PPTAgent：超越文字轉投影片的簡報生成與評估

Hao Zheng [1,2,*], Xinyan Guan [1,2,*], Hao Kong [3], Jia Zheng [1], Weixiang Zhou [1]
Hongyu Lin [1],Yaojie Lu [1], Ben He [1,2], Xianpei Han [1], Le Sun [1]
[1] 中國科學院軟體研究所中文資訊處理實驗室
[2] 中國科學院大學
[3] 上海捷鑫科技
{zhenghao2022,guanxinyan2022,zhengjia,
weixiang,hongyu,luyaojie}@iscas.ac.cn
{xianpei,sunle}@iscas.ac.cn haokong@knowuheart.com
benhe@ucas.edu.cn

摘要

從文件中自動生成簡報是一項具有挑戰性的任務，需要兼顧內容品質、視覺吸引力和結構連貫性。現有方法主要側重於單獨改進和評估內容品質，卻忽略了視覺吸引力和結構連貫性，這限制了它們的實際應用性。為了解決這些限制，我們提出了 PPTAgent，它透過受人類工作流程啟發的兩階段、基於編輯的方法，全面改進簡報生成。PPTAgent 首先分析參考簡報以提取投影片層級的功能類型和內容架構，然後草擬大綱並根據選定的參考投影片迭代生成編輯動作以創建新的投影片。為了全面評估生成簡報的品質，我們進一步引入了 PPTEval，這是一個評估框架，從三個維度評估簡報：內容、設計和連貫性。結果表明，PPTAGENT 在所有三個維度上都顯著優於現有的自動簡報生成方法。程式碼和資料可在 https://github.com/ icip-cas/PPTAgent 取得。

1 導論

簡報是一種廣泛使用的資訊傳遞媒介，因其在吸引和與觀眾溝通方面的視覺效果而受到重視。然而，製作高品質的簡報需要引人入勝的故事情節、精心設計的版面配置以及豐富、引人注目的內容（Fu 等人，2022）。因此，製作全面的簡報需要進階的簡報技巧和大量的精力。鑑於簡報製作固有的複雜性，人們對自動化簡報生成過程的興趣日益濃厚（Ge 等人，2025；Maheshwari 等人，2024；Mon-

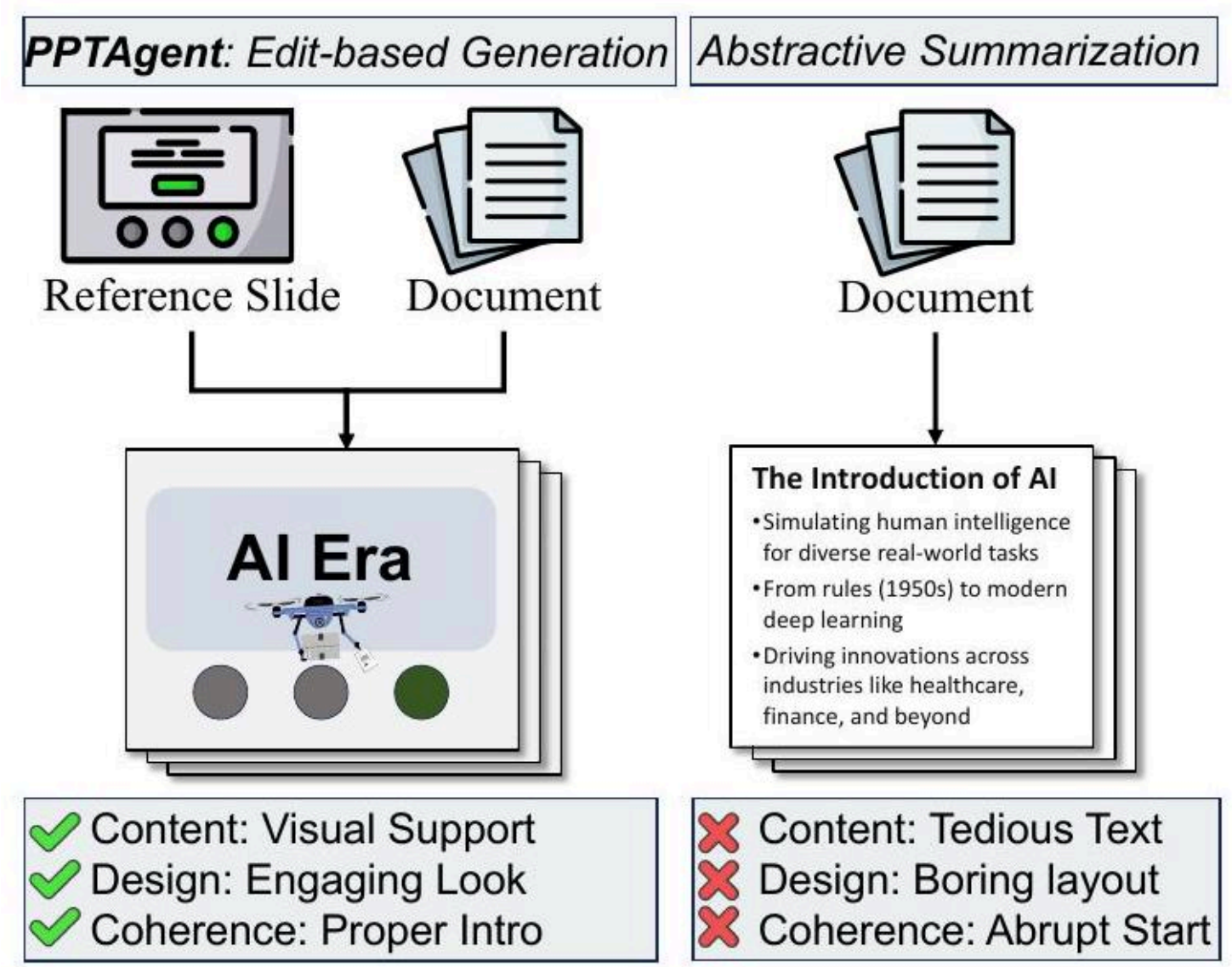


圖 1：圖 1：我們的 PPTAgent 方法（左）與傳統的抽象摘要方法（右）之比較。

dal 等人，2024）透過利用大型語言模型（LLMs）和多模態大型語言模型（MLLMs）的泛化能力。

現有方法通常遵循「文字轉投影片」的範式，即使用預定義的規則或範本將 LLM 輸出轉換為投影片。如圖 1 所示，先前的研究（Mondal 等人，2024；Sefid 等人，2021）傾向於將簡報生成視為一種抽象摘要任務，主要關注文字內容，而忽略了簡報以視覺為中心的本質（Fu 等人，2022）。這導致簡報內容過於冗長且單調，無法有效吸引觀眾（Barrick 等人，2018）。

人類的工作流程通常不是一次性從頭開始建立複雜的簡報，而是選擇範例投影片作為參考，然後將關鍵內容摘要並轉移到這些投影片上（Duarte，2010）。受此過程啟發，我們提出了 PPTAGENT，它將投影片生成分解為兩個階段：選擇參考投影片和逐步編輯。然而，實現這種基於編輯的簡報生成方法具有挑戰性。首先，由於

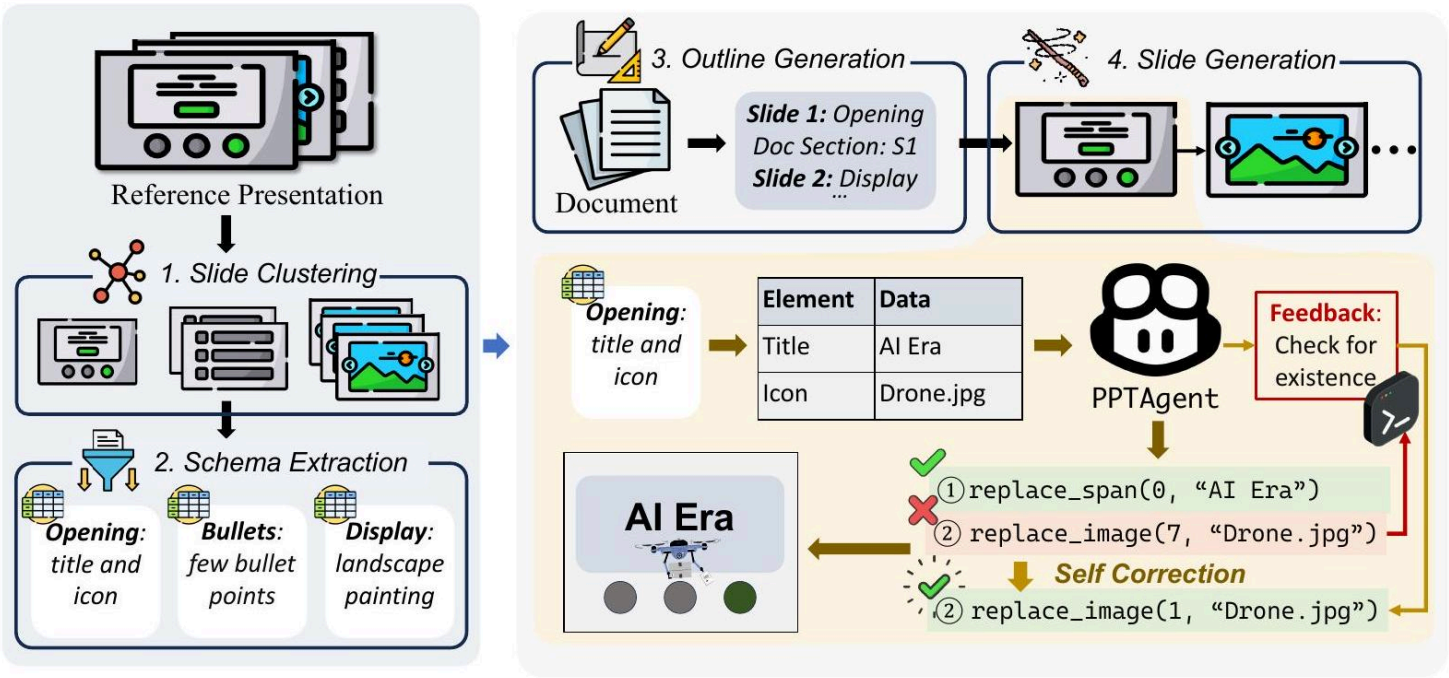


圖 2：PPTAgent 工作流程概述。階段一：簡報分析，涉及分析輸入簡報以將投影片分組並提取其內容綱要。階段二：簡報生成，根據大綱生成新簡報，並納入自我校正機制以確保穩健性。

簡報的版面配置和模式複雜性，使得 LLMs 難以直接判斷應參考哪些投影片。關鍵挑戰在於增強 LLMs 對參考簡報結構和內容模式的理解。其次，大多數簡報都以 PowerPoint 的 XML 格式儲存，如圖 11 所示，這種格式本質上冗長且多餘 (Gryk, 2022)，使得 LLMs 難以穩健地執行編輯操作。

為了解決這些挑戰，PPTAgent 分兩個階段運作。階段一對參考簡報進行全面分析，以提取投影片的功能類型和內容綱要，從而促進後續的參考選擇和投影片生成。階段二引入了一套帶有 HTML 渲染表示的編輯 API，透過程式碼互動簡化投影片修改 (Wang et al., 2024b)。此外，我們實施了自我校正機制 (Kamoi et al., 2024)，允許 LLMs 根據中間結果和執行回饋迭代地完善生成的編輯動作，確保穩健的生成。如圖 2 所示，我們首先分析並將參考投影片分組（例如，開場投影片、項目符號投影片）。對於每張新投影片，PPTAgent 會選擇一張適當的參考投影片（例如，第一張投影片的開場投影片）並生成一系列編輯動作（例如，replace_span）來修改它。

由於缺乏全面的評估
的評估框架中，我們提出了 PPTEval，它採用 MLLM-as-a-judge 範式（Chen et al., 2024a）來評估簡報的內容、設計和連貫性三個維度（Duarte, 2010）。人類評估驗證了 PPTEval 的可靠性和有效性。結果顯示，PPTAGENT 生成了高品質的簡報，在 PPTEval 的三個維度中平均得分為 3.67。

我們的主要貢獻可歸納如下：

我們提出了 PPTAgent，這是一個將自動簡報生成重新定義為以參考簡報為指導的編輯式流程的框架。
我們引入了 PPTEval，這是一個全面的評估框架，用於評估簡報的內容、設計和連貫性三個維度。
我們發布了 PPTAgent 和 PPTEval 程式碼庫，以及一個新的簡報資料集 Zenodo10K，以支援未來的研究。

2 PPTAgent

在本節中，我們闡述了簡報生成任務，並介紹了我們提出的 PPTAGENT 框架，該框架包含兩個不同的階段。在第一階段，我們透過投影片聚類和模式提取來分析參考簡報，從而全面理解
輸入簡報，這有助於後續的參考選擇和投影片生成。在第二階段，我們利用分析過的參考簡報來選擇參考投影片，並透過迭代編輯過程為輸入文件生成目標簡報。我們的工作流程概述如圖 2 所示。

2.1 問題闡述

PPTAGENT 旨在透過編輯流程生成引人入勝的簡報。我們提供傳統方法和 PPTAGENT 的正式定義，以突顯其主要差異。

用於建立每個投影片 $\boldsymbol{S}$ 的傳統方法（Bandyopadhyay 等人，2024；Mondal 等人，2024）在公式 1 中形式化。給定輸入內容 C ，它會生成 n 個投影片元素，每個元素都由其類型、內容和樣式屬性定義，例如（文字方塊，「Hello」，{邊框、大小、位置，...}）。

$$\boldsymbol{S} = \{e_1, e_2, \ldots, e_n\} = f(C)$$

儘管這種傳統方法很直接，但它需要手動指定樣式屬性，這對於自動生成來說具有挑戰性（Guo 等人，2023）。PPTAGENT 不會從頭開始建立投影片，而是生成一系列可執行動作來編輯參考投影片，從而保留其精心設計的版面配置和樣式。如公式 2 所示，給定輸入內容 C 和從參考簡報中選取的第 j 個參考投影片 R_j ，PPTAGENT 會生成一系列 m 個可執行動作，其中每個動作 a_i 都對應一行可執行程式碼。

$$\boldsymbol{A} = \{a_1, a_2, \ldots, a_m\} = g\left(C, R_j\right)$$

2.2 階段一：簡報分析

在此階段，我們分析參考簡報以引導參考選擇和投影片生成。首先，我們透過投影片聚類，根據其結構和版面配置特性對投影片進行分類。然後，我們提取內容綱要以識別每個聚類中投影片的內容組織，提供投影片元素的全面描述。

投影片聚類 投影片可根據其功能分為兩大類：支援簡報組織的結構性投影片（例如，開場投影片）和傳達特定資訊的內容投影片（例如，條列式
投影片）。為了區分這兩種類型，我們採用 LLMs 來相應地分割簡報。對於結構性投影片，我們利用 LLMs 的長上下文能力來分析輸入簡報中的所有投影片，識別結構性投影片，根據其文本特徵標記其結構角色，並相應地進行分組。對於內容投影片，我們首先將其轉換為圖像，然後應用分層聚類方法來分組相似的投影片圖像。隨後，我們利用 MLLMs 來分析轉換後的投影片圖像，識別每個聚類中的版面配置模式。更多詳細資訊請參閱附錄 D。

綱要提取 在分群之後，我們進一步分析了它們的內容綱要，以利投影片的生成。具體來說，我們定義了一個提取框架，其中每個元素都由其類別、描述和內容表示。這個框架能夠清晰且有條理地呈現每張投影片。詳細說明請參閱附錄 F，綱要範例如下所示。

類別	描述	資料
標題	主標題	範例庫
日期	活動日期	2018 年 2 月 15 日
圖片	主要圖片至 說明投影片	圖片：圖書館裡的孩童， 他們正在...

2.3 第二階段：簡報生成

PPTAGENT 首先會生成一份大綱，其中會為每個新投影片指定參考投影片和相關內容。然後，它會透過編輯 API 疊代編輯參考投影片中的元素，以建立目標簡報。

大綱生成 如圖 2 所示，我們利用 LLM 生成一個由多個條目組成的結構化大綱。每個條目代表一個新投影片，其中包含新投影片的參考投影片和相關文件內容。參考投影片是根據第一階段的投影片級功能描述來選擇的，而相關文件內容則是根據輸入文件來識別的。

投影片生成 在結構化大綱的引導下，投影片會根據相應的條目疊代生成。對於每個投影片，LLMs 會整合輸入文件中的文字內容和提取的圖片說明。新投影片會採用參考投影片的版面配置，同時確保內容的一致性和結構的清晰度。

具體來說，為了根據大綱中相應的條目生成新投影片，我們設計了基於編輯的 API，讓 LLMs 能夠編輯參考投影片。如下所示，這些 API 支援編輯、移除和複製投影片元素。此外，考量到簡報中 XML 格式的複雜性（如附錄 E 所示），我們將參考投影片渲染成 HTML 格式（Feng 等人，2024），提供更精確、更直觀的格式，以便於理解。這種基於 HTML 的格式，結合我們基於編輯的 API，使 LLMs 能夠對參考投影片執行精確的內容修改。

函數名稱	描述
del_span	刪除一個 span。
del_image	刪除圖片元素。
clone_paragraph	建立現有段落的副本 段落。
替換 span	替換 span 的內容。
替換圖片	替換圖片來源。

此外，為了在編輯過程中增強穩健性，我們實施了自我修正機制（Kamoi et al., 2024）。具體來說，生成的編輯動作會在 REPL 環境中執行。當動作未能應用於參考投影片時，REPL 會提供執行回饋，以協助 LLMs 完善其動作。LLM 隨後會分析此回饋，以調整其編輯動作（Guan et al., 2024; Wang et al., 2024b），從而實現迭代精煉，直到生成有效的投影片或達到最大重試限制。

3 PPTEval

我們推出了 PPTEval，這是一個全面的框架，可從多個維度評估簡報品質，解決了簡報缺乏無參考評估的問題。該框架提供數值分數（1 到 5 分）和詳細的理由，以證明每個維度的評估。

我們的評估框架以既定的簡報設計原則（Duarte, 2008, 2010）為基礎，著重於三個關鍵維度，如表 1 所示。具體而言，對於一份生成的簡報，我們評估投影片層級的內容與設計，同時評估整個簡報的連貫性。

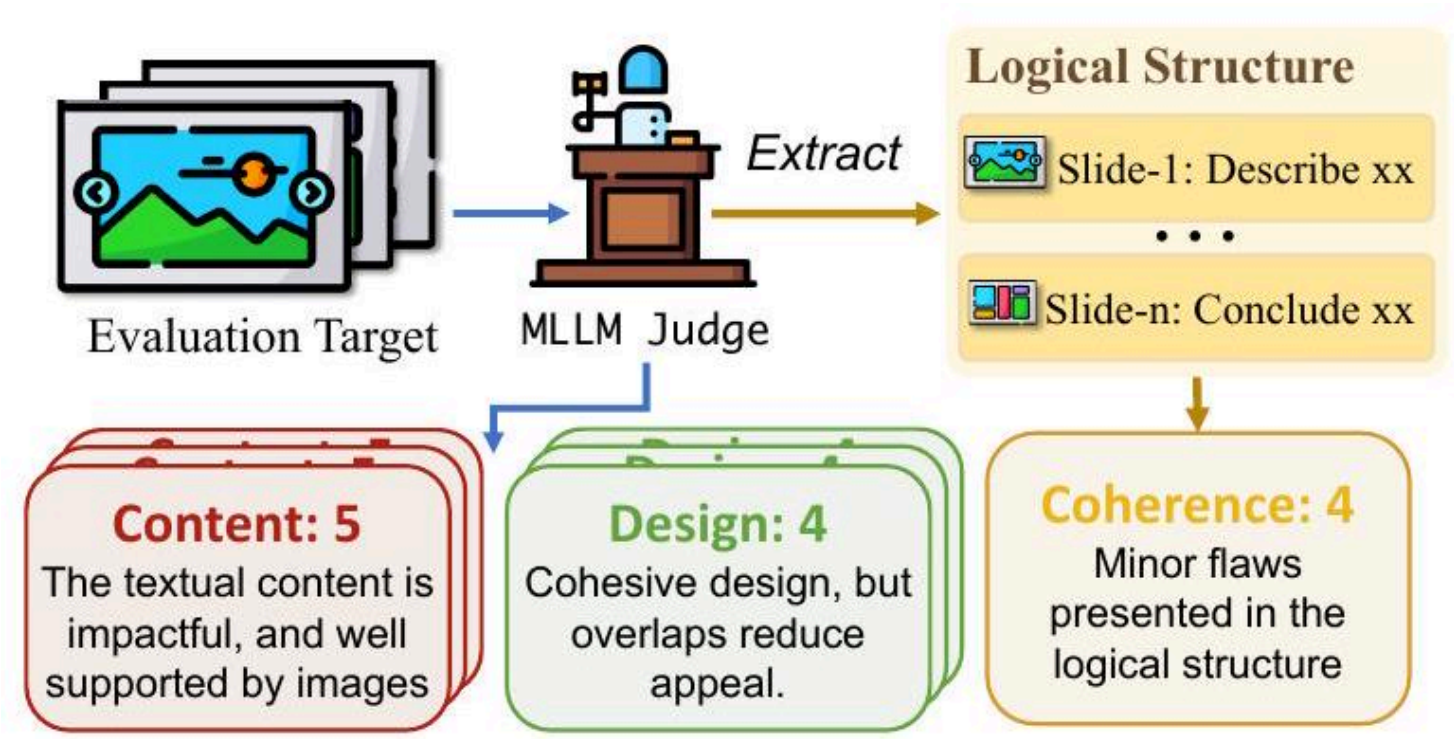


圖 3：圖 3：PPTEval 從內容、設計和連貫性三個維度評估簡報。

完整的評估流程如圖 3 所示，詳細的評分標準和代表性範例請參閱附錄 B。

維度	標準
內容	文字應簡潔且語法正確，並輔以相關圖片。
設計	和諧的色彩和適當的排版確保了可讀性，而幾何圖形等視覺元素則增強了整體吸引力。
連貫性	結構逐步發展，並納入必要的背景資訊。

表 1：表 1：PPTEVAL 中各維度的評分標準，所有評分均為 1-5 分。

4 實驗

4.1 資料集

現有的簡報資料集，例如 Fu 等人 (2022)；Mondal 等人 (2024)；Sefid 等人 (2021)；Sun 等人 (2021)，存在兩個主要問題。首先，它們大多以 PDF 或 JSON 格式儲存，這導致語義資訊的遺失，例如元素的結構關係和樣式屬性。此外，這些資料集主要由人工智慧領域的學術簡報組成，限制了其多樣性。為了解決這些限制，我們引入了 Zenodo10K，這是一個來自 Zenodo (歐洲核子研究組織和 OpenAIRE, 2013) 的新資料集，該平台託管了跨領域的各種文物，所有這些都具有明確的授權。我們從這個來源整理了 10,448 份簡報，並將其公開，以支持進一步的研究。

根據 Mondal 等人 (2024) 的做法，我們從五個領域中抽取了 50 份簡報作為參考簡報。此外，我們從相同領域收集了 50 份文件作為輸入文件。抽樣標準和預處理細節請參閱附錄 A，而資料集統計數據則彙總於表 2。

領域	文件		簡報		
	#Chars	#Figs	#Chars	#Figs	#Pages
文化	12,708	2.9	6,585	12.8	14.3
教育	12,305	5.5	3,993	12.9	13.9
科學	16,661	4.8	5,334	24.0	18.4
社會	13,019	7.3	3,723	9.8	12.9
科技	18,315	11.4	5,325	12.9	16.8

表 2：我們實驗中使用的資料集統計，詳細列出了字元數（'#Chars'）和圖形數（'#Figs'），以及頁數（'#Pages'）。

4.2 實作細節

PPTAGENT 採用三種模型實作：GPT-4o-2024-08-06 (GPT-4o)、Qwen2.5-72BInstruct (Qwen2.5, Yang et al., 2024) 和 Qwen2-VL-72B-Instruct (Qwen2-VL, Wang et al., 2024a)。這些模型根據其處理的特定模態（文字或視覺）進行分類，如其下標所示。具體來說，我們將配置定義為語言模型 (LM) 和視覺模型 (VM) 的組合，例如 Qwen2.5 LM + Qwen2-VLVM。

實驗資料涵蓋 5 個領域，每個領域有 10 個輸入文件和 10 個參考簡報，每個配置總共有 500 個簡報生成任務（5 個領域 $\times$10 輸入文件 $\times$10 參考簡報）。每個投影片生成最多允許兩次自我修正迭代。我們使用 Chen et al. (2024b) 和 Wu et al. (2020) 分別計算文字和圖像嵌入。所有開源 LLMs 均使用 VLLM 框架 (Kwon et al., 2023) 部署在 NVIDIA A100 GPU 上。實驗的總計算成本約為 500 GPU 小時。

4.3 基準線

我們選擇以下基準線方法：DocPres (Bandyopadhyay et al., 2024) 提出了一種基於規則的方法，透過多階段生成敘事豐富的投影片，並透過基於相似性的機制整合圖像。KCTV (Cachola et al., 2024) 提出了一種基於範本的方法，該方法在將投影片轉換為最終簡報之前，先以中間格式建立投影片，然後再使用預定義的範本。這些基準線方法在沒有視覺模型的情況下運作，因為它們不處理視覺資訊。每個配置都會生成 50 個簡報（5 個領域 $\times$ 10 個輸入文件），因為它們不需要參考簡報。因此，FID 指標被排除在它們的評估之外。

4.4 評估指標

我們使用以下指標評估簡報生成：

成功率（SR）評估簡報生成（Wu 等人，2024）的穩健性，計算方式為成功完成任務的百分比。對於 PPTAgent 而言，成功需要生成所有投影片，且在自我修正後沒有執行錯誤。對於 KCTV 而言，成功取決於生成之 LaTeX 檔案的成功編譯。DocPres 因其確定性的基於規則的轉換，故不納入此評估。
困惑度（PPL）衡量模型生成給定序列的可能性。我們使用 Llama-3-8B（Dubey 等人，2024）計算簡報中所有投影片的平均困惑度。較低的困惑度分數表示較高的文本流暢度（Bandyopadhyay 等人，2024）。
Rouge-L（Lin，2004）透過衡量生成文本與參考文本之間的最長共同子序列來評估文本相似度。我們報告 F1 分數以平衡精確度與召回率。
FID（Heusel 等人，2017）衡量生成簡報與參考簡報在特徵空間中的相似度。由於樣本量有限，我們使用 64 維輸出向量計算 FID。

PPTEval 採用 GPT-4o 作為評審模型，從內容、設計和連貫性三個維度評估簡報品質。我們透過計算各投影片的平均分數來得出內容和設計分數，而連貫性則在簡報層級進行評估。

4.5 整體結果

表 3 呈現了 PPTAgent 與基準線之間的效能比較，結果顯示：

PPTAgent 顯著提升了整體簡報品質。PPTAgent 在 PPTEval 的所有三個維度上，相較於基準線方法，展現出統計上顯著的效能提升。與基於規則的基準線 (DocPres) 相比，PPTAgent 在設計和內容維度上均有顯著改進 (3.34 對比 $2.37, +40.9\%; 3.34$ 對比 2.98， $+12.1\%$)，因為 DocPres 方法生成的簡報在設計方面投入極少。與基於範本的基準線 (KCTV) 相比，PPTAGENT 在設計和內容方面也取得了顯著改進 (3.34 對比 $2.95, +13.2\%; 3.28$ 對比 $2.55, +28.6\%$)，這突顯了編輯的有效性。

設定		現有指標				PPTEVAL			
語言模型	視覺模型	SR(%) ↑	PPL ↓	ROUGE-L ↑	FID ↓	內容 ↑	設計 ↑	連貫性 ↑	平均 ↑
文件呈現 (基於規則)									
GPT-4 $_{\text{LM}}$	-	-	76.42	13.28	-	2.98	2.33	3.24	2.85
Qwen2.5 $_{\text{LM}}$	-	-	100.4	13.09	-	2.96	2.37	3.28	2.87
KCTV (基於範本)									
GPT-4 o_{LM}	-	80.0	<u>68.48</u>	10.27	-	2.49	2.94	3.57	3.00
Qwen2.5 $_{\text{LM}}$	-	88.0	41.41	16.76	-	2.55	2.95	3.36	2.95
PPTAgent (我們)									
GPT-4 o_{LM}	GPT-4% $_{\text{VM}}$	97.8	721.54	10.17	7.48	<u>3.25</u>	3.24	<u>4.39</u>	<u>3.62</u>
Qwen2-VL $_{LM}$	Qwen2-VL $_{\text{VM}}$	43.0	265.08	13.03	<u>7.32</u>	3.13	3.34	4.07	3.51
Qwen2.5 $_{\text{LM}}$	Qwen2-VL $_{\text{VM}}$	95.0	496.62	<u>14.25</u>	6.20	3.28	3.27	4.48	3.67

表 3：簡報生成方法的效能比較，包括 DocPres、KCTV 和我們提出的 PPTAGENT。最佳/次佳分數以粗體/底線標示。結果使用現有指標報告，包括成功率 (SR)、困惑度 (PPL)、Rouge-L、Fréchet Inception Distance (FID) 和 PPTEval。

設定	成功率 (%)	內容	設計	連貫性	平均
PPTAGENT	**95.0**	**3.28**	3.27	**4.48**	**3.67**
無大綱	91.0	3.24	<u>3.30</u>	3.36	3.30
無 Schema	78.8	3.08	3.23	4.04	3.45
無 Structure	<u>92.2</u>	**3.28**	3.25	3.45	3.32
無 CodeRender	74.6	<u>3.27</u>	**3.34**	<u>4.38</u>	<u>3.66</u>

表 4：表 4：PPTAgent 使用 Qwen2.5 $_{\text{LM}}$+ Qwen2-VL $_{\text{VM}}$ 配置的消融分析，展示了每個組件的貢獻。

基礎典範。最值得注意的是，PPTAgent 在連貫性維度上顯示出顯著的提升（DocPres 為 4.48 對 $3.57, +25.5\%$ ；KCTV 為 4.48 對 3.28， $+36.6\%$)。這種改進可歸因於 PPTAGENT 對投影片結構角色的全面分析。

PPTAgent 展現強大的生成效能。我們的方法使 LLMs 能夠以卓越的成功率生成完善的簡報，Qwen2.5 $_{\text{LM}}$+ 、Qwen2-VL $_{\text{VM}}$ 和 GPT- 4_{LM}+ GPT 4_{OM} 的成功率均達到 $\geq 95\%$ ，與 KCTV 相比有顯著提升 （ 97.8% 對 88.0%)。此外，Qwen2.5 $_{\text{LM}} + \text{Qwen}_2 - \text{VL}_{\text{VM}}$ 在各個領域的詳細效能如表 8 所示，突顯了我們方法的多功能性和穩健性。

PPTEval 展現卓越的評估

能力。與 PPTEval 相比，傳統指標如 PPL 和 ROUGE-L 顯示出不一致的評估趨勢。例如，KCTV 實現了高 ROUGE-L (16.76) 但內容分數較低 (2.55)，而我們的方法則顯示出相反的趨勢，ROUGE-L (14.25) 和內容分數 (3.28)。此外，我們觀察到

ROUGE 分數過度強調文本與來源文件的對齊，可能會影響簡報的表達能力。最重要的是，PPTEval 透過其參考自由的設計評估和簡報連貫性的整體評估雙重能力，超越了現有的評估指標。進一步的一致性評估顯示在第 5.5 節。

5 分析

5.1 消融研究

我們在四種設定下進行了消融研究：(1) 隨機選擇一張投影片作為參考（無大綱），(2) 在大綱生成過程中省略結構性投影片（無結構），(3) 將投影片表示替換為 Guo 等人 (2023) 提出的方法（無 CodeRender），以及 (4) 移除內容綱要的指導（無綱要）。所有實驗均使用 Qwen2.5 LM＋ Qwen2-VL VM 配置進行。

如表 4 所示，我們的實驗揭示了兩個關鍵發現：1) 基於 HTML 的表示法顯著降低了互動複雜度，這可從移除程式碼渲染元件後，成功率從 95.0% 大幅下降至 74.6% 得到證明。2) 呈現分析對於生成品質至關重要，因為移除大綱和結構化投影片會顯著降低連貫性（從 4.48 降至 3.36/3.45 ），而移除投影片綱要則會使成功率從 95.0% 降至 78.8% 。

5.2 案例研究

我們在圖中展示了在不同配置下生成的簡報的代表性範例。

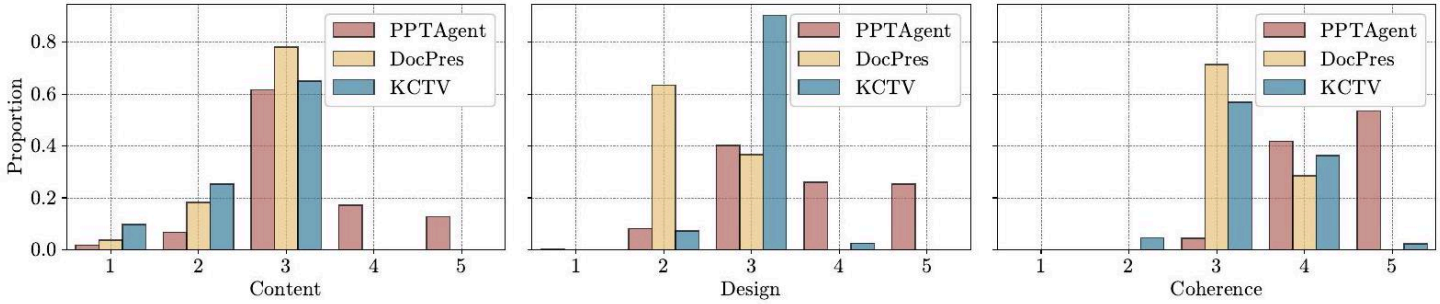


圖 4：圖 4：由 PPTAGENT、DocPres 和 KCTV 生成的簡報在內容、設計和連貫性這三個評估維度上的分數分佈，由 PPTEval 評估。

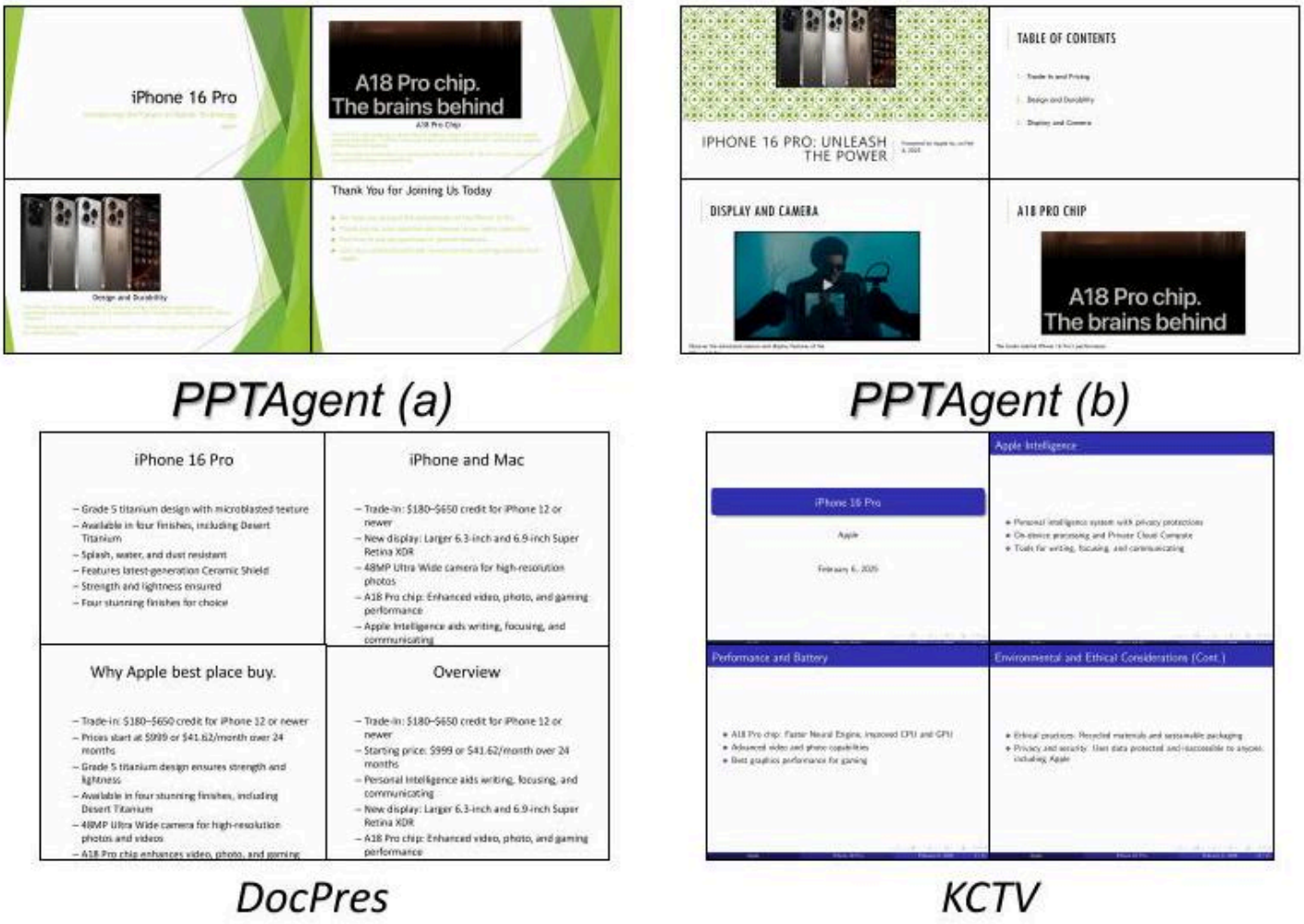


圖 5：不同方法簡報生成之比較分析。PPTAGENT 在不同參考簡報下生成，標示為 PPtAgent (a) 和 PPTAgent (b)。

圖 5. PPTAgent 在多個維度上展現出卓越的簡報品質。首先，它能有效地整合視覺元素，並將圖片放置於符合情境的位置，同時保持投影片內容的簡潔與結構良好。其次，它在多樣化的參考資料下，能生成視覺上引人入勝的投影片，展現出多樣性。相較之下，基準方法（DocPres 和 KCTV）主要生成以文字為主的投影片，視覺變化有限，受限於其基於規則或範本的模式。

5.3 分數分佈

我們進一步調查了生成簡報的分數分佈，以比較不同方法的性能特徵，如圖 4 所示。受限於其基於規則或範本的模式，基準方法在內容和設計維度上都表現出有限的多樣性，分數主要集中在 2 級和 3 級。相較之下，PPTAgent 展現出更分散的分數分佈，大多數簡報（$> 80\%$）在這些維度上獲得 3 分或更高的分數。此外，

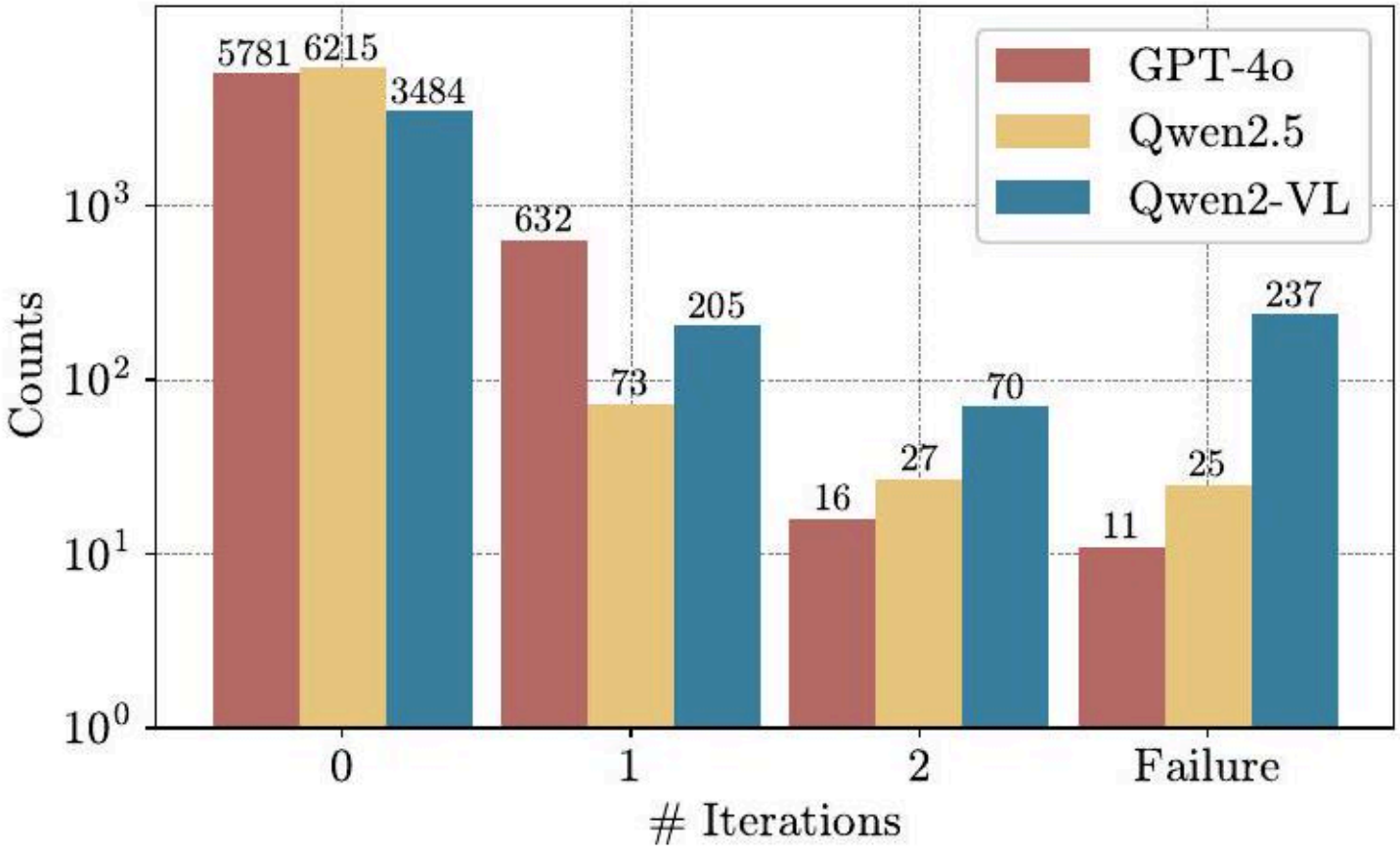


圖 6：不同模型生成單張投影片所需的迭代自我修正次數。

由於 PPTAgent 全面考量了結構化投影片，因此其在連貫性分數方面表現出色，超過 80% 的簡報獲得 4 分以上的成績。

5.4 自我修正的有效性

圖 6 說明了使用不同語言模型生成投影片所需的迭代次數。儘管 GPT-4o 展現出比 Qwen2.5 更優越的自我修正能力，但 Qwen2.5 在首次生成時遇到的錯誤較少。此外，我們觀察到 Qwen2-VL 更頻繁地出現錯誤，且自我修正能力較差，這可能是由於其多模態後訓練所致 (Wang et al., 2024a)。最終，所有三個模型都成功修正了超過一半的錯誤，這表明我們的迭代自我修正機制有效地確保了生成過程的成功。

5.5 評估一致性

PPTEval 與人類偏好 儘管 Chen 等人 (2024a) 強調了 LLMs 在各種生成任務中令人印象深刻的類人辨別能力。然而，在簡報的情境下，評估 LLM 評估與人類評估之間的相關性仍然至關重要。

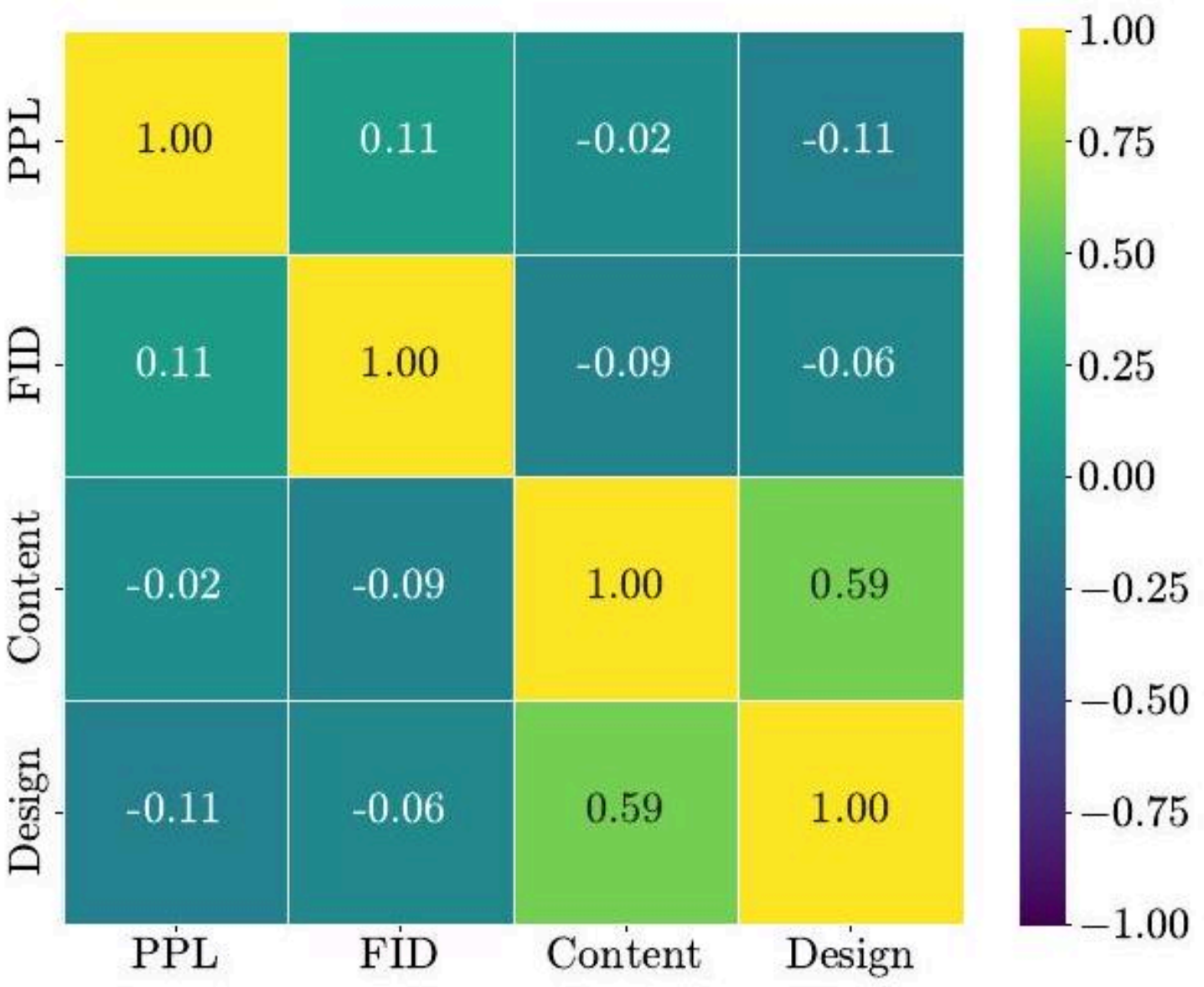


圖 7：圖 7：現有自動評估指標與 PPTEval 中內容和設計維度之間的相關性熱圖。

這種必要性源於 Laskar 等人 (2024) 的研究結果，該研究指出 LLMs 可能不足以評估複雜任務。表 5 顯示了人類與 LLMs 評分之間的相關性。平均 Pearson 相關係數為 0.71，超過了其他評估方法的分數 (Kwan 等人，2024)，這表明 PPTEval 與人類偏好高度一致。

PPTEval 與現有指標 我們透過皮爾森相關分析，分析了 PPTEval 的內容和設計維度與現有指標之間的關係，如圖 7 所示。皮爾森相關係數顯示，目前的指標對於簡報評估無效。具體來說，PPL 主要衡量文字流暢度，但由於投影片內容固有的碎片化性質，其在投影片內容上的表現不佳，經常產生異常測量結果。同樣地，雖然 ROUGEL 和 FID 分別量化了與參考文字和簡報的相似度，但這些指標未能充分評估內容和設計品質，因為高度符合參考資料並不能保證簡報的有效性。這些微弱的相關性突顯了 PPTEval 對於穩健且全面的簡報評估的必要性，該評估同時考慮了內容品質和設計有效性。

6 相關著作

自動簡報生成 近期提出的投影片生成方法可根據其處理元素放置和樣式的方式，分為基於規則和基於範本兩類。基於規則的方法，例如 Mondal 等人 (2024) 和 Bandyopadhyay 等人 (2024) 提出的方法，通常側重於增強文字內容，但忽略了簡報以視覺為中心的性質。

相關性	內容	設計	連貫性	平均
皮爾森	0.70	0.90	0.55	0.71
史皮爾曼	0.73	0.88	0.57	0.74

表五：人類評分與 LLM 評分在不同維度（連貫性、內容、設計）下的相關分數。所有呈現的相似性數據的 p 值均低於 0.05，表示統計顯著的置信水準。

，導致輸出缺乏吸引力。基於範本的方法，包括 Cachola 等人 (2024) 和像通義這樣的工業解決方案，依賴預定義的範本來創建視覺上吸引人的簡報。然而，它們對範本註釋的大量手動工作依賴性，顯著限制了可擴展性和靈活性。

LLM 代理人 許多研究 (Deng et al., 2024; Li et al., 2024; Tang et al., 2025) 探討了 LLMs 作為代理人協助人類完成各種任務的潛力。例如，Wang et al. (2024b) 展示了 LLMs 透過生成可執行動作來完成任務的能力。此外，Guo et al. (2023) 展示了 LLMs 透過 API 整合自動化簡報相關任務的潛力。

LLM 作為評審 LLMs 在指令遵循和語境感知方面展現出強大的能力，這使得它們被廣泛採用為評審 (Liu et al., 2023; Zheng et al., 2023)。Chen et al. (2024a) 展示了使用 MLLMs 作為評審的可行性，而 Kwan et al. (2024) 提出了一個多維度評估框架。此外，Ge et al. (2025) 研究了使用 LLMs 評估單張投影片品質的方法。然而，他們並未從整體角度評估簡報品質。

7 結論

在本文中，我們介紹了 PPTAgent，它將簡報生成概念化為一個兩階段的簡報編輯任務，透過 LLMs 理解和生成程式碼的能力來完成。此外，我們提出了 PPTEval，以提供評估簡報品質的量化指標。我們在多個領域的資料上進行的實驗證明了我們方法的優越性。這項研究為在無監督條件下生成投影片提供了一個新範例，並為簡報生成的未來工作提供了見解。

限制

儘管 PPTAgent 在簡報生成方面展現出有前景的能力，但仍存在一些限制。首先，儘管在我們的資料集上取得了高成功率（$> 95\%$），但模型偶爾會無法生成簡報，這可能會限制其可靠性。其次，雖然我們可以提供高品質的預處理簡報作為參考，但生成簡報的品質仍然受到輸入參考簡報的影響，這可能會導致次優的輸出。第三，儘管 PPTAgent 在版面最佳化方面比先前的方法有所改進，但它並未充分利用視覺資訊來完善投影片設計。這體現在偶爾的設計缺陷中，例如元素重疊，這可能會損害生成投影片的可讀性。未來的研究應著重於增強穩健性、減少對參考的依賴，並更好地將視覺資訊整合到生成過程中。

倫理考量

在建構 Zenodo10K 時，我們利用公開可用的 API 抓取資料，同時嚴格遵守每個工件相關的授權條款。具體來說，根據其各自授權不允許修改或商業使用的工件已被過濾掉，以確保符合智慧財產權。此外，所有參與專案的註釋人員都以超過其所在城市最低工資的費率獲得報酬，這反映了我們對公平勞動實踐和道德標準的承諾。

參考文獻

Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by 1 lms with a multi-staged end-to-end approach. arXiv preprint arXiv:2406.06556.

Andrea Barrick, Dana Davis, and Dana Winkler. 2018. Image versus text in powerpoint lectures: Who does it benefit? Journal of Baccalaureate Social Work, 23(1):91-109.

Isabel Alyssa Cachola, Silviu Cucerzan, Allen Herring, Vuksan Mijovic, Erik Oveson, and Sujay Kumar Jauhar. 2024. Knowledge-centric templatic views of documents. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages

15460-15476, 邁阿密，佛羅里達州，美國。計算語言學協會。

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. arXiv preprint arXiv:2402.04788.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36.

Nancy Duarte. 2008. Slide: ology: The art and science of creating great presentations, volume 1. O'Reilly Media Sebastapol.

Nancy Duarte. 2010. Resonate: Present visual stories that transform audiences. John Wiley & Sons.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

European Organization For Nuclear Research and OpenAIRE. 2013. Zenodo.

Weixi Feng、Wanrong Zhu、Tsu-jui Fu、Varun Jampani、Arjun Akula、Xuehai He、Sugato Basu、Xin Eric Wang 和 William Yang Wang。2024 年。Layoutgpt：使用大型語言模型進行組合式視覺規劃與生成。神經資訊處理系統進展，36。

Tsu-Jui Fu、William Yang Wang、Daniel McDuff 和 Yale Song。2022 年。Doc2ppt：從科學文獻自動生成簡報投影片。AAAI 人工智慧會議論文集，36(1):634-642。

Jiaxin Ge、Zora Zhiruo Wang、Xuhui Zhou、Yi-Hao Peng、Sanjay Subramanian、Qinyue Tan、Maarten Sap、Alane Suhr、Daniel Fried、Graham Neubig 等人。2025 年。Autopresent：從零開始設計結構化視覺效果。arXiv 預印本 arXiv:2501.00912。

Michael Robert Gryk。2022 年。資料檔案的人類可讀性。Balisage 標記技術系列，27。

Xinyan Guan、Yanjiang Liu、Hongyu Lin、Yaojie Lu、Ben He、Xianpei Han 和 Le Sun。2024 年。透過自主知識圖譜改造來緩解大型語言模型幻覺。收錄於《AAAI 人工智慧會議論文集》，第 38 卷，第 18126-18134 頁。

Yiduo Guo、Zekai Zhang、Yaobo Liang、Dongyan Zhao 和 Duan Nan。2023 年。Pptc 基準：評估大型語言模型在 PowerPoint 任務完成方面的表現。arXiv 預印本 arXiv:2311.01767。

Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017 年。透過雙時間尺度更新規則訓練的 GAN 會收斂到局部 Nash 平衡。收錄於《神經資訊處理系統進展》，第 30 卷。

Ryo Kamoi、Yusen Zhang、Nan Zhang、Jiawei Han 和 Rui Zhang。2024 年。LLMs 究竟何時能糾正自己的錯誤？對 LLMs 自我糾正的批判性調查。收錄於《計算語言學協會會刊》，第 12 卷，第 1417-1440 頁。

Wai-Chung Kwan、Xingshan Zeng、Yuxin Jiang、Yufei Wang、Liangyou Li、Lifeng Shang、Xin Jiang、Qun Liu 和 Kam-Fai Wong。2024 年。Mt-eval：大型語言模型的多輪能力評估基準。預印本，arXiv:2401.16745。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph Gonzalez、Hao Zhang 和 Ion Stoica。2023 年。使用分頁注意力實現大型語言模型服務的有效記憶體管理。在第 29 屆作業系統原理研討會論文集，第 611-626 頁。

Md Tahmid Rahman Laskar、Sawsan Alqahtani、M Saiful Bari、Mizanur Rahman、Mohammad Abdullah Matin Khan、Haidar Khan、Israt Jahan、Amran Bhuiyan、Chee Wei Tan、Md Rizwan Parvez、Enamul Hoque、Shafiq Joty 和 Jimmy Huang。2024 年。大型語言模型評估的系統性調查和批判性評論：挑戰、限制和建議。在 2024 年自然語言處理經驗方法會議論文集，第 13785-13816 頁，美國佛羅里達州邁阿密。計算語言學協會。

Yanda Li、Chi Zhang、Wanqi Yang、Bin Fu、Pei Cheng、Xin Chen、Ling Chen 和 Yunchao Wei。2024 年。Appagent v2：用於靈活行動互動的進階代理。arXiv 預印本 arXiv:2408.11824。

Chin-Yew Lin. 2004. ROUGE: 一個用於自動評估摘要的套件。收錄於《Text Summarization Branches Out》，第 74-81 頁，西班牙巴塞隆納。計算語言學學會。

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu 和 Chenguang Zhu. 2023. G-eval: 使用 GPT-4 進行 NLG 評估，具有更好的人類對齊。收錄於《Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing》，第 2511-2522 頁，新加坡。計算語言學學會。

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella 和 Anandhavelu Natarajan. 2024. 簡報不總是線性的！GNN 結合 LLM 實現文件到簡報的轉換與歸因。arXiv 預印本 arXiv:2405.13095。

Ishani Mondal, S Shwetha, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay 和 Jordan Boyd-Graber. 2024. 人類製作、人類使用的簡報：利用 LLMs 從文件中生成具備人物設定的投影片。收錄於《Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)》，第 2664-2684 頁。

Athar Sefid、Prasenjit Mitra 和 Lee Giles。2021 年。Slidegen：一種用於學術文件的抽象式分節投影片產生器。收錄於《第 21 屆 ACM 文件工程研討會論文集》，第 1-4 頁。

Edward Sun、Yufang Hou、Dakuo Wang、Yunfeng Zhang 和 Nancy XR Wang。2021 年。D2s：透過基於查詢的文字摘要進行文件到投影片的生成。arXiv 預印本 arXiv:2105.03664。

Hao Tang、Darren Key 和 Kevin Ellis。2025 年。Worldcoder，一個基於模型的 LLM 代理：透過撰寫程式碼並與環境互動來建構世界模型。《神經資訊處理系統進展》，第 37 卷，第 70148-70212 頁。

VikParuchuri。2023 年。marker。
Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better 11 m agents. arXiv preprint arXiv:2402.01030.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. Preprint, arXiv:2006.03677.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22227-22238.

An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei 等人。2024。Qwen2.5 技術報告。arXiv 預印本 arXiv:2412.15115。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric Xing 等人。2023。使用 MT-Bench 和 Chatbot Arena 評估 LLM 作為評審。神經資訊處理系統進展，36:46595-46623。

A 資料預處理

為了維持合理的成本，我們選擇了頁數介於 12 到 64 頁的簡報，以及文字長度介於 2,048 到 20,480 個字元的檔案。我們使用 VikParuchuri (2023) 從原始檔案中提取了文字和視覺內容。提取的文字隨後被整理成章節。對於視覺內容，我們生成了圖片說明，以透過文字描述協助選擇相關圖片。為了最大程度地減少冗餘，如果圖片嵌入的餘弦相似度分數超過 0.85，我們就會識別並移除重複的圖片。對於投影片層級的去重複，如果單張投影片的文字嵌入與前一張投影片的餘弦相似度分數超過 0.8，我們就會移除該投影片，如 Fu 等人 (2022) 所建議。

B PPTEval 的詳細資訊

我們透過上海的群眾外包平台招募了四名研究生，評估了總共 250 份簡報：其中 50 份是從 Zenodo 隨機選取，代表真實世界的簡報；另外兩組各 100 份簡報則分別由基準方法和我們的方法生成。根據 PPTEval 提出的評估框架，我們使用附錄 F 中詳述的評分標準，從三個維度進行評估。評估人員會收到轉換後的投影片影像，各自評分後再討論結果，以達成最終分數的共識。

此外，我們使用 Fleiss' Kappa 測量評分者間的一致性，三個維度的平均分數為 0.59（內容、設計和連貫性分別為 $0.61, 0.61, 0.54$），這表示評估者之間的一致性令人滿意（Kwan et al., 2024）。代表性的評分範例顯示在圖 8 中。

我們提供了詳細的說明如下：
內容：內容維度評估投影片上呈現的資訊，著重於文字和圖像。我們從三個角度評估內容品質：資訊量、文字內容的清晰度和品質，以及視覺內容提供的支援。高品質的文字內容特點是清晰、有影響力且傳達適量資訊的文字。此外，圖像應補充和強化文字內容，使資訊

更容易理解和引人入勝。為了評估內容品質，我們在投影片圖像上使用 MLLM，因為投影片無法以純文字格式輕鬆理解。

設計：良好的設計不僅能吸引注意力，還能提升內容傳達效果。我們從三個方面評估設計維度：配色方案、視覺元素和整體設計。具體來說，投影片的配色方案應具有清晰的對比度，以突顯內容，同時保持和諧。視覺元素（例如幾何圖形）的使用可以使投影片設計更具表現力。最後，良好的設計應遵循基本設計原則，例如避免元素重疊，並確保設計不會干擾內容傳達。

連貫性：連貫性對於維持觀眾在簡報中的參與度至關重要。我們根據邏輯結構和提供的上下文資訊來評估連貫性。當模型建構出引人入勝的故事情節，並輔以豐富的上下文資訊，使觀眾能夠無縫地理解內容時，就能實現有效的連貫性。我們透過分析從簡報中提取的邏輯結構和上下文資訊來評估連貫性。

C PPTAGENT 的詳細表現

我們在表 8 中呈現了 Qwen2.5LM+Qwen2-VLVM 在各個領域的詳細效能分析。此外，表 7 和表 6 顯示了成功率加權的效能，其中失敗的生成會獲得 0 分的 PPTEVAL 分數，這表明較低的成功率會顯著影響方法的整體有效性。

如表 6 所示，GPT-4o 在各種評估指標上始終表現出色，突顯了其先進的能力。儘管 Qwen2-VL 由於多模態後訓練的權衡而在語言能力方面表現出局限性，但 GPT-4o 在處理語言任務方面保持著明顯的優勢。然而，Qwen2.5 的引入成功地彌補了這些語言缺陷，使其性能與 GPT-4o 相當，並實現了最佳性能。這突顯了開源 LLMs 作為具有競爭力且能力強大的簡報代理的巨大潛力。

D 投影片分群

我們在演算法 1 中提出了用於版面分析的分層聚類演算法，其中投影片使用相似度閾值 θ 0.65 分組為叢集。為了專注於版面模式並最大程度地減少特定內容的干擾，我們透過將文字內容替換為佔位符字元（「a」）並將圖像元素替換為純色背景來預處理投影片。然後，我們使用餘弦相似度計算相似度矩陣，該矩度基於每個投影片對之間轉換後的投影片圖像的 ViT 嵌入。圖 9 說明了所得投影片叢集的代表性範例。

E 程式碼互動

為了視覺參考，圖 10 說明了以 HTML 格式呈現的投影片，而圖 11 顯示了其 XML 表示的摘錄（前 60 行）（總共 1,006 行）。

F 提示

F. 1 簡報分析提示

用於簡報分析的提示如圖 12、13 和 14 所示。

F. 2 簡報生成提示

用於生成簡報的提示如圖 15、16 和 17 所示。

F. 3 PPTEval 的提示詞

PPTEval 中使用的提示詞如圖 18, 19, 20, 21, 22 和 23 所示。

```
Algorithm 1 Slides Clustering Algorithm
    Input: Similarity matrix of slides \(S \in \mathbb{R}^{N \times N}\),
    similarity threshold \(\theta\)
    Initialize: \(C \leftarrow \emptyset\)
    while \(\max (S) \geq \theta\) do
        \((i, j) \leftarrow \arg \max (S) \quad \triangleright\) Find the most
```

```
    similar slide pair
        if \(\exists c_{k} \in C\) such that \(\left(i \in c_{k} \vee j \in c_
    then
            \(c_{k} \leftarrow c_{k} \cup\{i, j\} \triangleright\) Merge into e
    cluster
        else
            \(c_{\text {new }} \leftarrow\{i, j\} \quad \triangleright\) Create
            \(C \leftarrow C \cup\left\{c_{\text {new }}\right\}\)
        end if
        Update \(S\) :
            \(S[:, i] \leftarrow 0, S[i,:] \leftarrow 0\)
            \(S[:, j] \leftarrow 0, S[j,:] \leftarrow 0\)
    end while
    Return: \(C\)
```

Content

Score:1
Judgement:Lack of content

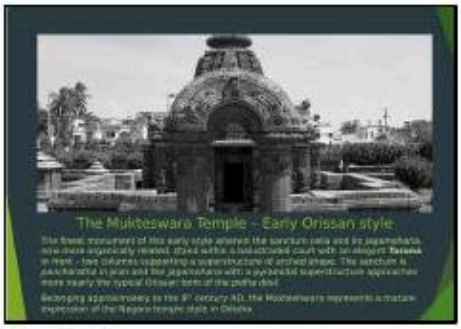
Score:3
Judgement: The content is somewhat tedious and lacks the support of images

Score:5
Judgement: The content is impactful with relevant images supports well

Design

Score:2
Judgement: Monochromatic colors without visual elements

Score:4
Judgement: Harmonious color with the use of geometric shapes; However some minor flaws diminished the overall design



Score:5
Judgement: Slide presents engaging design with consistent overall design

圖 8：圖 8：PPTEval 的評分範例。

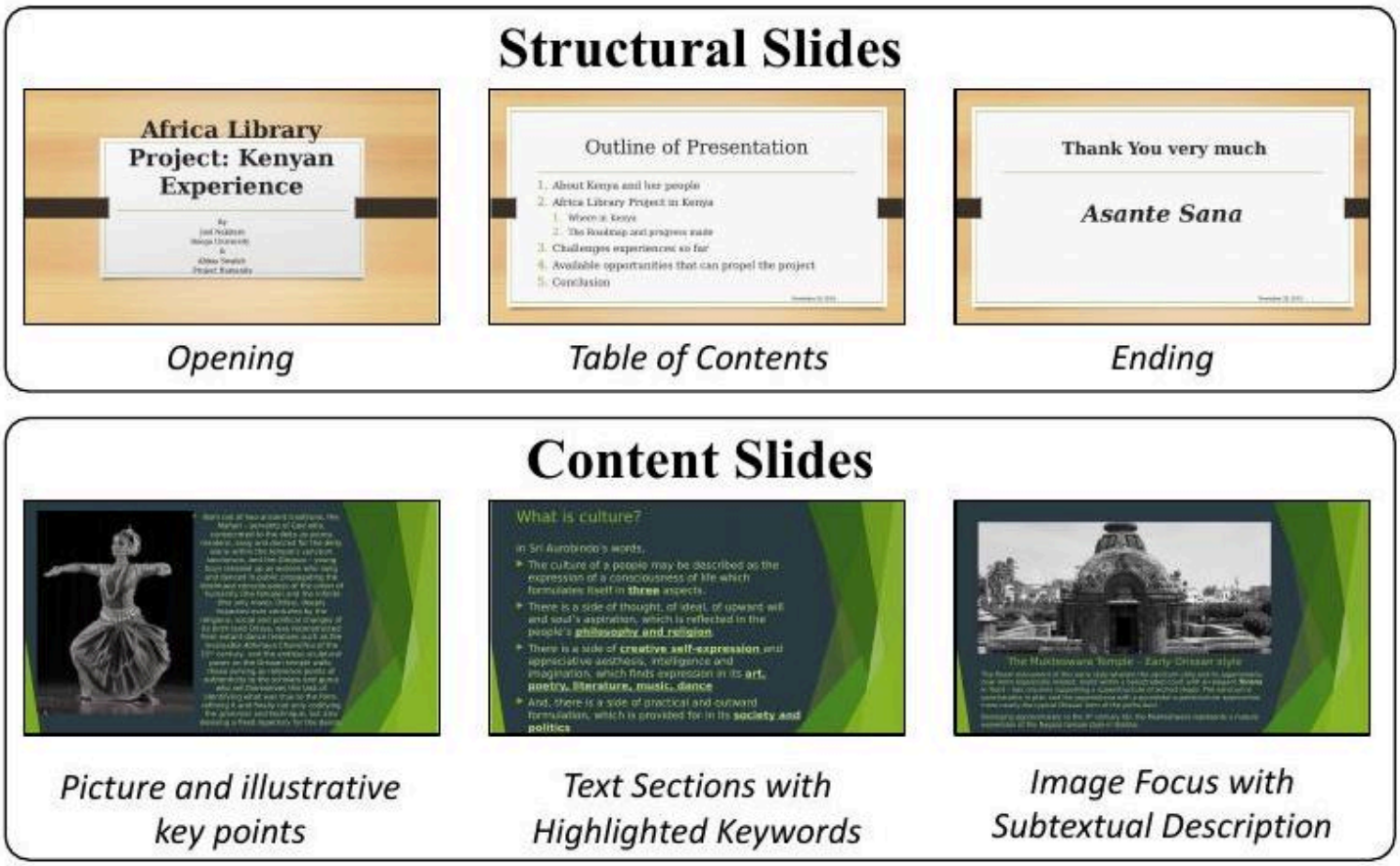


圖 9：圖 9：投影片群組範例。

```
<{DOCTYPE html>
<html>
<body style="width:720pt; height:540pt;">
<div id='0' style='font-size: 20pt; color: #595959'>
</div>
<div> checklists and infographics (e.g. IFLA, CILIP), widely disseminated in ou
<div id='2' style='font-size: 48pt; color: #595959; font-weight: bold'>
<p id='0'>Spotting fake news</p>
</div>
<div id='4' style='font-size: 16pt; color: #595959'>
<li id='0' style='font-size: 20pt; font-style: italic' bullet-type='§'>Melissa
clickbait-y, and/or satirical 'news' sources (https://goo.gl/vLsGQS)</li>
<div>
<div id='5' style='font-size: 20pt; color: #595959; font-weight: bold'>
<li id='0' bullet-type='s'>Imagine you're stuck with evaluating a piece of info
that source was listed as false or clickbait-y? </li>
</body>
</html>
```

圖 10：將投影片轉譯為 HTML 格式的範例。

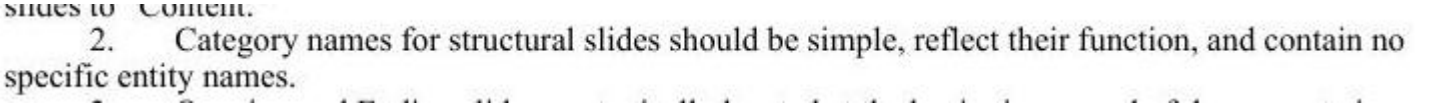
```
<p:sld xmlns:a="http://schemas.openxmlformats.org/drawingml/2006/main"
xmlns:p="http://schemas.openxmlformats.org/presentationml/2006/main"
xmlns:r="http://schemas.openxmlformats.org/officeDocument/2006/relationships">
  <p:cSld>
    <p:bg>
      <p:bgPr>
        <a:noFill/>
        <a:effectLst/>
      </p:bgPr>
    </p:bg>
    <p:spTree>
      <p:nvGrpSpPr>
        <p:cNvPr id="1" name=""/>
        <p:cNvGrpSpPr/>
        <p:nvPr/>
      </p:nvGrpSpPr>
      <p:grpSpPr/>
      <p:sp>
        <p:nvSpPr>
          <p:cNvPr id="6" name="TextBox 1"/>
          <p:cNvSpPr txBox="1">
            <a:spLocks noChangeArrowheads="1"/>
          </p:cNvSpPr>
          <p:nvPr/>
        </p:nvSpPr>
        <p:spPr bwMode="auto">
          <a:xfrm>
            <a:off x="308201" y="1422404"/>
            <a:ext cx="8367966" cy="1015663"/>
          </a:xfrm>
          <a:prstGeom prst="rect">
            <a:avLst/>
          </a:prstGeom>
          <a:noFill/>
          <a:ln>
            <a:noFill/>
          </a:ln>
          <a:extLst>
            <a:ext uri="{909E8E84-426E-40DD-AFC4-6F175D3DCCD1}">
              <a14:hiddenFill xmlns:a14="http://schemas.microsoft.com/office/drawing/2010/main">
                <a:solidFill>
                  <a:srgbClr val="FFFFFF"/>
                </a:solidFill>
              </a14:hiddenFill>
            </a:ext>
            <a:ext uri="{91240B29-F687-4F45-9708-019B960494DF}">
              <a14:hiddenLine xmlns:a14="http://schemas.microsoft.com/office/drawing/2010/main" w="9525">
                <a:solidFill>
                  <a:srgbClr val="000000"/>
                </a:solidFill>
                <a:miter lim="800000"/>
                <a:headEnd/>
                <a:tailEnd/>
              </a14:hiddenLine>
            </a:ext>
          </a:extLst>
        </p:spPr>
        <p:txBody>
          <a:bodyPr wrap="square">
            <a:spAutoFit/>
          </a:bodyPr>
```

圖 10：圖 11：簡報投影片的 XML 表示法的前 60 行（共 1,006 行）。

組態		現有指標				PPTEval			
語言模型	視覺模型	SR(%) ↑	PPL ↓	ROUGE-L ↑	FID ↓	內容 ↑	設計 ↑	連貫性 ↑	平均 ↑
文件呈現 (基於規則)									
GPT-4 LM	-	-	76.42	13.28	-	2.98	2.33	3.24	2.85
Qwen2.5 LM	-	-	100.4	13.09	-	2.96	2.37	3.28	2.87
KCTV (基於範本)									
GPT-4 LM	-	80.0	68.48	10.27	-	1.99	2.35	2.85	2.40
Qwen2.5 LM	-	88.0	41.41	16.76	-	2.24	2.59	2.95	2.59
PPTAGENT (我們)									
GPT-4 LM	GPT-4% VM	97.8	721.54	10.17	7.48	3.17	3.16	4.20	3.54
Qwen2-VL LM	Qwen2-VL VM	43.0	265.08	13.03	7.32	1.34	1.43	1.75	1.50
Qwen2.5 LM	Qwen2-VL VM	95.0	496.62	14.25	6.20	3.11	3.10	4.25	3.48

表 6：簡報生成方法的加權效能比較，包括 DocPres、KCTV 和我們提出的 PPTAgent。結果使用成功率 (SR)、困惑度 (PPL)、Rouge-L、Fréchet Inception Distance (FID) 和 SR 加權的 PPTEval 進行評估。

```
System Message:
You are an expert presentation analyst specializing in categorizing PowerPoint
identifying structural slides (such as Opening, Transitions, and Ending slides)
Prompt:
Objective:Analyze a set of slides provided in plain text format. Your task is t
(such as Opening and Ending) based on their content and categorize all other sl
Instructions:
slides to "Categorize structural slides in the presentation (such as Opening, E
slides to "Content."
```


slides to "Content."

2. Category names for structural slides should be simple, reflect their function, and contain no specific entity names.

```
specific entity names.
may 3. Opening and Ending slides are typically located at the beginning or end
may consist of only one slide.
4. Other transition categories must contain multiple slides with partially iden
Output format requirements:
Use the Functional key to group all categorized structural slides, with categor
    the slide's function (e.g., "Opening," "Ending") and do not describe any sp
Use the Content key to list all slides that do not fall into structural categor
```

```
Example output:
    json
    functional":
    "opening": [1],
    "table of contents": [2, 5],
    "section header": [3,6],
    "ending": [10]
    "content": [4, 7, 8, 9]
}
Ensure that all slides are included in the categorization, with their corresponding
output.
Input: {{slides}}
Output:
```

圖 12：用於聚類結構化投影片的提示說明。

系統訊息： 您是一位樂於助人的助理

提示：

分析所提供投影片影像中的內容佈局和媒體類型。
您的目標是建立一個簡潔、描述性的標題，純粹捕捉內容元素的呈現模式和結構排列。
要求：
專注於「如何」
已去中心化並呈現，而非內容是
等)
避免：
任何提及特定主題或科目的內容
商業或產業專用術語
實際內容描述

您不能使用以下版面名稱：
{{ existed_layoutnames }}
範例輸出：
帶有中央圖片的階層式項目符號
透過時間軸呈現演進過程
使用結構化表格顯示分析結果
成長概覽圖表
圖片與重點說明
輸出：提供單行版面配置模式標題。

設定	SR(%)	內容	設計	連貫性	平均
PPTAGENT	**95.0**	3.11	3.10	4.25	3.48
無大綱	91.0	2.94	3.00	3.05	3.00
無架構	78.8	2.42	2.54	3.18	2.71
無結構	<u>92.2</u>	**3.02**	2.99	3.18	3.06
不含 CodeRender	74.6	2.43	2.49	3.26	2.73

表 7：PPTPresenter 在 Qwen2.5 LM＋ Qwen2-VL VM 配置下進行的消融分析，PPTEval 分數根據成功率加權，以展示每個組件的貢獻。

領域	SR (%)	PPL	FID	PPTEval
文化	93.0	185.3	5.00	3.70
教育	94.0	249.0	7.90	3.69
科學	96.0	500.6	6.07	3.56
社會	95.0	396.8	5.32	3.59
科技	97.0	238.7	6.72	3.74

表 8：在不同領域中，Qwen2-VL LM＋ Qwen2-VL VM 配置下的評估結果，使用成功率 (SR)、PPL、FID 和三個評估維度上的平均 PPTEval 分數。

圖 13：用於推斷版面配置模式的提示說明。

系統訊息： \section*{**您是 hefpiu 助理**}

提示：
請分析投影片元素並以 JSON 格式建立結構化範本綱要。該綱要
應：

識別構成投影片的關鍵內容元素（包括文字和圖片）
對於每個元素，請指定：
“description”: 清楚說明元素用途，請勿提及任何細節

「類型」:「文字」或「圖片」是根據元素的標籤判斷的：「圖片」指定給 <img> 標籤 - ”data”:

對於文字元素：實際的文字內容為字串或段落層級的陣列（<p> 或 ），合併行內文字片段（）
對於圖片元素：請使用 `<img>` 標籤的 `alt` 屬性作為圖片的資料

範例格式：

「element_name」：{
「description」:「此元素的用途」，# 請勿提及任何細節，僅說明用途
「type」:「text」或「image」，
"data": "實際文字" 或 "<類型>:<50字描述>" # 詳細說明在此，不可為空或 null 或["文字 1", "文字 2"] # 多個文字元素或 [“logo:...”, “logo:...”] # 多個圖片元素
}
輸入：
請提供一個可用作範本的綱要，以建立類似的投影片。

圖 14：用於提取投影片綱要的提示說明。

系統訊息：

您是一位專業的簡報設計師，負責建立結構化的 PowerPoint 大綱。每個投影片大綱都應包含投影片標題、從提供的選項中選擇的合適版面配置，以及簡潔的說明性註解。您的目標是確保大綱符合指定的投影片數量，並且僅使用提供的版面配置。最終交付的成果應格式化為 JSON 物件。請確保大綱中不使用除所提供版面配置以外的任何版面配置。

提示： 步驟。

理解 JSON 內容：

仔細分析所提供的 JSON 輸入。
識別主要區塊和子區塊。
{{json_content }}
2. 產生大綱：

確保投影片數量符合指定要求。
保持投影片之間的邏輯流暢，並確保投影片順序能增進理解。透過功能性版面配置，確保各部分之間的轉場流暢。仔細分析所提供版面配置中指定的內容和媒體類型。

針對每張投影片，提供：

一個能清楚表達內容的投影片標題。
從提供的版面配置中選擇一個符合投影片功能的版面。
投影片說明，其中應包含簡潔明瞭的重點描述。
請以 JSON 格式提供您的輸出。
範例輸出：
XX 的開幕式：{
「佈局」：「佈局1(媒體類型)」，
「子區塊鍵」：[]，
“description”: “...”
“description”:
},
“layout”: “layout2(media
layout”: “layout2(media_type)”, # 從給定的版面配置中選擇（功能性或內容性）
subsection keys": [“子章節 1.1 標題”, “子章節 1.2 標題”],
“description”:
}
輸入：
投影片數量：{{ num_slides }}
圖片資訊：
{{ image_information }}
您只能使用以下版面配置
內容版面配置：
{{ layouts }}
功能性佈局：
{{functional_keys }}

輸出：

圖 15：用於生成大綱的提示說明。

系統訊息：

您是簡報內容的編輯代理人。您將參考文字和可用圖片轉換為符合架構的結構化投影片內容。您擅長遵循架構規則，例如內容長度，並確保所有內容都嚴格來自所提供的參考資料。您絕不會生成新內容或使用未明確提供的圖片。

提示：

根據所提供的架構生成投影片內容。

每個綱要元素都指定其用途及其預設數量。

需求：

內容生成規則：

遵循元素的預設數量，必要時進行調整
所有產生的內容都必須基於參考文字或圖片資訊。
- Generated text should use concise and impactful presentation style
- For image elements, data should be the image path # eg: "images/lo

對於圖片元素，資料應為圖片路徑，例如：「images/logo.png」。

圖片類型應是圖片選擇的關鍵因素，如果沒有提供相關圖片（相似類型或用途），請留空。

2. 核心元素：

必須從參考文字中提取必要內容（例如，投影片標題、主要內容），並保持語義一致性
在文字中討論）

支援元素（例如，簡報者、標誌圖片）：

僅在參考文字或圖片資訊中存在相關內容時才生成

為每個元素產生內容，並以以下格式輸出：
「元素 1」：{
「資料」：[「文字 1」
資料」：[「文字 1」、「文字 2」] 用於文字元素
或 ["/path/to/image" "
["/path/to/image", "..."] 用於圖片元素
} ”
輸入：
綱要
綱要
{{schema}}

簡報大綱：
{{outline}}
簡報中繼資料：
{{metadata}}

參考文字： {{text}}}

可用圖片：
{{images_info }}
{{images_info}}
輸出：生成內容中的鍵應與架構中的鍵相同

圖 16：用於產生投影片內容的提示說明。

系統訊息：

您是一位專精於投影片內容操作的程式碼生成代理程式。您能精確地將內容編輯指令轉換為 API 呼叫，方法是遵循 HTML 結構、區分標籤並維持

關係以確保精確的元素定位

提示：

根據提供的指令產生 API 呼叫序列，確保符合指定規則並精確執行。
You must determine the parent-child relationships of elements based on indentation and ensure that all and <img> elements are processed, leaving no unhandled content.
元素皆已處理，沒有未處理的內容。

每個指令都遵循以下格式：(element_class, type, quantity_change: int, old_data, new_data)。
步驟

1. **數量調整：**

quantity_change 規則：
如果 quantity_change $= 0$ ，
內容 delspan 操作。僅替換
如果 quantity_change 為 > 0 ，請使用 clone_paragraph 新增對應數量的段落：
複製時，請優先選取已具特殊樣式的相同 element_class 段落
（例如，粗體、顏色），如果有的話。
新複製段落的段落 ID 應為父元素目前最大段落 ID 加上 $\overline{1}$ ，同時保留複製段落內的 span ID 不變。
如果 quantity_change 為 < 0 ，請使用 del_span 或 del_image 來減少對應的元素數量。請務必先從段落末端移除 span 元素。
每個指令的 API 呼叫只能根據「數量變更」使用 clone_paragraph 或 del_span/del_image，不能兩者都用。
文字內容：使用 replace span 將新內容依序分發到段落中的一個或多個
元素。為強調的內容選擇適當的標籤（例如粗體、特殊顏色、較大字體）。

- 圖片內容：使用 replace_image **來替換圖片資源。**

輸出格式：
相關的元素類別。API 呼叫群組，解釋原始指令的意圖以及 - 用於複製操作

。計算段落。新建立段落的 ID。

可用的 API
{{api_docs}}
輸入範例：
請僅輸出 API 呼叫序列，每行一個呼叫，以「‘python 和 ’’」包裝，並附上對應指令的註解。

圖 17：用於生成編輯動作的提示說明。

系統訊息： \section*{**您是 AI 助理**}

提示：
請根據以下三個維度描述輸入投影片：
無論傳達的資訊是否滑動
沒有顏色或圖片，傳達的資訊過於冗長或過少，導致大片空白
2. 內容清晰度與語言品質

檢查文字內容是否有文法錯誤或語意不清之處。
3. 圖片與相關性
評估視覺輔助工具（例如圖片或圖示）的使用、其存在性，以及它們與投影片主題和內容的相關程度。
提供客觀且簡潔的描述，不帶任何評論，僅專注於上述維度。

圖 18：用於描述 PPTEval 中內容的提示說明。

系統訊息：

您是一位得力助手。

提示：

請根據以下三個維度描述輸入投影片：
描述是否有任何樣式維度
去飽和。
雜訊。 2
分析投影片中顏色的使用，識別所使用的顏色，並判斷設計是單色（黑白）還是彩色（灰色也算）。 3. 視覺元素的使用

描述投影片是否包含輔助視覺元素，例如圖示、背景、圖片或幾何圖形（矩形、圓形等）。

請根據上述概述的維度，提供客觀且簡潔的描述，不帶任何評論。

圖 19：PPTEval 中用於描述風格的提示範例。

系統訊息：

您是一位專業的簡報內容擷取者，負責分析和總結簡報的關鍵元素和中繼資料。您的任務是擷取並提供以下資訊：

提示：

評分標準（五點量表）：

投影片說明：簡要概述每張投影片的內容和重點。段落段落，不包含在其他段落中），例如作者、演講者、日期以及其他在開場和結尾投影片中直接說明的細節。

```
Example Output:
    "slide_1": "This slide introduces the xx, xx.",
    "slide_2": "...",
    "background": \{
        "speaker": "speaker x",
        "date": "date x"
        date": date x"
\} \({ }^{\text {\} }}\)
```

輸入：
{{p}
輸出：

圖 20：PPTEval 中用於提取內容的提示說明。

系統訊息：

您是一位公正的簡報分析評審，負責評估投影片內容的品質。請仔細審閱所提供的投影片影像，評估其內容，並以分數形式提供您的判斷。每個分數等級都要求所有評估標準都符合

程度。

提示：
評分標準（五點量表）：

1 分（差）：

投影片上的文字有嚴重的文法錯誤或結構不佳，導致難以理解。

2 分 (低於平均)：

投影片缺乏明確的重點，文字表達生硬，整體組織鬆散，難以吸引觀眾。

3 分 (平均)：

投影片內容清晰完整，但缺乏視覺輔助，導致整體吸引力不足。
4 分 (佳):
投影片內容清晰且發展良好，但圖片與主題的相關性較弱，限制了簡報的有效性。

5 分 (優):

和文字有效地互補，成功傳達資訊。

範例輸出：

「原因」:「xx」,「分數」： 「分數」: 整數

輸入：{{descr}}
讓我們一步一步思考並提供您的判斷。
圖 21：用於評估 PPTEval 中內容的提示說明。

系統訊息：

您是一位公正的簡報分析評審，負責評估投影片的視覺吸引力。請仔細審閱所提供的投影片描述，僅評估其美學，並提供符合標準分數的評估標準。每個分數等級都要求所有

提示：

評分標準（五點量表）：
1 分（差）：
投影片樣式之間存在衝突，導致內容難以閱讀。
2 分（尚可）：
投影片使用單調的顏色（黑白），確保了可讀性，但缺乏視覺吸引力。3 分（普通）：
投影片採用基本的配色方案；然而，它缺乏輔助的視覺元素，例如圖示、背景、圖片或幾何圖形（如矩形），使其看起來很樸素。

4 分（良好）：

投影片使用和諧的配色方案，並包含一些視覺元素（如圖示、背景、圖片或幾何圖形）；然而，整體設計可能存在一些小瑕疵。

5 分 (極佳)：

投影片的風格和諧且引人入勝，輔助視覺元素（如圖片和幾何圖形）的運用提升了投影片的整體視覺吸引力。

輸出範例：

"reason": "xx", "score": **整數**

}

輸入。[{描述)]

輸入：{{描述}}
輸入：｛描述｝
讓我們一步一步思考並提供您的判斷。

圖 22：用於評估 PPTEval 中風格的提示說明。

系統訊息：

您是一位公正的簡報分析評審，負責評估簡報的連貫性。請仔細審閱所提供的簡報摘要，評估其邏輯流程和上下文資訊，每個分數等級都要求所有評估標準符合該等級的標準。
程度。

提示：

評分標準（五點量表）

1 分（差）：

術語不一致，或邏輯結構不清晰，導致觀眾難以理解。

2 分 (尚可)：

術語一致，邏輯結構大致合理，但轉場有些微問題。

3 分 (普通)：

邏輯結構健全，轉場流暢；然而，缺乏基本的背景資訊。

4分（良好）：

邏輯流程合理，並包含基本的背景資訊（例如，演講者或致謝/結論）。

5分（優秀）：

敘事結構引人入勝且組織嚴謹，包含詳細而全面的背景資訊。

範例輸出：

「原因」：「xx」

「分數」：整數

}

輸入：{{presentation}}

讓我們一步一步思考並做出判斷，只專注於上述維度並嚴格遵守標準。

圖 23：用於評估 PPTEval 中連貫性的提示範例。

這些作者貢獻相同

[1] https://zh.wikipedia.org/wiki/REPL

[2] https://docs.python.org/3/tutorial/errors. html