



**字幕中指示代名詞解析系統：
結合多模態視覺感知技術**

**DEMONSTRATIVE PRONOUN RESOLUTION SYSTEM IN
SUBTITLES: A MULTIMODAL VISUAL PERCEPTION APPROACH**

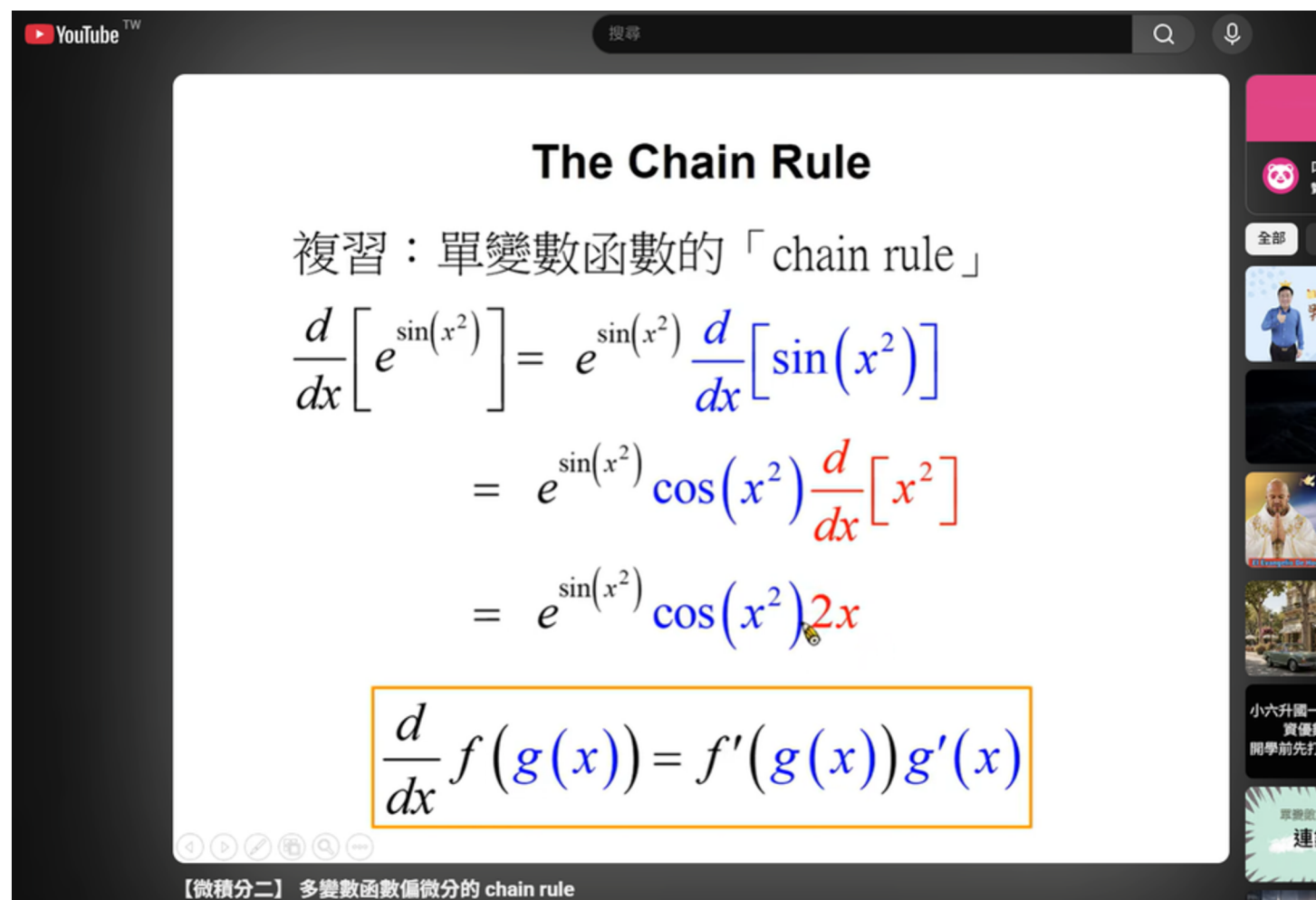
B-5

專題組員：練亭蘭

專題教授：李信杰 教授

Motivation - 研究動機與目的

1. 字幕無法表達畫面視覺資訊



The Chain Rule

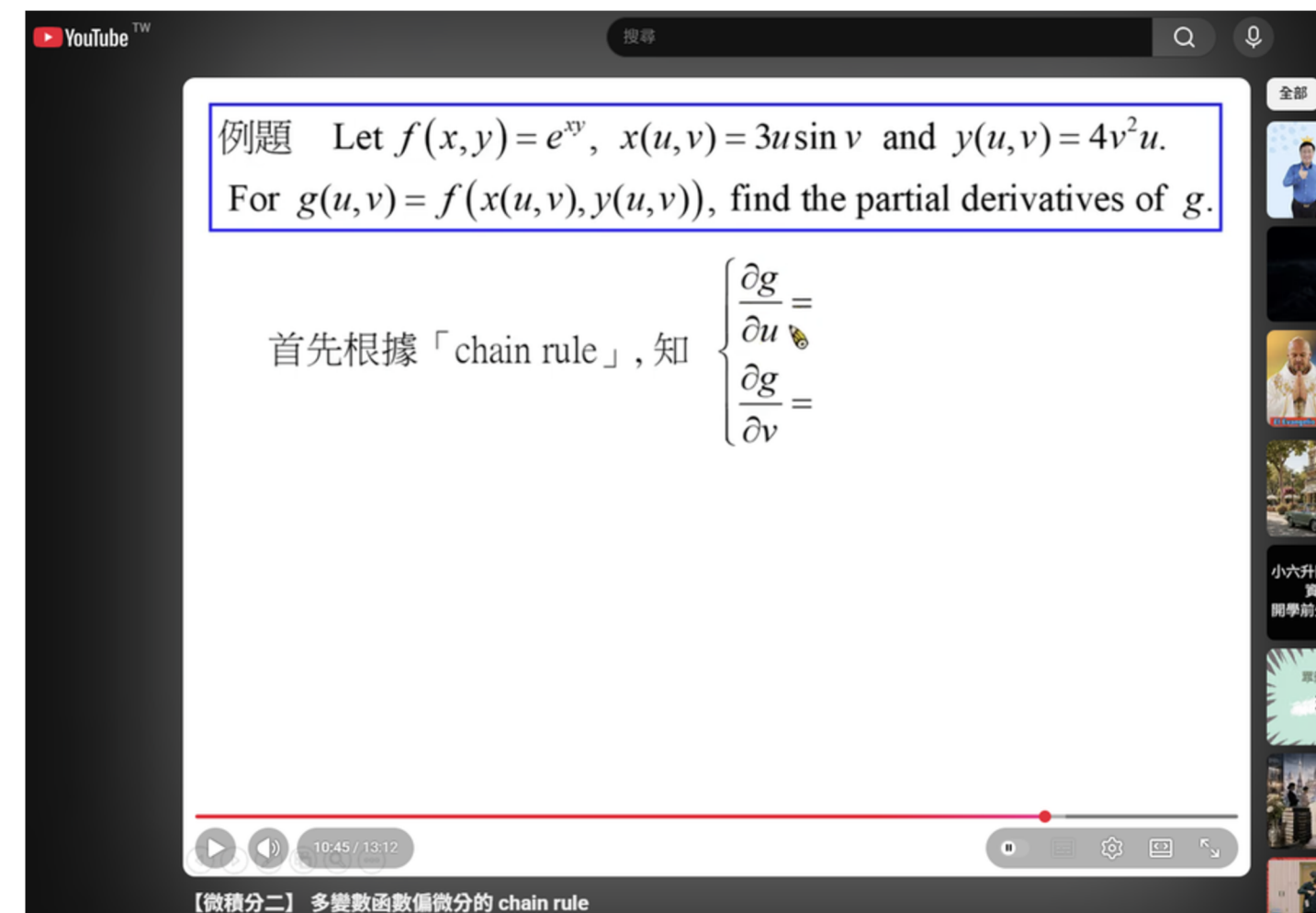
複習：單變數函數的「chain rule」

$$\begin{aligned}\frac{d}{dx} \left[e^{\sin(x^2)} \right] &= e^{\sin(x^2)} \frac{d}{dx} [\sin(x^2)] \\ &= e^{\sin(x^2)} \cos(x^2) \frac{d}{dx} [x^2] \\ &= e^{\sin(x^2)} \cos(x^2) 2x\end{aligned}$$
$$\frac{d}{dx} f(g(x)) = f'(g(x)) g'(x)$$

【微積分二】多變數函數偏微分的 chain rule

- 字幕殘缺
- 講者：「接下來我們再對這個藍色的部分進行微分，」無法透過文字了解「這個藍色的部分」代表什麼

2. 沒有完整前後文，無法理解內容



例題 Let $f(x, y) = e^{xy}$, $x(u, v) = 3u \sin v$ and $y(u, v) = 4v^2 u$.
For $g(u, v) = f(x(u, v), y(u, v))$, find the partial derivatives of g .

