

人工智慧

NTU, Spring 2025, homework1 Basic MLLM Implementation

資工碩一 渠景量 R13922193

Task 1

1. Briefly describe how you implement the two models:

BLIP :

透過 BlipProcessor 和 BlipForConditionalGeneration 載入模型與處理圖像。讀取圖像資料後，利用 processor 將圖像轉換成模型輸入格式，接著使用 .generate() 方法產生 caption，最後透過 .decode() 將模型輸出轉回自然語言文字。可針對不同資料集（如 MSCOCO 或 Flickr30k）進行大規模圖片描述生成與自動化評估，方便比較模型表現。

Phi-4:

我透過 AutoProcessor 和 AutoModelForCausalLM 載入模型與處理模態資料。為了配合 Phi-4 的格式，我使用 `<|user|>...<|end|><|assistant|>` 作為 prompt 模板，將圖像與指令組合後送入模型。推理時我使用 generate() 方法產生對應文字輸出。考量到效率以及我的 GPU 記憶體有 12GB 以上，我在處理 Flickr30k 資料集時採用 **batch size = 6** 的批次推理方式，一次處理多張圖像以提升效能；但在 MSCOCO 中則採用逐張推理以確保穩定性與推理準確度。

2. Experiment table of (2 models) X (2 datasets)

| | MSCOCO-Test | | | |
|-------|-------------|---------|---------|--------|
| | BLEU | ROUGE-1 | ROUGE-2 | METEOR |
| BLIP | 0.2552 | 0.5685 | 0.3349 | 0.4207 |
| Phi-4 | 0.0457 | 0.2157 | 0.1029 | 0.3253 |
| | flickr30k | | | |
| | BLEU | ROUGE-1 | ROUGE-2 | METEOR |
| BLIP | 0.1614 | 0.4779 | 0.2526 | 0.3232 |
| Phi-4 | 0.0353 | 0.2044 | 0.0864 | 0.2991 |

3. Analysis: describe what is observed from the table and what causes the difference in metric between the two models.

BLIP 在 MSCOCO 與 Flickr30k 的所有指標 (BLEU、ROUGE、METEOR) 均明顯優於 Phi-4，顯示其在圖像描述任務上的表現更佳。這主要是因為 BLIP 是針對圖像字幕任務訓練的模型，能夠產生與標準答案更相符的精簡描述。而 Phi-4 作為多模態大型語言模型，雖具備語意理解能力 (從 METEOR 分數可見)，但缺乏針對性的微調，使得生成內容與標準答案差異較大，導致 BLEU 與 ROUGE 分數偏低。

4. Case study: qualitative analysis of interesting samples in both models.

MSCOCO-Test:

image#2



Ground Truth: ["A young girl inhales with the intent of blowing out a candle.", "A young girl is preparing to blow out her candle.", "A kid is to blow out the single candle in a bowl of birthday goodness. ", "Girl blowing out the candle on an ice-cream ", "A little girl is getting ready to blow out a candle on a small dessert."]

BLIP: a woman sitting at a table with a candle in her mouth

Phi-4: The image features a woman sitting at a dining table, eating a bowl of ice cream. She is using a spoon to enjoy her dessert. The table is set with a few other items, including a wine glass and a cup.

There are two other people in the scene, one sitting next to the woman and another person further back. A chair is also visible in the background, placed near the dining table. The atmosphere appears to be casual and relaxed, as the woman enjoys her ice cream.

分析：在這張圖片中，MSCOCO 的標註正確指出主體為一位正準備吹蠟燭的小女孩，背景呈現生日慶祝的氛圍。BLIP 模型錯誤地將蠟燭位置理解為「在嘴裡」，產生語意錯誤。相比之下，Phi-4 雖然正確識別了冰淇淋、湯匙與桌面擺設，

但忽略了小女孩與「吹蠟燭」這一關鍵動作，且將主角誤判為成年女性。整體來看，Phi-4 描述較細緻但缺乏情境核心，而 BLIP 雖精簡但錯誤率較高，兩者皆未能完整捕捉圖片中的「生日時刻」語意。

flickr30k:

image#2



Ground Truth: ["Several men in hard hats are operating a giant pulley system.", "Workers look down from up above on a piece of equipment.", "Two men working on a machine wearing hard hats.", "Four men on top of a tall structure.", "Three men on a large rig."]

BLIP: a metal tower

Phi-4: The image features a group of people working on a tall metal structure, possibly a tower or a crane. They are standing on a platform or a ladder, and there are several other people visible in the scene, some of them closer to the structure and others further away. The workers appear to be engaged in a task that requires them to be at a considerable height, and they are likely using various tools and equipment to complete their job.

分析：這張圖呈現多位工人攀爬於高空金屬結構上作業，推測為吊塔或起重機。Flickr30k 的標註正確捕捉到人數、裝備（安全帽）、與工業設備等細節。BLIP 僅回應為「a metal tower」，忽略了人物與其活動，語意過於簡略。相較之下，Phi-4 能辨識出「高空作業」、「多人」、「工具設備」等元素，描述較具上下文與結構性，但未明確指出「工人」或「安全帽」，亦未掌握具體作業內容。整體而言，Phi-4 較能呈現場景語境，而 BLIP 顯得過度簡化。

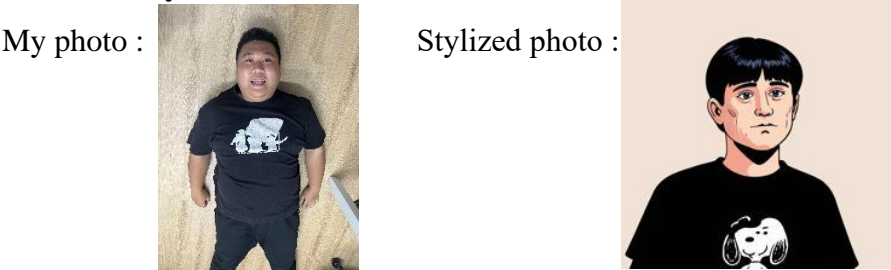
總結：根據量化評估指標與質性案例分析，BLIP 與 Phi-4 在圖像描述任務中展現出明顯的取向差異。從分數上看，BLIP 在 MSCOCO-Test 與 Flickr30k 的 **BLEU**、**ROUGE-1/2**、**METEOR** 指標皆顯著高於 Phi-4，顯示其在生成與標註描述高度一致的文本上表現穩定。然而，Case Study 中我們觀察到 BLIP 常忽略關鍵語意或出現明顯誤解（如誤將蠟燭在嘴裡），反映其對場景的理解較淺層；相對地，Phi-4 雖然在指標上得分偏低，但在場景理解與細節描寫上更具豐富性與結構性，例如能準確掌握人物數量、位置與場景氛圍。整體而言，BLIP 偏重詞彙對齊與形式準確，Phi-4 則更傾向語意完整與敘述自然。

Task 2-1


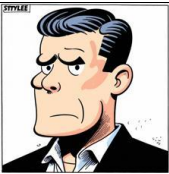










1. Briefly describe how you implement





我使用 Phi-4 輸入人像圖片與指令，要求其描述該人物的外觀特徵（Use a sentence to describe the person’s appearance, including hairstyle, face shape, and clothing.），並將其輸出作為圖像生成模型的 prompt。為了使生成 snoopy style，我在 prompt 後面加上風格描述：「Peanuts comic style」。考量到 Stable Diffusion 的 CLIP tokenizer 限制（最大 77 個 tokens），我加入了自動裁切機制，若 prompt 長度超過上限則逐字刪減描述直到符合條件。最後，將最終 prompt 輸入至 stable-diffusion-3-medium 模型生成圖像，並 resize 成 224x224 尺寸，輸出儲存。整個流程完全不訓練模型，僅透過 prompt engineering 與自動化 pipeline 完成風格轉換任務。

2-1 The style transfer on YOUR PROFILE PHOTO



2-2 5 success samples and 5 failure samples of CeleFaces and describe

| 5 success | | | 5 failure | | |
|-----------|---|---|-----------|--|---|
| Image id | Content image | Stylized image | Image id | Content image | Stylized image |
| 008 |  |  | 005 |  |  |
| 015 |  |  | 031 |  |  |
| 018 |  |  | 041 |  |  |

| | | | | | |
|-----|---|---|-----|--|---|
| 020 |  |  | 056 |  |  |
| 077 |  |  | 087 |  |  |

分析：成功樣本多半具有明確的外觀特徵，如捲髮、鬍子或穿著鮮明的衣物，描述簡潔具辨識度，便於模型準確生成對應風格。失敗樣本則常見於描述模糊、圖像光線不佳、側臉構圖或 prompt 長度過長被截斷，導致生成圖與原圖落差大，人物特徵無法有效保留。

2-3 Compare different instruction strategies

我 phi-4 的 prompt 都是相同的，但是 stable-diffusion 的 style description 不一樣的

Style 1: "a cartoon character in Peanuts comic style, flat colors, thick outlines, simple shapes, cute and minimalist" ArtFID = 22.169498443603516

Style 2: "Peanuts comic style" ArtFID = 20.654003143310547

分析：實驗結果顯示，Style 2 在風格一致性上優於 Style 1。可能原因為模型已在訓練階段對 "Peanuts comic style" 建立較清晰概念，而過度描述反而可能模糊風格重心。

Task 2-2

1. Briefly describe how you implement

這題要加入原始的人像圖片，搭配 Phi-4 生成的人物外觀描述，一起輸入到 stable-diffusion-v1-5 的 Image-to-Image 模型中，讓模型在保留原圖輪廓的同時，進行 Snoopy 風格轉換。風格提示我使用簡潔的「Peanuts comic style」，也嘗試搭配具體的風格描述，觀察效果差異。為了避免 prompt 過長超出模型限制，我也加入了自動裁切機制，以防止超過 77 個 token，如果超過會使得模型無法生成圖片。

2-1 The style transfer on YOUR PROFILE PHOTO



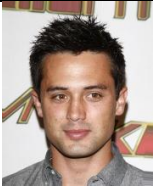




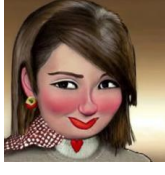




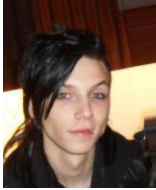







My photo :



Stylized photo :



2-2 5 success samples and 5 failure samples of CeleFaces and describe

| 5 success | | | 5 failure | | |
|-----------|---|---|-----------|--|---|
| Image id | Content image | Stylized image | Image id | Content image | Stylized image |
| 094 |  |  | 007 |  |  |
| 034 |  |  | 009 |  |  |
| 018 |  |  | 011 |  |  |
| 056 |  |  | 044 |  |  |
| 078 |  |  | 068 |  |  |

分析：成功樣本通常具備明確特徵（如髮型、膚色、配件），而失敗樣本則可能因 prompt 不夠具體、圖片過於模糊或背景干擾，導致風格偏移或內容丟失。相較於 Task 2-1，Task 2-2 雖能保留原圖輪廓，但更容易受到原圖構圖與風格描述衝突的影響。

2-3 Compare different instruction strategies

下圖為 094，使用 "a cartoon character in Peanuts comic style, flat colors, thick outlines, simple shapes, cute and minimalist" 的 prompt 產生出來的圖片，而上一題的成功例子的 prompt，則為 "Peanuts comic style"。從結果可以看出，簡潔的「Peanuts comic style」所生成的圖像在風格一致性與人物特徵保留上表現較好；而較長的描述雖然詳細，但容易造成語意模糊或模型誤解，導致五官比例失真。簡潔明確的指令在此次作業中更有效。

