽事影片之應用，特別聚焦於《英雄聯盟》2023 年歐洲、中東與非洲冠軍聯賽（League of Legends EMEA Championship, LEC）春季賽，作為主要資料來源。研究透過結合大型視覺語言模型（LVLM）生成之敘事語句與人工標註資料，建立適用於片段檢索（MR）與精彩片段偵測（HD）任務之多層次資料集。

研究共收集 LEC 春季賽期間的 46 場比賽影片，涵蓋 9 天賽程（每日 5 場，最終日 6 場），影片來源均為 Twitch 平台之官方直播。每場比賽由不同戰隊進行對戰，轉播內容包含主播解說與多角度遊戲畫面切換，具有高度視覺資訊豐富性。

為促進研究再現性與後續學術應用，本研究已將所使用之《英雄聯盟》賽事影片資料集與人工標註之語句（含複雜與簡單 query）公開至 GitHub，網址如下：

https://github.com/victor-wang0125/lol-highlight-dataset/tree/master

(一).完整性與品質穩定：

LEC 為全球四大職業聯賽之一，賽事製作品質高，錄影內容清晰、無剪接，有助於模型學習真實事件之時序邏輯與上下文語境。

(二).語言相容性佳：

影片解說與字幕主要以英文為主，方便結合主流視覺語言模型進行語意建模。

(三).具觀眾代表性與應用潛力：

作為全球主要聯賽之一，LEC 資料具備外部推廣性，可應用於其他電競或賽事 highlight 自動剪輯任務。

三、資料預處理

為建構可供模型訓練與推論的輸入資料格式，本研究首先針對收集之《英雄聯盟》賽事影片進行資料預處理，主要步驟如下：

(一).影片切割與片段化處理：

所有賽事影片均來自官方 Twitch 頻道之完整賽局錄影。為利於語句對應、特徵抽取與模型訓練，影片會先進行統一長度的切割。本研究分別以 10 秒為單位產生弱標註訓練資料、以 150 秒為單位產出用於人工敘事與精準標註之片段。

(二).代表畫面擷取：

針對每段 10 秒的影片片段，本研究擷取該片段的最後一幀作為代表畫面（代表影格），提供給大型視覺語言模型（LVLM）進行敘事語句的自動生成。由於不同擷取策略（如首幀或中間幀）並不保證能捕捉到事件的關鍵畫面，因此本研究採取較為簡單且易於實現的固定取最後一幀方式，以確保資料處理流程的一致性與可操作性。

(三).語句生成與資料對齊：

透過 Gemini 模型對每個代表畫面生成敘事語句（query），並與對應的 10 秒片段進行綁定。這些資料將作為 VideoLights 預訓練階段的輸入，構成弱監督資料集。

(四).特徵抽取與格式轉換：

本研究使用 SlowFast 模型與 CLIP 模型對影片片段進行視覺特徵抽取。所有特徵皆儲存為.npz 格式，以供 VideoLights 模型載入。每段資料在轉換後需包含以

下欄位：

1.qid：該 query 對應的編號。

2.query：由 LVLM 或人工產生的語句。

3.vid：對應的影片片段編號。

4.duration：片段長度（秒）。

5.relevant_clip_ids：該 query 所對應之重要影格索引。

6.relevant_windows：對應的時間區間（以兩秒為一單位）。

7.saliency_scores：標註者所給予的關注程度分數。

四、標註方式

本研究共設計兩類標註資料來源：LVLM 所產生之弱監督資料，與人工撰寫之高語意敘事資料。其中僅有用於模型微調與測試階段的人工資料，進一步區分為「簡單敘事」與「複雜敘事」兩類。

(一).LVLM 弱監督標註（用於預訓練）：

預訓練階段的資料來源為將原始比賽影片切分為長度為 10 秒的短片段，每段搭配使用 Gemini 生成的畫面敘述作為查詢語句（query，作為檢索影片中對應片段的文字描述輸入），並使用 BLIP 模型中的 Image-Text Matching（ITM）技術，透過敘事語句與片段代表畫面的相似度比對，產出每段影片的顯著性分數（saliency score，用於衡量影片片段對任務的重要程度，數值越高代表越關鍵）。

(二).人工精華標註（用於微調）：

微調階段則使用人工標註資料。我們將影片切分為 150 秒的片段，針對每段進行「精彩片段」的判斷與註記，標註依據包括以下幾類：

1.擊殺行為（如連殺、首殺、反殺）。

2.團隊交戰（多人交戰或團戰起始）。

3.目標物爭奪（如巴龍、飛龍、兵營、防禦塔）。

4.關鍵轉折點（如殘血反殺、逆轉勝）。

此外，人工標註的資料片段均以 150 秒為單位，並由標註者針對每段片段進行 saliency 標註與時間區段（relevant window）定位，形成最終用於微調階段的資料集。

其中僅有用於模型微調與測試階段的人工資料，進一步區分為「簡單敘事」與「複雜敘事」兩類：

(三).簡單敘事語句：

通常僅描述角色之間的擊殺事件，例如“VIT's Bo has slain FNC's Oscarinin.”。

(四).複雜敘事語句：

除了描述擊殺事件外，亦會補充角色位置、發生地點與支援角色資訊，如“Bo (Jungle) secures a kill against Oscarinin (Top) in the top lane, assisted by Photon (Top).”

將該區段內所有 clip_id 依每 2 秒對應編號展開為 relevant_clip_ids 標註每個 CLIP 的 saliency score（標註者所評之分數）將上述資訊儲存為 JSONL 格式。

五、模型設計與訓練策略

本研究採用 VideoLights 作為核心模型進行片段檢索（MR）與精彩片段偵測（HD）任務。該模型具備支援多任務學習與跨模態語意對齊的架構，能同時處理影片與語句之對應關係預測。

VideoLights 模型具備高度模組化設計，其核心結構整合三大組件：語意強化的 FRA 模組（Fine-grained Region Attention）、跨模態融合的 Bi-CMF（Bi-directional Cross-modal Fusion）注意力機制，以及統合多任務學習輸出的 Uni-JFM（Unified Joint Feature Module）架構，支援 MR 與 HD 任務的同時學習與預測。模型基礎特徵萃取部分則搭配 SlowFast 進行動作特徵分析，以及 CLIP/BLIP 進行圖文語意特徵嵌入。

訓練流程採雙階段設計，預訓練階段使用由大型視覺語言模型（LVLM）自動生成之語句與 BLIP ITM 推估之 saliency 分數執行弱監督式訓練以預先對齊跨模態語意，微調階段則使用人工標註資料進行監督式學習，並探討敘事語句語意層次對模型預測表現之影響。

本研究重點在於驗證：現有多模態模型在電競語境下的表現、資料標註自動化的可行性，以及敘事語意強度對任務準確度的影響。

六、評估指標

為全面評估模型於片段檢索（MR）與精彩片段偵測（HD）任務中的效能，本研究採用多種主流評估指標進行量化分析：

(一).片段檢索評估指標：

1.R@1@IoU（Recall at 1 with IoU threshold）：評估模型是否能在第 1 個預測中正確定位事件區段，涵蓋多個 IoU 門檻（如 0.3、0.5、0.7 等）。

2.mAP@IoU（mean Average Precision）：衡量模型在多個 IoU 門檻下的精度平均值，反映定位準確性。

3.mIoU（mean Intersection over Union）：計算預測區段與標註區段之平均重疊程度。

4.分段長度分析：依據事件片段長短分為 Long（30 – 150 秒片段）, Middle（10–30 秒片段）, Short（10 秒以下片段）三類，比較其 MR 表現差異。

(二).精彩片段偵測評估指標：

1.HIT@1：評估模型於 Top-1 預測中是否成功命中精華片段。

2.mAP：衡量模型在分類為精華與否的預測準確性。

3.HL-min-Fair/Good/Very Good：本研究依據 saliency_scores 欄位之分數，將片段品質分為三層：Fair（2 分以下）、Good（3 分）、Very Good（4 分），以比較模型在低、中、高強度精華片段的預測差異，並評估其難易分辨性。

肆、實驗與結果

一、實驗設計

本研究核心任務為結合多模態輸入進行《英雄聯盟》賽事影片之片段檢索（MR）與精彩片段偵測（HD）預測。本研究聚焦於資料構建與應用成效驗證，故採用現有模型 VideoLights 作為主體架構，並依實驗需求調整其訓練策略與資料流。

本研究延續原始模型之雙階段訓練策略設計，分為：

(一).預訓練階段（synthetic training）：

使用由 LVLM（Gemini 模型）根據每 10 秒片段之代表畫面生成的簡潔敘事語句，搭配 BLIP ITM 模型進行語句與畫面語意對齊，計算相似度分數作為 saliency label，建構弱監督式資料集。此階段可有效擴充訓練規模，並促使模型學習語意模糊但結構規律的跨模態映射關係。

(二).微調階段（fine-tuning）：

以人工標註之高語意片段作為精緻樣本，對模型進行監督式微調。標註資料來自 150 秒長度片段之人工語句與 saliency scores，並進一步區分為簡單敘事語句與複雜敘事語句，以驗證語句語意強度對模型表現之影響。該階段為模型導入領域知識與精細語境結構的關鍵。

(三).資料分割與訓練設定：

最終訓練資料按照 70%、15%、15%比例切分為訓練集、驗證集與測試集，此配置與 QVHighlights（Lei et al., 2021）的資料配置相符，目的在於確保模型能有足夠的資料進行參數學習，同時保留驗證與測試資源以監控過擬合情形與評估泛化效能。並統一使用 VideoLights 提供之標準訓練流程與多任務損失函數，輸出片段預測結果（事件時間區段與顯著性等級）。

透過上述雙階段訓練設計，本研究旨在驗證：

1.現有多模態檢索模型能否適用於高結構語境的電競影片 。

2.弱監督資料是否足以取代人工標註。

3.敘事語句的語意複雜度是否會影響跨模態對齊與片段預測效能。

二、實驗組別與變項設計

為深入驗證不同資料條件與模組配置對模型效能的影響，本研究設計多組實驗進行比較，實驗變項涵蓋語句型態、資料規模與特徵來源等三個面向，具體說明如下：

(一).語句型態比較：

表（一）呈現複雜敘事語句與簡單敘事語句在模型微調訓練階段的表現差異，目的在於檢驗語句語意強度是否影響跨模態對齊與預測效能。兩組實驗皆使用相同的 20 筆人工標註資料，唯一差異為 query 句型的敘述複雜度。

整體而言，使用複雜語句的模型在大多數片段檢索（MR）任務指標上略優於簡單語句，包含 MR-full-R1@0.3（56.47 vs. 51.14）、MR-full-R1@0.5（51.76 vs. 46.59）、MR-full-mAP（32.74 vs. 32.50）與 mIoU（41.95 vs. 39.49）。這顯示語句若提供更多語意線索，能幫助模型更準確地對齊影片片段，提升片段定位精度。然而，在 MR-full-R1@0.7（29.41 vs. 31.82）與 MR-middle-mAP（37.01 vs. 39.12）等指標中，簡單語句表現略勝一籌，顯示模型在語意過於豐富時可能面臨對齊困難。

在精彩片段偵測（HD）任務中，複雜語句組於 HL-min-Fair-mAP（58.91 vs. 56.51）與 HL-min-Good-mAP（58.89 vs. 56.33）表現優於簡單語句組；而在 HL-min-VeryGood-mAP（37.06 vs. 35.00）與 HL-min-VeryGood-Hit@1（29.41 vs. 27.27）指標上，也維持穩定的優勢。此結果顯示，即使在高閾值的顯著事件偵測任務中，豐富敘事語句仍具備一定貢獻，有助於模型更完整地掌握語意關鍵，提升事件辨識能力。

表 1　語句型態比較表

（N=20）

指標	複雜語句	簡單語句
MR-full-R1@0.3	56.47	51.14
MR-full-R1@0.5	51.76	46.59
MR-full-R1@0.7	29.41	31.82
MR-full-mAP	32.74	32.50
MR-full-mAP@0.5	56.57	53.72
MR-full-mAP@0.75	31.02	31.76
MR-full-mIoU	41.95	39.49
MR-long-mAP	59.09	57.92
MR-middle-mAP	37.01	39.12
MR-short-mAP	21.48	18.90
HL-min-Fair-mAP	58.91	56.51
HL-min-Fair-Hit1	52.94	47.73
HL-min-Good-mAP	58.89	56.33
HL-min-Good-Hit1	52.94	47.73
HL-min-VeryGood-mAP	37.06	35.00
HL-min-VeryGood-Hit1	29.41	27.27

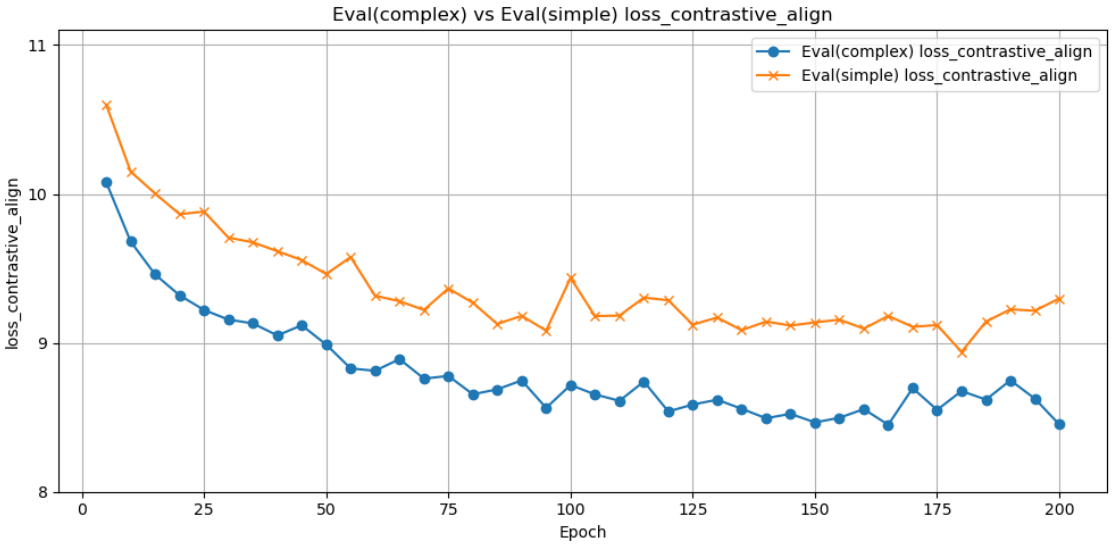

圖 4　複雜語句 vs 簡單語句之損失變化圖

為進一步分析語句型態對語意對齊效果之影響，本研究繪製圖 4，比較複雜與簡單語句在驗證階段之 loss_contrastive_align 損失變化。此損失項用於衡量模型在跨模態語意對齊過程中語句與影片特徵的相似度學習效果，數值越低代表對齊效果越佳。由圖中可見，兩組模型的對齊損失皆隨訓練進程逐漸下降，惟複雜語句組自起始便穩定低於簡單語句組，並持續維持顯著差距。最終，複雜語句組的 evalloss_contrastive_align 收斂於 8.5 附近，而簡單語句則停留於約 9.2 水準。

此結果顯示，語句中若包含更多語意元素與結構關係，模型能更有效地與多模態特徵進行語意對齊，尤其在驗證集上展現出更佳的跨模態理解泛化能力。這亦與表 1 中 MR 指標表現趨勢相符，進一步支持複雜語句在跨模態檢索任務中的語意優勢。

表 2 進一步比較在移除 CLIP 特徵後，複雜與簡單語句在模型微調表現上的差異，本研究設計實驗情境，將原有多模態特徵中的 CLIP 特徵移除，僅保留 BLIP 與 SlowFast 特徵作為輸入，並比較在此條件下使用複雜 query 與簡單 query 微調模型後的任務表現。表格中列出片段檢索（MR）與精彩片段偵測（HD）任務之主要評估指標，數據皆基於 20 筆資料進行訓練與測試。從結果觀察，複雜語句組在大多數指標上仍展現優勢，特別是在片段檢索（MR）任務中，其 mAP（35.27 vs. 32.36）、mAP@0.75（36.23 vs. 30.41）、mIoU（44.14 vs. 41.73）與長區間 mAP（73.64 vs. 58.58）等指標皆顯著優於簡單語句組，顯示語意豐富的描述有助於模型在較長或語意密度較高的片段中，獲得更精準的對齊能力。

在精彩片段偵測任務中，兩組在 HL-min-Fair 與 HL-min-Good 等級 mAP 與

Hit@1 指標表現相近，複雜語句組略佔優勢。而在 HL-min-VeryGood-mAP 上，簡單語句組則以 36.93 略高於複雜語句組之 35.91，惟其 Hit@1 指標(28.24 vs. 26.14) 仍由複雜語句組領先。整體來看，移除 CLIP 特徵後，語句本身的語意承載能力扮演了更關鍵的角色，複雜語句提供的語境資訊可在多數任務下補足模態刪減所造成的語意落差，進一步提升模型的語句對齊與事件辨識能力。

表 2　語句型態比較表（無 CLIP 特徵）

（N=20）

指標	複雜語句	簡單語句
MR-full-R1@0.3	60.00	59.09
MR-full-R1@0.5	48.24	50.00
MR-full-R1@0.7	35.29	30.68
MR-full-mAP	35.27	32.36
MR-full-mAP@0.5	54.10	54.19
MR-full-mAP@0.75	36.23	30.41
MR-full-mIoU	44.14	41.73
MR-long-mAP	73.64	58.58
MR-middle-mAP	40.50	42.55
MR-short-mAP	19.76	15.02
HL-min-Fair-mAP	60.90	60.30
HL-min-Fair-Hit1	60.00	56.82
HL-min-Good-mAP	60.85	60.19
HL-min-Good-Hit1	60.00	56.82
HL-min-VeryGood-mAP	35.91	36.93
HL-min-VeryGood-Hit1	28.24	26.14

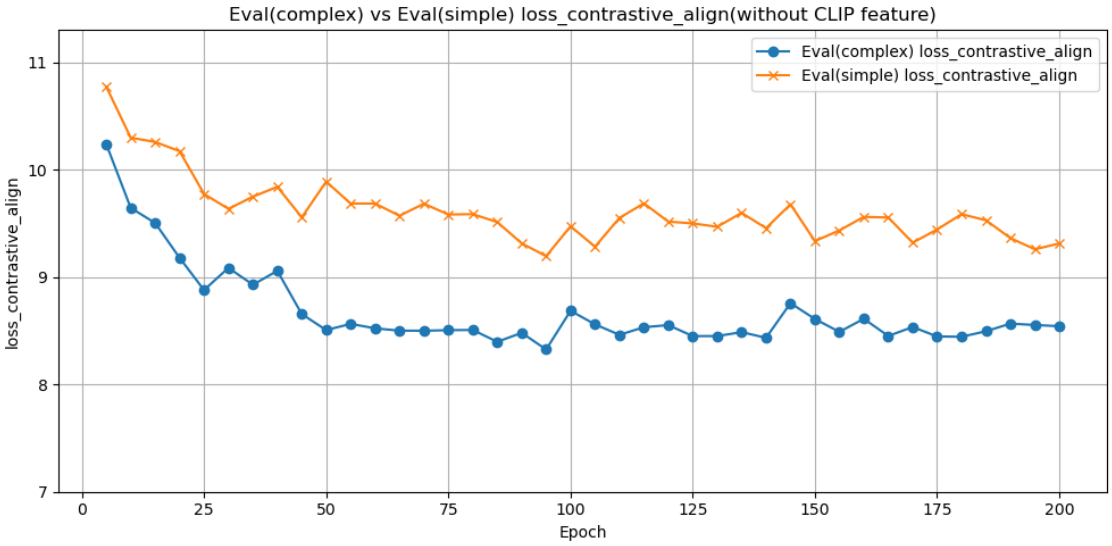

圖 5　複雜語句 vs 簡單語句之損失變化圖（無 CLIP 特徵）

圖 5 呈現複雜語句與簡單語句於移除 CLIP 特徵後，其語意對齊損失 loss_contrastive_align 在驗證階段的變化趨勢，旨在觀察不同語句型態於單一語言模態設定下的跨模態對齊能力。從圖中可見，兩組語句皆在初期快速收斂，惟整體表現仍存在差異。複雜語句在訓練後期穩定維持於較低損失水準（約 8.3～8.7），顯著低於簡單語句組（約 9.3～9.6），顯示在少模態配置下，複雜語句所提供的語意線索仍有助於語句與視覺特徵間的對齊學習。

此結果延續表 2 中 MR 與 HD 任務上的指標觀察，進一步驗證複雜語句在移除 CLIP 特徵的條件下，仍能展現語意支撐效果，有效提升模型對語句內容與視覺特徵之語義對應能力，並促進整體語意對齊學習穩定性。

在表 3 實驗設定中，模型同時採用 SlowFast 與 CLIP 作為視覺特徵來源，並僅使用 CLIP 模型進行文字特徵提取，旨在探討語句型態於此圖文模態配置下對模型效能之影響。實驗結果如表所示，整體而言，複雜敘事語句組在大多數評估指標上仍略勝一籌。特別是在片段檢索（MR）任務中，複雜語句於 R1@0.5（48.24 vs. 44.32）、mAP（31.24 vs. 30.55）與 mAP@0.75（31.37 vs.29.20）等高閾值指標上皆展現優勢，顯示豐富描述有助於模型在較長或語意密度較高的片段中取得更精準的對齊能力。

HighlightDetection 任務中，兩組語句在 Fair 與 Good 水準下的 mAP 與 Hit@1 表現相當，差距較小，但在 VeryGood 階段的 Hit@1 指標方面，複雜語句則明顯優於簡單語句（37.65 vs. 27.27），顯示其在處理語意密度更高之重要事件預測任務上更具助益。

整體來看，即使文字模態僅依賴 CLIP 特徵進行語意表徵，複雜語句仍能於高精度場景中發揮語意輔助作用，突顯語句內容設計在跨模態檢索任務中所扮演的重要角色。

表 3　語句型態比較表（無 BLIP 特徵）

（N=20）

指標	複雜語句	簡單語句
MR-full-R1@0.3	50.59	50.00
MR-full-R1@0.5	48.24	44.32
MR-full-R1@0.7	32.94	31.82
MR-full-mAP	31.24	30.55
MR-full-mAP@0.5	54.02	50.35
MR-full-mAP@0.75	31.37	29.20
MR-full-mIoU	38.97	38.86
MR-long-mAP	50.91	51.67
MR-middle-mAP	36.83	38.58
MR-short-mAP	20.68	16.98
HL-min-Fair-mAP	57.19	56.86
HL-min-Fair-Hit1	54.12	53.41
HL-min-Good-mAP	57.17	57.07
HL-min-Good-Hit1	54.12	53.41
HL-min-VeryGood-mAP	38.45	36.90
HL-min-VeryGood-Hit1	37.65	27.27

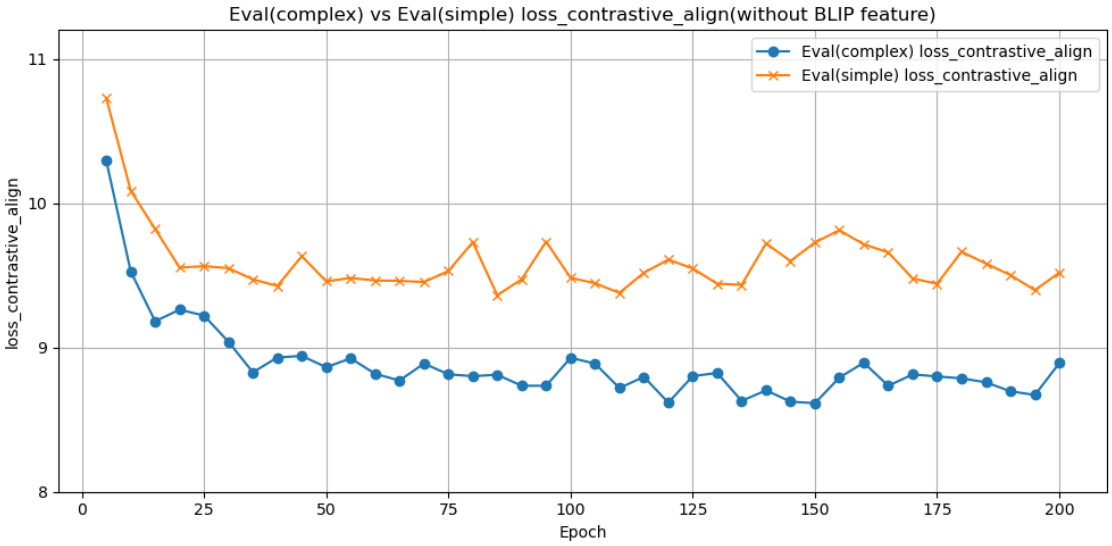

圖 6　複雜語句 vs 簡單語句之損失變化圖（無 BLIP 特徵）

為進一步佐證語句型態於語意對齊階段之實際差異，圖 6 呈現複雜語句與簡單語句於驗證階段的 loss_contrastive_align 損失變化曲線，對照模態設定為：移除 BLIP，僅保留 CLIP 作為語言模態。圖中可見，儘管整體損失水準高於其他模態設定，顯示語意對齊表現受限，但複雜語句於大多數訓練階段仍維持低於簡單語句的損失值，顯示其在語意引導上具穩定優勢。尤其在後期趨於穩定的階段，複雜語句組的 loss_contrastive_align 明顯較低，對應於表 3 中在 MomentRetrieval 指標皆優於簡單語句的表現。此結果強化了即便語言模態資訊受限，語句本身的語意結構仍對模型的跨模態對齊效果產生正面作用。

在僅保留 CLIP 特徵的實驗條件下，表 4 比較了複雜敘事語句與簡單敘事語句在片段檢索（MR）與精彩片段偵測（HD）任務中的微調表現。所有模型皆以 20 筆標註資料進行訓練與測試，僅語句內容有所差異。

整體觀察顯示，簡單敘事語句在多數 MR 指標上略優於複雜語句。例如，MR-full-R1@0.3（46.59 vs. 38.82）、MR-full-R1@0.5（39.77 vs. 34.12）與 mIoU（33.72 vs. 31.12）均呈現簡單語句具更高召回與定位精度。在 HL 任務中亦呈現類似趨勢，簡單語句在 HL-min-Fair-mAP（49.26 vs. 47.12）、HL-min-Good-mAP（49.24 vs. 47.12）、HL-min-VeryGood-mAP（31.16 vs. 25.07）等評估指標上均優於複雜語句，顯示在高閾值條件下其事件判斷精度也更佳。

此結果反映出 CLIP 特徵在處理語意複雜度較高的語句時仍具侷限性，可能無法充分解析複雜語句中的多層語意關係，導致其潛在優勢未能有效轉化為模型表現。該現象突顯出語言特徵提取器與語句型態之間的耦合關係，亦提供未來在語言

模組設計與語句風格調整上的優化參考。

表 4　語句型態比較表（僅 CLIP 特徵）

（N=20）

指標	複雜語句	簡單語句
MR-full-R1@0.3	38.82	46.59
MR-full-R1@0.5	34.12	39.77
MR-full-R1@0.7	24.71	25.00
MR-full-mAP	25.63	26.04
MR-full-mAP@0.5	42.10	45.70
MR-full-mAP@0.75	25.28	23.88
MR-full-mIoU	31.12	33.72
MR-long-mAP	53.64	54.17
MR-middle-mAP	36.48	35.27
MR-short-mAP	7.99	9.29
HL-min-Fair-mAP	47.12	49.26
HL-min-Fair-Hit1	40.00	42.05
HL-min-Good-mAP	47.12	49.24
HL-min-Good-Hit1	40.00	42.05
HL-min-VeryGood-mAP	25.07	31.16
HL-min-VeryGood-Hit1	14.12	19.32

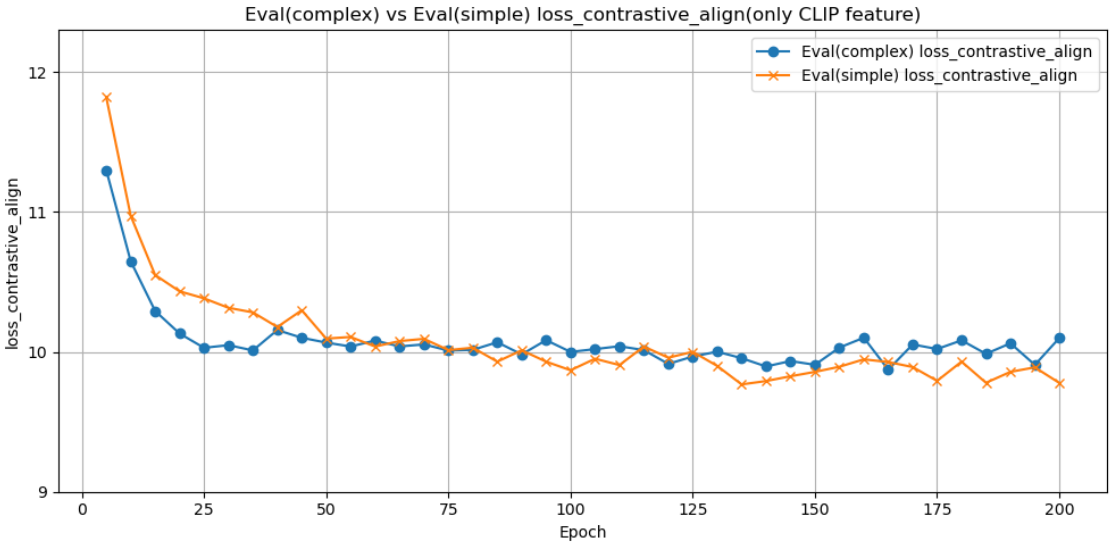

圖 7　複雜語句 vs 簡單語句之損失變化圖（僅 CLIP 特徵）

為補充表 4 中定量指標的分析，本研究進一步繪製圖 7 比較在僅使用 CLIP 作為語言模態特徵的情境下，複雜語句與簡單語句於驗證階段的 loss_contrastive_align 變化趨勢。從圖中可見，在訓練初期，複雜語句組之 loss_contrastive_align 略低於簡單語句，表示模型初始能夠從較豐富語句中提取一定語意結構資訊，進行語言與視覺間的初步對齊。然而，隨著訓練輪次增加，簡單語句組的損失迅速下降並與複雜語句組收斂至相近範圍，甚至在中後期出現多個區間損失更低的現象，顯示其語意表示更為穩定，且更易與視覺模態建立一致關聯。此一趨勢與表 4 中定量評估指標一致，簡單語句組於多數 MR 與 HD 指標中均略優於複雜語句，尤其在 HL-min-VeryGood-mAP（31.16vs.25.07）與 R1@0.3（46.59vs.38.82）等指標中展現明顯差距。這顯示 CLIP 作為通用型圖文對齊模型，在處理語法結構較簡單、語意層級較低的語句時表現更為穩定，其對於複雜語句中的角色關係、事件因果與多層語義線索較難準確建模與嵌入，導致複雜語句未能發揮其應有的資訊密度優勢。

此外，從圖中變異趨勢亦可推論：CLIP 所提取的語言向量未能對複雜語句進行足夠語意拆解與內部結構建模，造成模型難以捕捉句中多主詞或多事件的語意重點，而僅能依賴句中關鍵詞進行表面對齊。這種對語句內容的語意壓縮，使得複雜語句未僅未帶來額外效益，反而可能引入雜訊。

整體而言，圖 7 反映了在文字模態僅倚賴 CLIP 的設定下，簡單語句相較複雜語句具有更高的穩定性與語意對齊效率，突顯語言模態的選擇將直接影響語句型態的適配性。此結果提醒我們，未來進行跨模態檢索任務時，不僅需考量語句內容

本身之資訊量與語意深度，也須關注語言編碼模組的語意表徵能力與語句風格之耦合關係，進一步提升語句設計與模型特徵選擇之匹配程度。

我們探討了複雜敘事語句與簡單敘事語句在僅使用 BLIP 特徵提取器的情況下，對模型在片段檢索（MR）與精彩片段偵測（HD）任務表現的影響。比較結果如表 5 所示。

在 MR 任務中，兩組語句在整體指標上表現相近，但複雜語句於部分關鍵指標上略占優勢。例如，在 MR-full-R1@0.5（38.82 vs. 37.50）、MR-full-mAP（26.91 vs. 26.55）、以及 MR-full-mAP@0.75（28.03 vs. 26.20）等項目中，複雜語句提供的語意資訊似乎略有助於提升模型在精準召回上的能力。然而在 mIoU（32.45 vs. 33.50）項目中，簡單語句略優的現象，顯示語句型態對時間重疊預測的影響仍相對有限。

在 HD 任務中，簡單語句組的表現則較為穩定優越，特別是在 HL-min-Fair-mAP（51.02 vs. 48.80）、HL-min-Good-mAP（50.96 vs. 48.40）與對應的 Hit@1 指標（44.32 vs. 40.00）皆略優於複雜語句組。僅在 HL-min-VeryGood-Hit@1 指標上兩組表現十分接近，差距不大。

綜合而言，本實驗顯示在僅使用 BLIP 特徵的情況下，複雜語句於 MR 任務具一定優勢，而 HD 任務則較傾向支持簡單語句。此現象與 BLIP 所擅長的圖文對齊特性有關，當缺乏其他語言或視覺模態輔助時，簡潔明確的語句反而能更直接對應模型所抽取的關鍵語意，對 HD 任務更為有利。此結果對未來在特徵選擇與語句建構策略的配置設計，提供了實證基礎與參考方向。

表 5　語句型態比較表（僅 BLIP 特徵）

（N=20）

指標	複雜語句	簡單語句
MR-full-R1@0.3	42.35	40.91
MR-full-R1@0.5	38.82	37.50
MR-full-R1@0.7	27.06	27.27
MR-full-mAP	26.91	26.55
MR-full-mAP@0.5	45.79	42.32
MR-full-mAP@0.75	28.03	26.20
MR-full-mIoU	32.45	33.50
MR-long-mAP	56.82	54.44
MR-middle-mAP	34.18	34.51
MR-short-mAP	11.95	11.01
HL-min-Fair-mAP	48.80	51.02
HL-min-Fair-Hit1	40.00	44.32
HL-min-Good-mAP	48.40	50.96
HL-min-Good-Hit1	40.00	44.32
HL-min-VeryGood-mAP	24.93	27.19
HL-min-VeryGood-Hit1	16.47	15.91

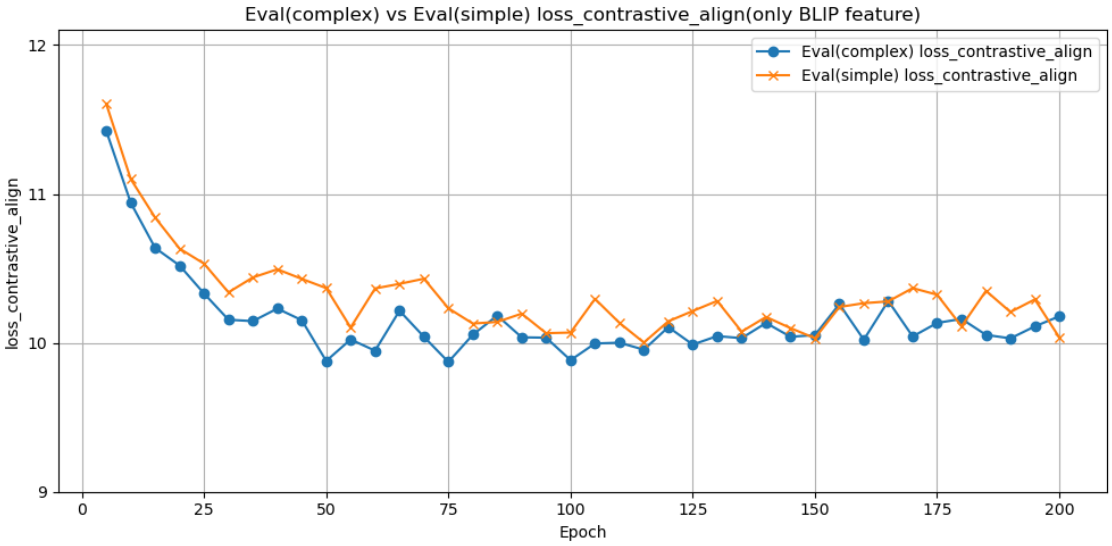

圖 8　複雜語句 vs 簡單語句之損失變化圖（僅 BLIP 特徵）

圖 8 呈現複雜語句與簡單語句於僅使用 BLIP 模態時，在 loss_contrastive_align 損失上的比較結果。整體趨勢顯示，雖然兩者對齊損失差異有限，但複雜語句在訓練後期表現出更穩定且略低的損失值，反映其在跨模態對齊學習中具有些微優勢。這意味著，當模型僅依賴 BLIP 進行圖文對齊時，複雜語句的語意密度仍可帶來一定程度的正向效果，協助模型更精確地對應視覺片段。

然而，此優勢並未在 HD 任務中完全展現。根據表 5 結果，簡單語句於多數 HD 評估指標上仍略勝一籌，說明在缺乏其他輔助語言或視覺模態時，BLIP 模型對高語意複雜度語句的處理能力有限，難以將其語意優勢完整轉化為事件辨識效能。圖表中的穩定差距與 HD 任務表現不一致，也突顯語句型態與模態配置之間的交互作用需視任務而定。

整體來看，本圖顯示複雜語句在語意對齊層面仍具一定潛力，尤其在 MR 任務中可能對片段定位有助益。但若缺乏多模態支援，其優勢將難以被語言模態單獨發揮。此觀察進一步強化了語句設計與模態選擇需搭配考量的論點，亦為多模態系統的設計策略提供了實證依據。綜整各項語句型態實驗結果可見，語句的語意強度對模型效能具有一定影響，但其效果高度依賴於特徵模態的組合。在多模態配置中（如同時結合 SlowFast、CLIP 與 BLIP 特徵的設定），複雜語句相較簡單語句在片段檢索（MR）任務中的 R@1 與 mAP 指標表現更佳，顯示語句所承載的豐富語意在多源特徵支援下能有效轉化為片段定位的精度提升，尤其在高閾值評估（如 mAP@0.75、R1@0.7）上更為明顯。

然而，當模型僅使用單一語言模時，複雜語句的優勢則相對縮小。特別是在精

彩片段偵測（HD）任務中，簡單語句組在多數指標中反而表現略優或相當，顯示在語意處理能力較受限的模態條件下，語句簡潔反而有助於減少語意對齊負擔與資訊干擾。

整體來看，簡單語句在單一模態條件下具穩定性與解碼優勢，而複雜語句則需倚賴多模態支援才能充分釋放語意潛力。因此，語句型態設計不宜脫離模態結構獨立考量，應與多模態特徵提取器相輔相成，方能實現最佳語意對齊與事件識別效果。

(二).弱監督預訓練效果比較：

為探討預訓練階段是否有助於提升模型在《英雄聯盟》賽事影片任務上的表現，本研究進行有無使用 Gemini 模型進行弱監督預訓練之比較實驗。兩組皆採用相同數量（20 筆）之複雜敘事語句進行微調，差異僅在於是否經歷合成資料之預訓練流程。

從表 6 觀察，有預訓練組於片段檢索（MR）任務中表現整體優於未預訓練組。其在 mAP（34.88 vs. 32.74）、高召回門檻（R1@0.7 為 35.29 vs. 29.41）以及 MR-short-mAP（27.52 vs. 21.48）皆有明顯提升，顯示預訓練有助於模型建構更穩定的跨模態對齊能力。

在精彩片段偵測（HD）任務中，兩組表現雖接近，但預訓練組仍展現穩定優勢。於 Fair 與 Good 標準下，Hit@1 與 mAP 指標皆有領先，如 HL-min-Good-Hit@1 為 60.00%，高於無預訓練組的 52.94%。此外，在 HL-min-VeryGood 項目中，預訓練組在 Hit@1（34.12 vs. 29.41）與 mAP（40.47 vs. 37.06）皆優於對照組，顯示即便在資料量較稀疏的高品質標準下，LVLM 預訓練仍具正向效果。

整體而言，Gemini 弱監督預訓練流程雖非全面性提升所有指標，但對 MR 任務的整體召回與精度提升具明顯助益，對 HD 任務亦有一定推進效果。此結果顯示 LVLM 自動生成之合成語句與弱標註資料，能有效輔助模型建立初步語意映射基礎，進而提升後續微調學習的穩定性與泛化能力。

表 6　弱監督預訓練效果比較表

（N=46）

指標	預訓練	無預訓練
MR-full-R1@0.3	60.00	56.47
MR-full-R1@0.5	51.76	51.76
MR-full-R1@0.7	35.29	29.41
MR-full-mAP	34.88	32.74
MR-full-mAP@0.5	57.47	56.57
MR-full-mAP@0.75	32.96	31.02
MR-full-mIoU	44.37	41.95
MR-long-mAP	63.04	59.09
MR-middle-mAP	34.23	37.01
MR-short-mAP	27.52	21.48
HL-min-Fair-mAP	60.74	58.91
HL-min-Fair-Hit1	60.00	52.94
HL-min-Good-mAP	60.92	58.89
HL-min-Good-Hit1	60.00	52.94
HL-min-VeryGood-mAP	40.47	37.06
HL-min-VeryGood-Hit1	34.12	29.41

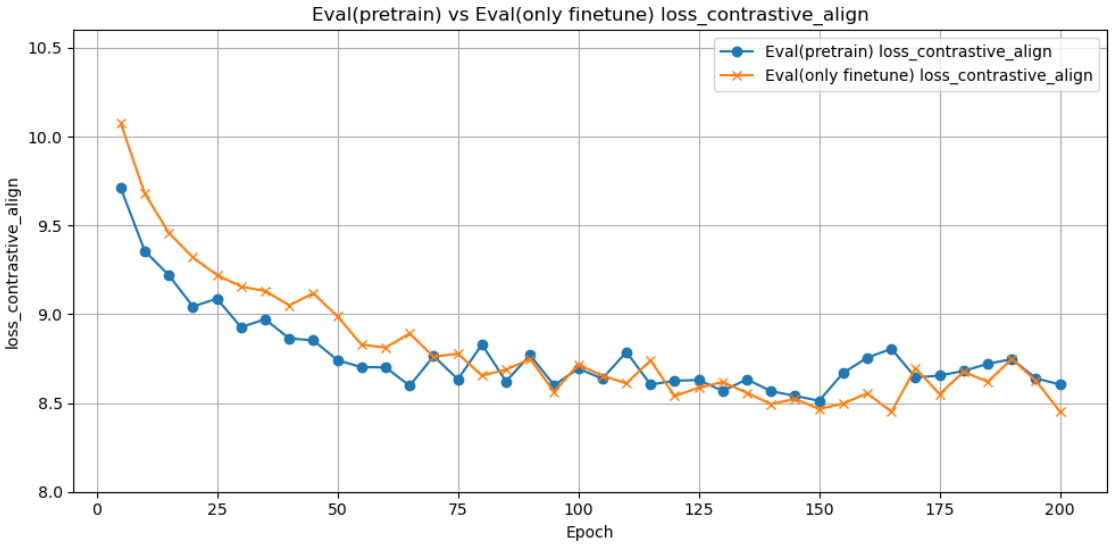

圖 9　預訓練 vs 無預訓練之損失變化圖

在探討弱監督預訓練對模型語意對齊能力的影響時，本研究進一步比較有無預訓練模型於 loss_contrastive_align 損失函數上的變化情形，結果如圖 9 所示。整體而言，進行預訓練的模型在訓練初期與中期展現出較低的對齊損失值，代表其語句與視覺片段間的語意匹配學習更為快速且穩定，顯示預訓練確實能提供有效的語意初始化。然而，值得注意的是，在訓練後期，僅進行微調的模型在部分區段的 loss_contrastive_align 表現出略低於預訓練組的趨勢，並伴隨較高的波動幅度。此一結果顯示，儘管微調模型初期學習較慢，仍可能在長時間訓練後持續優化語意對齊能力，縮短與預訓練模型的差距，甚至在個別階段出現優於預訓練組的表現。

整體而言，該圖表明弱監督預訓練雖有助於模型語意對齊的快速收斂與穩定性，但其最終效果仍與訓練步數、資料規模、語句型態等因素密切相關。預訓練並非絕對優勢，其實際效益仍需搭配下游任務評估指標（如 MR/HD 之 mAP 與 R@1）進行整合分析。

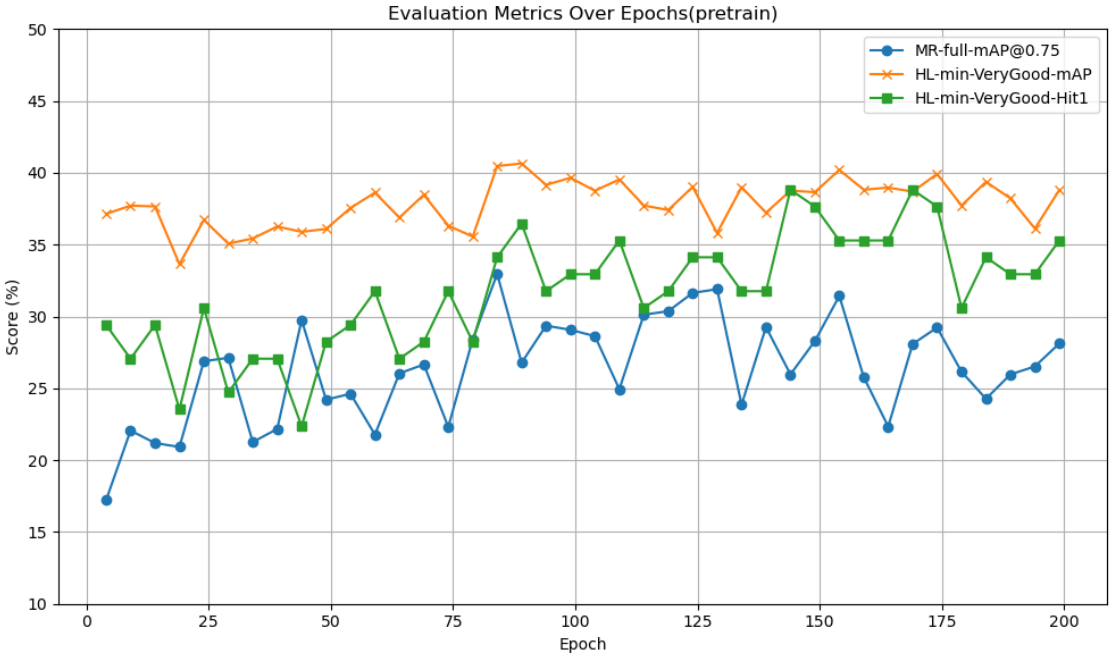

圖 10　預訓練之評估指標圖

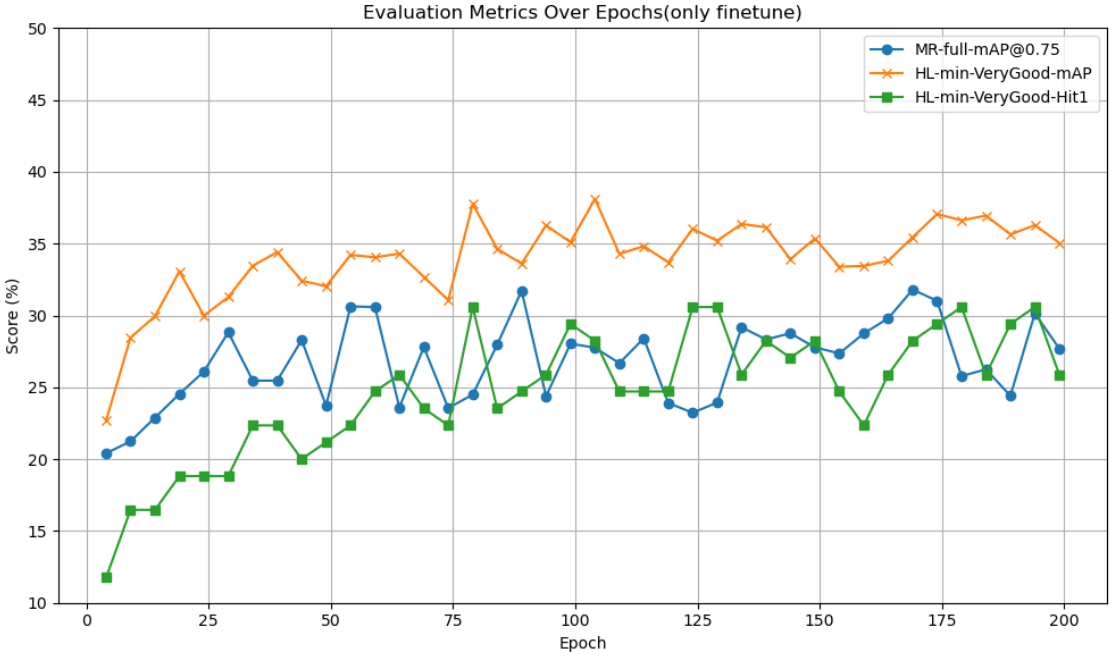

圖 11　無預訓練之評估指標圖

雖然本研究於實驗中同時採用了多組評估指標（如 R1@0.3/0.5/0.7、mAP@0.5、mIoU 等），但為避免圖表過於繁複，並強調模型在高標準情境下的表現，後續圖表展示將以 MR-full-mAP@0.75 以及 HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit@1 三項指標為主要觀察對象。此舉能更清楚凸顯模型在嚴格條件下的檢索精度與精華事件辨識能力。

進行弱監督預訓練的模型在多數評估指標上優於僅做微調的模型，且其訓練過程中的指標曲線更加平滑穩定、波動幅度較小。如圖 10 所示，三個評估指標均呈現逐步上升後趨於穩定的趨勢。具體而言，HL-min-VeryGood-mAP 在訓練初期即位於較高水平，並隨著迴圈增加略微上升到約 38%，曲線波動幅度較小且在高位區間內平穩震盪；HL-min-VeryGood-Hit1 則從約 29%起始，隨訓練進行整體向上躍升，最高達到約 38%，最終穩定在 30~40%之間，但呈現一定程度的起伏；相對而言，MR-full-mAP@0.75 起始值僅約 16%~18%，隨著訓練迴圈逐漸爬升，在中後期最高約達 35%左右，最終保持在 20%~30%之間。圖 11 中三項指標的變化走勢存在顯著差異。HL-min-VeryGood-mAP 初始值相對較低，但在前 25 個迴圈迅速上升到約 30%-35%，隨後進一步提升並在 35%~40%範圍內波動，訓練後期與預訓練階段的 HL-min-VeryGood-mAP 曲線僅有微小差距。HL-min-VeryGood-Hit1 起點更低，在中後期最高約達 31%，但也經歷較大波動，最終位於約 25%~30%區間。相比之下，MR-full-mAP@0.75 變化較為平緩，從約 20%左右起步，逐步升至約 25%-30%，中後期穩定在這個水平。可以觀察到，純微調階段的 HL 指標起初較低，需較多迴圈才能達到與預訓練階段相當的表現，而 MR 指標的提升趨勢則與預訓練階段相似但略遲緩。最後，根據表 6 的最終最佳結果，預訓練模型在高門檻評估指標上（如 MR-full-R1@0.7、MR-full-mAP@0.75 以及 HL-min-VeryGood 系列）均明顯優於未經預訓練之模型，進一步凸顯了以語意初始化所帶來的效能增益。

(三).模態特徵消融實驗（Ablation Study）：

為探討不同視覺語言模態對《英雄聯盟》賽事影片中片段檢索（MR）與精彩片段偵測（HD）任務效能之影響，本研究進行特徵模態消融實驗，分別移除 CLIP 與 BLIP 特徵，並與全模態基準組（包含 SlowFast、CLIP 與 BLIP 特徵）進行比較，結果如表 7 所示。

實驗結果顯示，在多數指標上，移除 CLIP 特徵後模型表現不僅未下滑，甚至明顯提升。例如 MR 任務中的 MR-full-mAP 自 32.74 上升至 35.27，MR-full-R1@0.7 則從 29.41 增加至 35.29，顯示模型即使不依賴 CLIP 所提供之圖文嵌入資訊，仍能有效捕捉語句與片段之對應關係。此外，MR-full-mIoU 也由 41.95 提升至 44.14，突顯模型對片段邊界的預測更加精確。

HD 任務方面，在高標準判定條件（如 HL-min-Good、HL-min-Fair）下，移除 CLIP 特徵後的模型於 mAP 與 Hit@1 表現亦與完整模態組相當甚至略優，如 HL-min-Good-mAP 提升至 60.85，HL-min-Fair-Hit@1 提升至 60.00，顯示 CLIP 特徵在此應用場域可能並非關鍵模態。

相較之下，移除 BLIP 特徵所產生的影響更為多元，部分片段檢索（MR）與精彩片段偵測（HD）指標略有下降，例如 MR-full-mAP 從 32.74 下降至 31.24，MR-long-mAP 亦明顯下滑，顯示 BLIP 特徵在語句語意理解與跨模態對齊中仍具一定貢獻。然而值得注意的是，在部分高閾值指標中，如 MR-full-R1@0.7（由 29.41 升至 32.94）與 HL-min-VeryGood-Hit1（由 29.41 提升至 37.65）等，反而呈現上升趨勢，顯示移除 BLIP 特徵在某些情境下可能有助於提升模型對關鍵片段的判斷能力。

綜合上述，實驗結果指出，在語意結構複雜且專業術語密集的電競影片任務中，CLIP 特徵並非必要模態，移除後不僅可有效降低多模態運算成本，亦有助於提升模型預測穩定性。另一方面，BLIP 特徵則對整體語意對齊具備基礎貢獻，但其存在在特定場景下亦可能引入模態干擾。此發現可作為未來多模態模型設計與模態選擇的重要參考依據。

表 7　模態特徵消融實驗比較表(無預訓練)

（N=20）

指標	無移除特徵	移除 CLIP 特徵	移除 BLIP 特徵
MR-full-R1@0.3	56.47	60.00	50.59
MR-full-R1@0.5	51.76	48.24	48.24
MR-full-R1@0.7	29.41	35.29	32.94
MR-full-mAP	32.74	35.27	31.24
MR-full-mAP@0.5	56.57	54.10	54.02
MR-full-mAP@0.75	31.02	36.23	31.37
MR-full-mIoU	41.95	44.14	38.97
MR-long-mAP	59.09	73.64	50.91
MR-middle-mAP	37.01	40.50	36.83

MR-short-mAP	21.48	19.76	20.68
HL-min-Fair-mAP	58.91	60.90	57.19
HL-min-Fair-Hit1	52.94	60.00	54.12
HL-min-Good-mAP	58.89	60.85	57.17
HL-min-Good-Hit1	52.94	60.00	54.12
HL-min-VeryGood-mAP	37.06	35.91	38.45
HL-min-VeryGood-Hit1	29.41	28.24	37.65

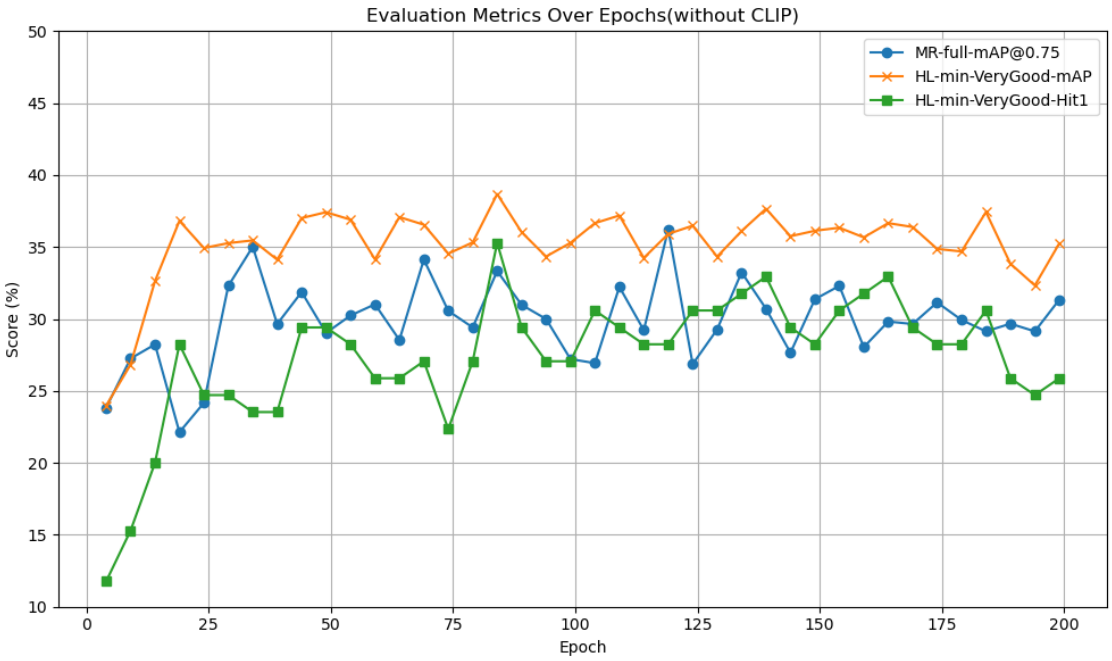

圖 12　無預訓練且移除 CLIP 特徵之評估指標圖

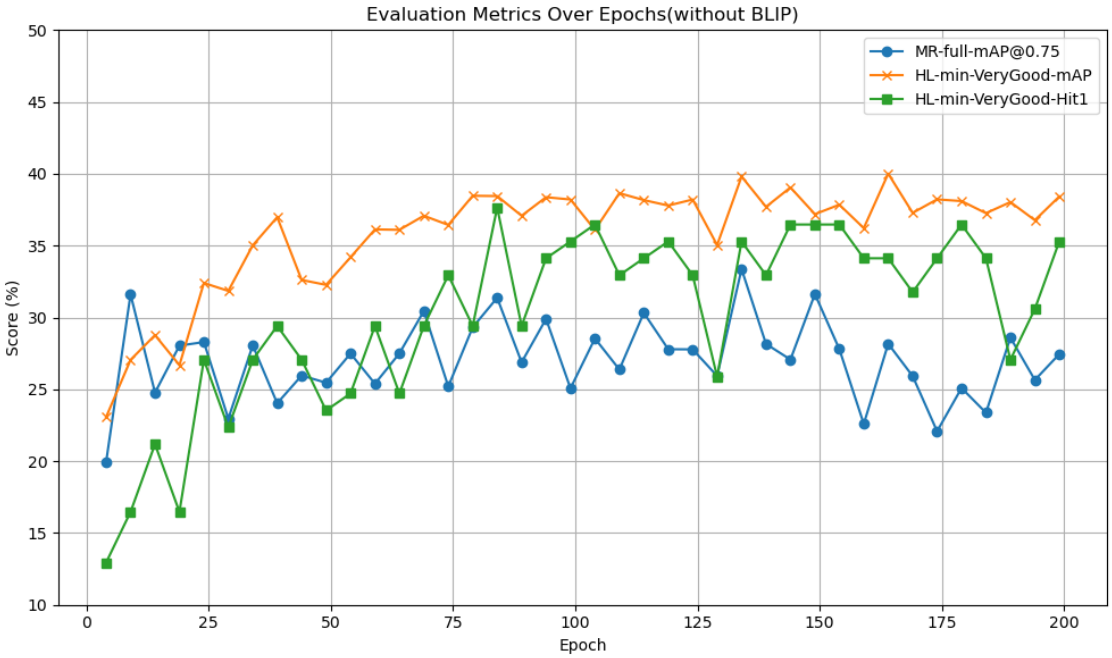

圖 13　無預訓練且移除 BLIP 特徵之評估指標圖

在無預訓練條件下，圖 11 呈現完整模態組於三項高閾值任務指標（MR-full-mAP@0.75、HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit1）之訓練變化情形，三者皆隨訓練輪次逐步上升，並於約第 100 輪後趨於穩定。其中，以 HL-min-VeryGood-mAP 表現最佳，整體曲線上升穩定且最早收斂，其次為 MR-full-mAP@0.75，HL-min-VeryGood-Hit1 則相對較低，顯示模型在高精度事件召回上尚具挑戰。

當移除 BLIP 特徵後，如圖 13 所示，HD 任務指標明顯改善，特別是 HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit1 皆出現顯著上升，分別可達約 40%與 37%，整體優於完整模態組；而 MR-full-mAP@0.75 則維持與完整模態相近之水準。從訓練曲線觀察，HL-min-VeryGood-mAP 穩定上升且波動幅度小，顯示該模態組合於 HD 任務中更具學習穩定性；而 MR-full-mAP@0.75 則於前期呈現較大震盪，後期收斂趨勢與完整模態相仿。此結果與先前觀察一致，即在無預訓練情況下移除 BLIP 特徵，能提升部分高門檻 HD 任務之指標表現。

另一方面，移除 CLIP 特徵時，如圖 12 所示，模型於 MR 任務之收斂與效能皆顯著改善。MR-full-mAP@0.75 於訓練初期即快速達到約 35%的高水準，HL-min-VeryGood-mAP 表現亦與完整模態相近，而 HL-min-VeryGood-Hit1 則介於完整模態與移除 BLIP 模態之間，顯示其整體表現兼具穩定與精準。從訓練曲線分析，

CLIP 移除後之模型於收斂速度與指標峰值均表現優異，與先前分析一致，即 CLIP 特徵可能於無預訓練設定下引入冗餘訊息，降低語意對齊效率。

綜合以上分析可知，在無預訓練設定下，模態組合對高閾值任務表現與收斂特性影響顯著。完整模態組合整體穩定，適用於平衡各任務需求；移除 CLIP 有助於提升 MR 任務表現與加速收斂；而移除 BLIP 則可強化高階 HD 任務表現。此結果顯示不同視覺與語言模態間資訊價值不一，研究者應依據實際任務目標審慎配置模態特徵，以達最佳訓練成效。

進一步考察在經過弱監督預訓練條件下，各模態特徵對任務表現的影響，表 8 呈現模型於保留全部特徵、移除 CLIP 特徵與移除 BLIP 特徵三種設定下的效能比較。

在片段檢索（MR）任務中，整體而言，完整模態組與移除 CLIP 特徵組的表現不相上下，甚至在若干指標上，移除 CLIP 組表現略優。例如 MR-full-mAP 提升至 37.76（原為 34.88），MR-full-R1@0.7 提升至 36.47，且 MR-full-mAP@0.75 提升至 38.20，顯示模型於高相似區段的檢索能力因移除 CLIP 而略有增強，推測可能是 CLIP 所引入之語意資訊與其他模態特徵在本任務中存在某種干擾或重疊。

然而，在 MR-long-mAP 上，移除 CLIP 特徵後大幅提升至 74.55，較完整組的 63.04 明顯優異，顯示其對長距區段具有更強的整體語意整合能力；但在 MR-short-mAP 上則下降至 20.18，說明其在短片段檢索上表現有所削弱。至於移除 BLIP 特徵的結果，整體呈現明顯下滑趨勢。例如 MR-full-mAP 降至 31.30、MR-full-mAP@0.5 與@0.75 分別下降至 49.84 與 29.43，MR-long-mAP 也明顯下滑至 50.91。顯示在弱監督預訓練情境下，BLIP 所提供的語意嵌入特徵依然對模型理解語句與影片片段之間的對齊關係具關鍵作用。

在精彩片段偵測（HD）任務方面，完整模態組整體表現仍略優於其他兩組，特別是在 HL-min-Good 及 HL-min-VeryGood 條件下，mAP 與 Hit@1 指標皆達最高（如 HL-min-Good-Hit@1 為 60.00，HL-min-VeryGood-Hit@1 為 34.12）。移除 CLIP 或 BLIP 特徵後雖略有下滑，但差異不大，顯示 HD 任務可能相對仰賴視覺模態中 SlowFast 提取之動態資訊，而非 CLIP 或 BLIP 提供的視覺特徵。

整體來看，在有預訓練條件下，BLIP 特徵對 MR 任務具明顯貢獻，而 CLIP 特徵反而在部分情境下可能產生資訊冗餘或干擾，其移除甚至提升模型在長片段檢索的穩定性與精度。此一觀察與無預訓練情境下的結果形成對照，無預訓練時，移除 CLIP 可提升 MR 和部分 HL 表現；但有預訓練時，完整模態對 HL 輕微有利，而移除 CLIP 只在某些 MR 指標上略優。這說明預訓練後模型已部分融合 CLIP 訊息，使 CLIP 的相對貢獻降低甚至成為冗餘。顯示模態設計需視任務類型與資料特

性進行彈性調整。

表 8　模態特徵消融實驗比較表(預訓練)

（N=46）

指標	無移除特徵	移除 CLIP 特徵	移除 BLIP 特徵
MR-full-R1@0.3	60.00	50.59	51.76
MR-full-R1@0.5	51.76	47.06	45.88
MR-full-R1@0.7	35.29	36.47	35.29
MR-full-mAP	34.88	37.76	31.30
MR-full-mAP@0.5	57.47	55.71	49.84
MR-full-mAP@0.75	32.96	38.20	29.43
MR-full-mIoU	44.37	40.93	40.59
MR-long-mAP	63.04	74.55	50.91
MR-middle-mAP	34.23	45.79	38.34
MR-short-mAP	27.52	20.18	19.46
HL-min-Fair-mAP	60.74	57.92	56.76
HL-min-Fair-Hit1	60.00	52.94	54.12
HL-min-Good-mAP	60.92	57.91	56.76
HL-min-Good-Hit1	60.00	52.94	54.12
HL-min-VeryGood-mAP	40.47	35.23	36.30
HL-min-VeryGood-Hit1	34.12	27.06	34.12

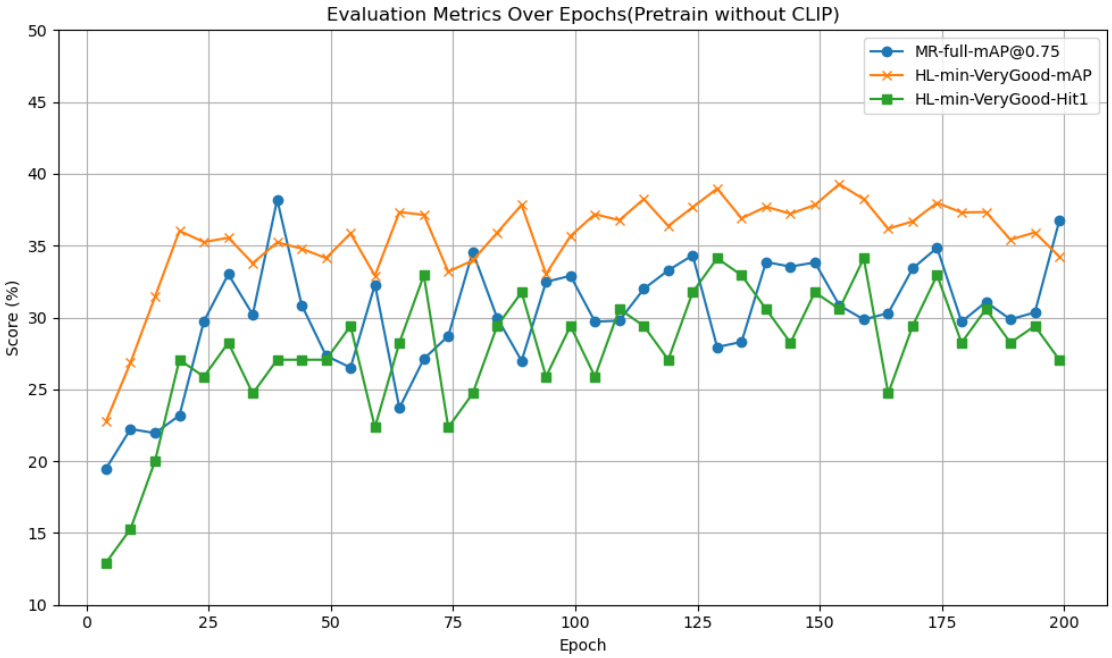

圖 14　有預訓練且移除 CLIP 特徵之評估指標圖

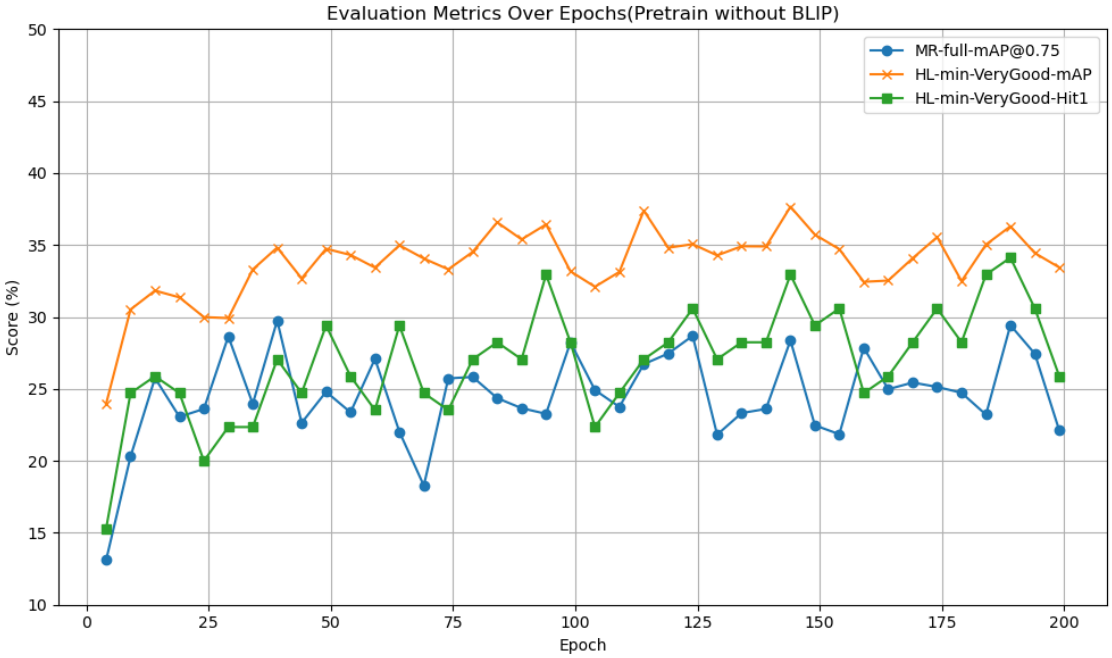

圖 15　有預訓練且移除 BLIP 特徵之評估指標圖

在預訓練條件下，圖 10 呈現完整模態組於三項高閾值任務指標（MR-full-mAP@0.75、HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit1）之訓練變化情形，三者皆隨訓練輪次逐步上升，其中，以 HL-min-VeryGood-mAP 表現最佳，整體曲線上升穩定且最早收斂，其次為 HL-min-VeryGood-Hit1，顯示模型在此條件下具備良好的語意對齊與高品質事件檢索能力。而 MR-full-mAP@0.75 則相對較低，顯示模型在高精度事件召回上尚具挑戰。

當移除 BLIP 特徵後，如圖 15 所示，HD 任務指標明顯改善，三項指標表現均較完整模態明顯下滑，尤其 MR-full-mAP@0.75 曲線呈現明顯波動，後期甚至出現下降趨勢，顯示缺乏 BLIP 語意嵌入會削弱模型對敘事語句的理解與片段對齊能力。

另一方面，移除 CLIP 特徵時，如圖 14 所示，移除 CLIP 特徵所造成的影響則較小甚至略具正向效應。HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit1 介於完整模態與移除 BLIP 模態之間。MR-full-mAP@0.75 則於訓練早期即快速上升，收斂速度甚至優於完整模態，顯示在模型已接受預訓練之條件下，則 CLIP 所提供的靜態圖像特徵對整體效能影響相對有限。

綜合上述分析結果，我們發現完整模態搭配預訓練設定可提供最穩定且優越的 HD 任務效能，特別是在 HL-min-VeryGood-mAP 與 HL-min-VeryGood-Hit1 等高閾值指標上表現最為突出。另一方面，移除 CLIP 特徵後雖略微犧牲部分 HD 指標穩定性，但在 MR-full-mAP@0.75 上則呈現明顯上升趨勢，反映 CLIP 模態在特定視覺語意任務中可能產生冗餘訊息，適度移除反而有助於模型聚焦於動態語意對齊學習。綜合而言，BLIP 特徵為實現高語意語句理解與跨模態精準檢索的關鍵模態，而 CLIP 特徵則應視任務需求調整配置。

三、結果總結與跨實驗比較

本研究透過多組變項操弄，全面檢驗了語句型態、預訓練機制與視覺模態對於 VideoLights 模型在《英雄聯盟》賽事影片任務中的影響。以下彙整各實驗結果，進行整體歸納與分析。

首先，在語句型態比較方面，複雜敘事語句於片段檢索（MR）任務中的 R1@0.5 及 mAP 指標大致優於簡單語句，顯示豐富的語意資訊有助於模型更精確地對應影片片段。在精彩片段偵測（HD）任務中，當模型僅採用單一模態特徵時，簡單敘事語句的 Hit@1 與各級 mAP 表現往往略勝複雜語句。當模型採用多模態特徵時，複雜語句的指標優於簡單語句，顯示豐富敘事能與多源特徵更有效結合，提升語意對齊與事件識別精度。此一結果突顯出語句設計與特徵模態間的交互效果，亦即在多模態架構下，豐富語意才能發揮其潛力，反之在模態受限的情況下則可能導致干

擾。

在預訓練效果比較方面，觀察結果顯示使用 Gemini 預訓練語句與 BLIP ITM 推估 saliency 分數進行弱監督學習之模型，其在片段檢索（MR）任務的多數指標表現皆不亞於無預訓練組，尤以高閾值評估指標（如 R1@0.7、mAP@0.75）提升最為明顯。例如，在複雜語句下的完整模態設定中，有預訓練組的 MR-full-R1@0.7 為 35.29，相較無預訓練組之 29.41 明顯優越；而 MR-full-mAP@0.75、HL-min-VeryGood-mAP 等指標亦展現出相同趨勢。

此結果顯示，弱監督預訓練雖無需高成本人工標註，卻仍能有效強化模型對語句與片段間細緻語意關聯的捕捉能力，尤其有助於提升模型在嚴格判準下的預測準確性。這也顯示出，大型視覺語言模型在預訓練階段所提供的語意資訊，有助於提升模型對跨模態資料的理解與對齊能力，特別是在語句與影片片段語意關聯較為複雜時，能發揮輔助作用。

模態消融實驗顯示不同視覺語言模態對於模型效能具有明確影響，且其效果會隨是否經歷預訓練而改變。在無預訓練的情況下，移除 CLIP 特徵反而能提升模型整體表現：MR-full-mAP 由 32.74 提升至 35.27，MR-full-R1@0.7 自 29.41 增至 35.29，顯示 CLIP 在未經語意對齊學習下，可能產生語意干擾或冗餘訊息，限制模型對片段的準確辨識。在無預訓練時移除 CLIP 可提升整體效能，在有預訓練情境下，確實完整模態在大部分精彩片段偵測（HD）指標上最佳，但對於片段檢索（MR）任務的某些指標，移除 CLIP 反而提升了績效（例如 MR-middle-mAP 和 MR-long-mAP 在移除 CLIP 時明顯提高）。值得注意的是，在進行預訓練的條件下，BLIP 特徵確實展現出穩定且關鍵的貢獻。移除 BLIP 特徵將明顯導致多數 MR 任務指標下降，特別是 MR-full-mAP 與 MR-long-mAP 表現顯著低於完整模態組，顯示 BLIP 作為文本對齊模組，在處理語句語意嵌入與跨模態對齊上具備不可取代的重要性，尤其在面對複雜敘事語句時更為關鍵。然而，在未進行預訓練的情況下，BLIP 特徵的效能貢獻則相對不一。如前所述，雖然部分指標（如 MR-mAP）仍呈下降趨勢，惟亦觀察到少數高閾值指標（如 HL-min-VeryGood-Hit1）於移除 BLIP 特徵後略有提升，顯示其在未預訓練設定下可能存在模態干擾的情形。

綜合而言，語句設計與預訓練策略對模型表現具有顯著影響。模態選擇應依據任務需求與語境特性進行調整，而非一味追求模態多元化。本研究結果指出，在特定情境下精簡模態反而有助於提升模型效能，為多模態檢索系統的設計與資料建構策略提供了實證依據。下一章將進一步討論本研究的理論貢獻、限制與未來應用方向。

伍、結論與未來發展

本研究針對《英雄聯盟》賽事影片進行多模態片段預測任務之探討，透過語句型態、預訓練策略與特徵模態這三種實驗變項，系統性檢驗了 VideoLights 模型於高語意密度語境中的適應能力與挑戰。經由多組實驗與跨組比較，研究結果指出語意豐富之敘事語句與 LVLM 弱監督預訓練可有效提升模型準確性，而視覺語言模態的選擇亦須依任務屬性調整，未必愈多愈好。

本章將進一步總結本研究之主要發現與貢獻，反思研究過程中的限制，並針對未來多模態模型在領域特化應用上的發展方向提出建議。

一、結論

本研究旨在探討多模態語意檢索模型於電競賽事影片中之應用潛力，透過結合大型視覺語言模型（LVLM）與 VideoLights 架構，進行片段檢索（MR）與精彩片段偵測（HD）任務實驗，並聚焦於語句型態、預訓練策略及特徵模態等變項的影響效果。綜合實驗結果，歸納以下三項主要結論：

(一).複雜語句於多模態條件下展現優勢：

當採用多模態特徵時，複雜敘事語句能顯著提升 MR 任務的對齊與預測效果，尤其在中高閾值指標（如 R1@0.5、mAP@0.75）上表現優異。然而，僅採單一模態時，簡單語句反而更穩定於 HD 任務中取得較佳 Hit@1 表現，顯示語句語意強度與特徵模態間具有交互效應。此結果指出，在多模態架構中，語句的語意豐富度（semantic richness）對模型的跨模態對齊能力有直接影響。複雜語句提供更多語境與事件細節，使模型能從視覺特徵中提取對應線索，進一步強化語意連結與時間區段的精準定位，顯示跨模態推理能力會隨語句內容提升。

在實際部署中，若系統能獲得結構化或詳盡的事件描述，可搭配多模態模型進行精細剪輯，但若僅能取得簡短描述或標籤，則建議優化語句設計或轉向輕量級特徵模態以降低語意負擔。

(二).弱監督預訓練能穩定提升模型對高閾值任務之表現：

透過 Gemini 模型自動生成語句並搭配 BLIP ITM 推估 saliency 分數之弱監督預訓練方式，能在無需人工標註的情況下，提升模型對語意關聯的捕捉能力。特別在 MR 的高閾值與 HD 的 Very Good 指標上皆有顯著進步，突顯大語言模型在預訓練階段的價值。針對資源有限或人工標註成本高的應用場景，建議可優先導入

LVLM 為基礎的語句生成與 saliency 標註方法，作為訓練資料擴充的首選機制，再進行針對性微調以達最佳效能。

(三).模態組合需依任務調整，非所有特徵模態皆有助益：

結果指出模態組合的效果並非線性疊加，特定特徵（如 CLIP）在專業且語意結構密集的任務中，可能與任務語境不匹配而造成語意干擾。這提醒我們在設計多模態系統時，特徵模態的組合亦須審慎調整，不同視覺語言模態之跨模態語意對齊程度與特徵重複性將直接影響模型效能。

綜合各實驗組別結果觀察，於片段檢索（MR）任務中"結合弱監督預訓練且移除 CLIP 特徵"的模型配置整體表現最為優異。，其 MR-full-R1@0.3 達 61.18、R1@0.7 為 36.47、mAP 為 37.76，mAP@0.75 亦達 38.20；在精彩片段偵測（HD）任務方面則是"結合弱監督預訓練且無移除特徵" 的模型配置表現最為優異，HL-min-Good-mAP 與 HL-min-VeryGood-mAP 分別為 60.92 與 40.47，HL-min-VeryGood-Hit@1 更達 34.12，整體成效明顯優於其他組別。此結果顯示，透過語意引導的弱監督學習搭配適當模態篩選，不僅能提升模型之跨模態對齊能力，亦有助於強化對高品質片段的辨識與預測表現。

二、研究貢獻

本研究的主要貢獻可歸納如下：

(一).驗證 HD/MR 多模態模型於高結構電競語境中的可行性與挑戰

本研究首次針對《英雄聯盟》這類高結構語境的電競影片，實證分析精彩片段偵測（HD）與片段檢索（MR）模型在其上的應用效果，即使採用目前表現優異的 VideoLights 架構，在缺乏語境專屬處理的情況下仍面臨語意掌握與事件邊界對齊等瓶頸。此發現對後續語境特化的模型優化策略提供了依據。

(二).評估 LVLM 弱監督生成資料於實務場景中的應用潛力與限制

本研究利用 26 部英雄聯盟影片搭配 Gemini 模型生成弱標語句，並作為 VideoLights 預訓練資料，驗證其對模型語意對齊能力的幫助與侷限。實驗結果顯示，儘管 LVLM 可自動生成語句並減少人工負擔，然而其對高語意密度場景的理解力仍有限，顯示未來仍需設計具任務特化能力的視覺語言模型。

(三).建構並釋出電競專屬語意資料集設計流程

本研究手動標註 20 部比賽影片、超過 600 段事件資料，包含文字敘述、時間範圍與顯著性標籤，建立符合電競語意邏輯的微調資料。此資料集之設計流程可為

未來建立高語境資料集提供參考，並補足目前主流影片資料庫多聚焦生活場景的限制。

(四).實證指出現有多模態模型在高語意任務下的應用邊界

本研究雖未對 VideoLights 架構進行修改，但透過應用於特定電競任務後，觀察其在長事件、中等顯著性任務上具穩定性，在短事件與 Very Good 類別表現下降，揭示其模型架構在語意密集、高節奏場景下的理解瓶頸，具實務意涵與學術啟示。

(五).比較不同語意細緻度敘事語句對多模態精華檢測之影響

本研究於標註資料階段設計兩類語句輸入策略，作為多模態模型中 query 的語言模態來源。第一類為簡單敘事語句，僅包含選手間的擊殺關係；第二類為複雜敘事語句，額外加入選手所屬位置、發生地點、助攻角色等細節，形成更高語意密度的描述。

實驗結果顯示，複雜敘事語句在模型檢測表現上略優於簡單語句，顯示於《英雄聯盟》這類語意結構高度依賴上下文的領域中，語句設計對多模態語意對齊能力具實質影響。此結果亦說明，未來在語句生成、提示設計（prompt design）與 query 編碼策略上，應考慮語意細緻度與資料結構之間的匹配關係，以提升精華檢測任務之效果。

三、討論與研究限制

本研究嘗試結合大型視覺語言模型（LVLM）與多模態檢索架構，用以處理語意密度高、結構邏輯嚴謹的《英雄聯盟》電競賽事影片。實驗透過語句型態、預訓練策略與模態組合等變項，全面檢視模型於精彩片段偵測（HD）與片段檢索（MR）任務上的表現。然而，整體指標準確率仍未達實務應用門檻，顯示目前主流 HL/MR 模型與 LVLM 於高結構語境中的適應性仍存在明顯限制。

相較於常見生活化影片或通用視覺資料集，《英雄聯盟》賽事影片具備高度結構特性。該語境內的語意判斷須依賴事件順序、角色定位與比賽進程等背景資訊。例如「搶巴龍」雖為關鍵事件，但若無法整合隊伍經濟差、陣容佈局與前後文轉折，其精華價值即難以正確辨識。這類語境理解對多模態檢索模型而言，已不僅是視覺分類或片段定位問題，更是對跨模態語意對齊與上下文推理能力的實質考驗。

LVLM 雖在開放場景下展現優異語句生成與語意整合能力，但本研究中採用的 Gemini 模型生成之敘事語句，仍偶有語意不清、邏輯錯置的情況，對模型訓練品質產生潛在干擾。此外，即使進行了弱監督預訓練或多模態特徵擴增，整體模型

表現亦僅有小幅提升，顯示現有方法難以充分捕捉高結構語境下的語意脈絡與事件邏輯。

另一方面，模態組合策略仍有待精煉。雖實驗顯示移除 CLIP 特徵後表現反而上升，可能來自該模態與任務目標間存在語意落差或冗餘資訊干擾，但尚無法明確界定其原因。後續研究可進一步探索模態選擇標準與融合方法，並搭配語境感知機制進行調整。

雖本研究於多組變項下建立初步驗證框架，並實證現有方法於高結構任務下之挑戰性，惟在實驗資料規模、語句語意品質與語境建模能力上仍有明顯進步空間。未來研究可朝向下列方向深化：（1）建構專為電競賽事語境設計的語句生成模型；（2）發展具語境記憶與邏輯推理能力的模態融合策略；（3）擴充賽事資料來源與標註規模，以提升訓練覆蓋與泛化能力。如此方能逐步突破現有模型於結構化語境中所面臨的瓶頸，實現更準確、語意一致的多模態語句匹配與影片精華識別功能。

四、未來發展

本研究揭示在《英雄聯盟》賽事影片等高結構語境中，現有多模態模型與大型視覺語言模型（LVLM）所面臨的適應性挑戰。儘管透過語句設計、模態組合與弱監督預訓練等策略，可部分提升模型效能，但整體表現距離實務需求仍有顯著落差。因此，未來研究可從以下三個方向進一步深化：

(一).語句生成機制之領域專化：

現有 LVLM（如 Gemini、BLIP）雖能生成結構正確的描述句，但對電競賽事中專業術語與語境邏輯掌握力仍顯不足，可能導致語句語意與片段事件不符。未來可考慮微調現有 LVLM 或設計具賽事語境感知能力的語句生成架構，提升敘述語句的語意準確度與任務相關性，進而增強弱監督學習的信號品質。

(二).語境感知與語意推理能力的融合機制：

多數現有 HL/MR 模型雖在通用任務上表現優異，但缺乏針對結構化事件的語境建模與語意推理能力。在電競賽事中，片段的重要性常仰賴上下文脈絡與時序發展。未來可強化模型對賽事邏輯的掌握與跨時間片段的推理能力。

(三).資料維度優化與特徵強化策略：

本研究結果顯示，並非納入更多模態即能穩定提升表現，部分模態如 CLIP 在高語境結構的任務中可能引入資訊干擾。因此，未來研究應重視特徵模態的選擇與精緻化處理策略。

陸、參考文獻

Sun, H., Zhou, M., Chen, W., & Xie, W. (2024, March). Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 5, pp. 4998-5007).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., ... & Li, H. (2022, October). Frozen clip models are efficient video learners. In *European Conference on Computer Vision* (pp. 388-404). Cham: Springer Nature Switzerland.

Ju, C., Han, T., Zheng, K., Zhang, Y., & Xie, W. (2022, October). Prompting visual-language models for efficient video understanding. In *European conference on computer vision* (pp. 105-124). Cham: Springer Nature Switzerland.

Rasheed, H., Khattak, M. U., Maaz, M., Khan, S., & Khan, F. S. (2023). Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6545-6554).

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2022). Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, *508*, 293-304.

Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., & Chen, C. W. (2024, September). r 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *European Conference on Computer Vision* (pp. 421-438). Cham: Springer Nature Switzerland.

Liu, M., Wang, X., Nie, L., He, X., Chen, B., & Chua, T. S. (2018, June). Attentive moment retrieval in videos. In *The 41st international ACM SIGIR conference on*

research & development in information retrieval (pp. 15-24).

Moon, W., Hyun, S., Park, S., Park, D., & Heo, J. P. (2023). Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23023-23033).

Lei, J., Berg, T. L., & Bansal, M. (2021). Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, *34*, 11846-11858.

Liu, Y., Li, S., Wu, Y., Chen, C. W., Shan, Y., & Qie, X. (2022). Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3042-3051).

Paul, D., Parvez, M. R., Mohammed, N., & Rahman, S. (2024). VideoLights: Feature Refinement and Cross-Task Alignment Transformer for Joint Video Highlight Detection and Moment Retrieval. *arXiv preprint arXiv:2412.01558*.

Chu, W. T., & Chou, Y. C. (2017). On broadcasted game video analysis: event detection, highlight detection, and highlight forecast. *Multimedia Tools and Applications*, *76*(7), 9735-9758.

Wang, L., Sun, Z., Yao, W., Zhan, H., & Zhu, C. (2019). Unsupervised Multi-stream Highlight detection for the Game" Honor of Kings". *arXiv preprint arXiv:1910.06189*.

Liaw, C. M., & Dai, B. R. (2020, June). Live stream highlight detection using chat messages. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)* (pp. 328-332). IEEE.

Chen, C. C., Lo, L. W., & Lin, S. J. (2022). COHETS: A highlight extraction method using textual streams of streaming videos. *Knowledge-Based Systems*, *258*, 110000.

Ringer, C., Nicolaou, M. A., & Walker, J. A. (2022). Autohighlight: Highlight detection in League of Legends esports broadcasts via crowd-sourced data. *Machine Learning with Applications*, *9*, 100338.

Song, Y. (2016). Real-time video highlights for yahoo esports. *arXiv preprint arXiv:1611.08780*.

Han, H. K., Huang, Y. C., & Chen, C. C. (2019, December). A deep learning model for extracting live streaming video highlights using audience messages. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (pp. 75-81).

Liu, Y., Ni, X., & Niu, G. (2021). Perceived stress and short-form video application addiction: a moderated mediation model. *Frontiers in Psychology*, *12*, 747656.

Wang, Y. (2020). Humor and camera view on mobile short-form video apps influence user experience and technology-adoption intent, an example of TikTok (DouYin). *Computers in human behavior*, *110*, 106373.

Wright, C. (2017). Are beauty bloggers more influential than traditional industry experts?. *Journal of Promotional Communications*, *5*(3).

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, *1*, 25-36.

Sun, M., Farhadi, A., & Seitz, S. (2014, September). Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision* (pp. 787-802). Cham: Springer International Publishing.

Mahasseni, B., Lam, M., & Todorovic, S. (2017). Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 202-211).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.