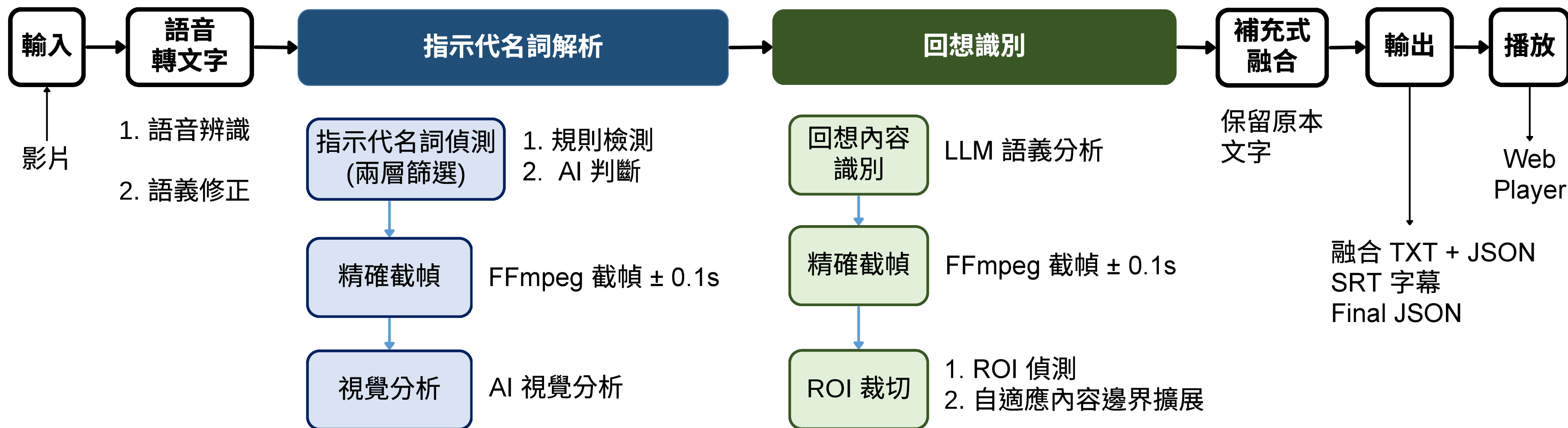
首先根據「chain rule」, 知 $\begin{cases} \frac{\partial g}{\partial u} = \\ \frac{\partial g}{\partial v} = \end{cases}$

【微積分二】多變數函數偏微分的 chain rule

- 講者：「根據前面的連鎖率」，沒有前後文，無法知道「前面的」代表什麼

Methodology - 技術與系統架構

使用開源框架與模型：Faster-Whisper (base), FFmpeg, Gemini 3.1 Flash Lite, Gemini 2.5 Flash Image



Results - 成果展示

The Chain Rule

複習：單變數函數的「chain rule」

$$\frac{d}{dx} \left[e^{\sin(x^2)} \right] = e^{\sin(x^2)} \frac{d}{dx} [\sin(x^2)]$$

$$\frac{d}{dx} f(g(x)) = f'(g(x)) g'(x)$$

然後再乘上，裡面這一層就是 $\sin x$ 平方的微分，接下來我們再對這個 $\frac{d}{dx} [\sin(x^2)]$ 藍色的部分進行微分，

回想內容

- 講者說「這個」
→ 系統解析為畫面中的 $\frac{d}{dx} \sin(x^2)$ ，並補充至字幕中

Results - 成果展示

例題 Let $f(x, y) = e^{xy}$, $x(u, v) = 3u \sin v$ and $y(u, v) = 4v^2 u$.
For $g(u, v) = f(x(u, v), y(u, v))$, find the partial derivatives of g .

首先根據「chain rule」, 知 $\begin{cases} \frac{\partial g}{\partial u} = \\ \frac{\partial g}{\partial v} = \end{cases}$

回想內容

定理 (chain rule)

若 $z = f(x, y)$,

且 $x = x(s, t)$, $y = y(s, t)$ 則

$$\frac{\partial z}{\partial s} = \frac{\partial f}{\partial x}(x, y) \frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}(x, y) \frac{\partial y}{\partial s}$$

Direct

「二元自變數連鎖率公式」

信心: 100%

- 講者說「根據前面所說的」
→ 系統自動定位講者回想的哪一個
先前段落
- 於字幕處標註時間 (↑ [0:08:24])
- 於回想內容欄位附上可跳轉的畫
面截圖、概念簡述與信心度
- 藍色框暗示使用者可以點擊跳轉
影片回該概念(0:08:24)

▶ 10:42 / 13:11

[↑ [0:08:24] z 現在是代表這整個合成函數，對於自變數 s 的偏微，就是 ...]

根據前面所說的連鎖率，

Conclusion - 結論與未來展望

結論：

- 首次整合語音辨識 + 多模態視覺感知
 - 解決逐字稿遺失畫面資訊的問題
 - 系統透過視覺分析解析指示代名詞所指對象
 - 回想識別可定位回想內容並提取對應畫面與區域裁切圖供參考
- 系統搭配網頁播放器呈現影片、字幕與回想內容

未來展望：

- 提升指示代名詞描述的精準度
- 提升回想畫面截圖區域的精準度
- 指標偵測（如滑鼠或雷射筆位置）
- 更多輸出格式支援

Demo Video:



GitHub:

