HW1

Student id : r13922154 章程睿

Task 1: Image Captioning Evaluation

1. Briefly describe how you implement the two models:

❖ BLIP Model Implementation :

1. Model Loading: load the pre-trained BLIP model (BlipForConditionalGeneration) and its processor (BlipProcessor) from Salesforce's "blip-image-captioning-base" checkpoint.
2. Dataset Handling:load either MSCOCO or Flickr30k dataset from Hugging Face,extract the image and its reference captions.
3. Caption Generation Process:Each image is processed through the BLIP processor to create inputs for the model.

❖ Phi-4 Model Implementation :

1. Model Loading: load Microsoft's Phi-4 multimodal model with optimizations:Flash Attention 2 is used when available on GPU,The model uses the appropriate device mapping and torch data types.
2. Dataset Handling : load either MSCOCO or Flickr30k dataset,process the dataset in batches rather than one image at a time.
3. Caption Generation Process : define a specific prompt template: <|user|><|image_1|>Generate a detailed caption for this image.<|end|><|assistant|>.

Batches of images are processed through the processor together. Generation parameters include max_new_tokens=20 and model-specific generation config.

❖ Common Evaluation Implementation :

BLEU-1(unigram precision),BLEU-4(combined n-gram precision with n = 1,2,3,4).

METEOR Score : Uses NLTK's implementation with reference alignment

ROUGE-1(unigram overlap) ROUGE-2(bigram overlap)

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

Blip2					
	bleu-1	bleu-4	meteor	rouge-1	rouge-2
mscoco	0.6017	0.2422	0.4242	0.3638	0.1301
flickr30k	0.5250	0.1498	0.3218	0.2497	0.0603

(1000)					
flickr30k(full)	0.5326	0.1564	0.3288	0.2500	0.0651

註:flickr30k(1000) 為由flickr30k["test"]裡面的"split"這個column中所歸類爲"test"的所有資料.

phi4							
	bleu-1	bleu-2	bleu-3	bleu-4	meteor	rouge-1	rouge-2
mscoco	0.7799	0.5939	0.4311	0.2999	0.5703	0.6016	0.3564
flickr30k (full)	0.7560	0.5619	0.3999	0.2768	0.5277	0.5017	0.2632

3. **Analysis : describe what is observed from the table and what causes the difference in metric between the two models.**
   Ans : The experimental results clearly demonstrate that Phi-4 consistently outperforms BLIP2 across all evaluation metrics on both datasets. Both models perform better on MSCOCO than on Flickr30k.
   Causes of metric Differences Between the Two Models:
   - Phi-4 is a more recent model likely featuring more advanced architecture design and a larger parameter count.
   - Phi-4 probably employs more sophisticated attention mechanisms that better capture image-text relationships.
   - Phi-4 was likely pre-trained on a larger, more diverse dataset of image-text pairs
   - Phi-4 has probably undergone instruction tuning, making it more adept at understanding and executing caption generation tasks
4. Case study : qualitative analysis of interesting samples in both models

Task 2-1 : MLLM Image Style Transfer (Text-to-image) :
Task 2-2 : MLLM Image Style Transfer (Image-to-Image) :

1.Briefly describe how you implement task 2-1 & 2-2:

In this implementation, I adopted a "two-stage generation" instruction strategy:

1. **Description Generation Stage**: Using the Phi-4 multimodal model to interpret facial images and generate targeted textual descriptions.
2. **Style Transfer Stage**: Combining the textual descriptions with Snoopy-style prompts to generate stylized images through Stable Diffusion.

Key Instruction Design Details

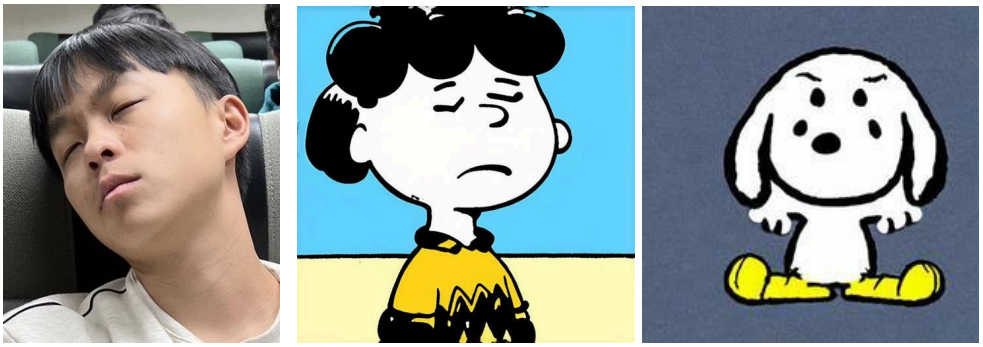
Description Generation Prompt : "Describe this person's appearance in detail. Focus on facial features, hairstyle, expression, and any notable characteristics."

Style Transfer Prompt :Used a fixed style prefix combined with individual feature descriptions: "In the style of Snoopy from Peanuts, cartoon, cute, rounded character designs, simple black line art, bold colors, minimalist backgrounds..."

This two-stage strategy allows the model to first understand facial features, then perform targeted style transfer, ensuring the generated Snoopy-style images still retain the unique characteristics of the original faces.
The difference between the two tasks is whether they provide an original input to the diffusion model.

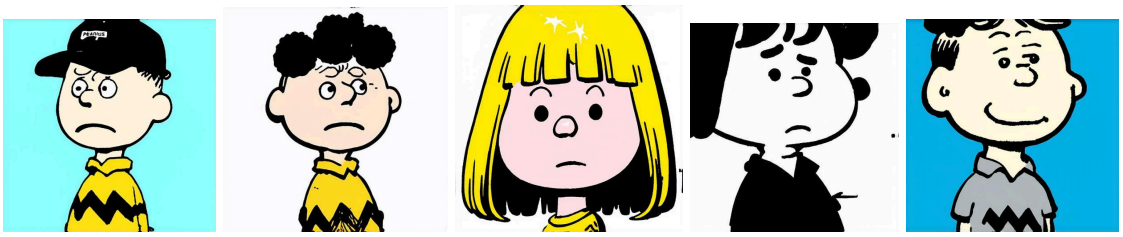
2.Visualization

- The style transfer on your profile photo:
  input_image : 2-1 : 2-2 :

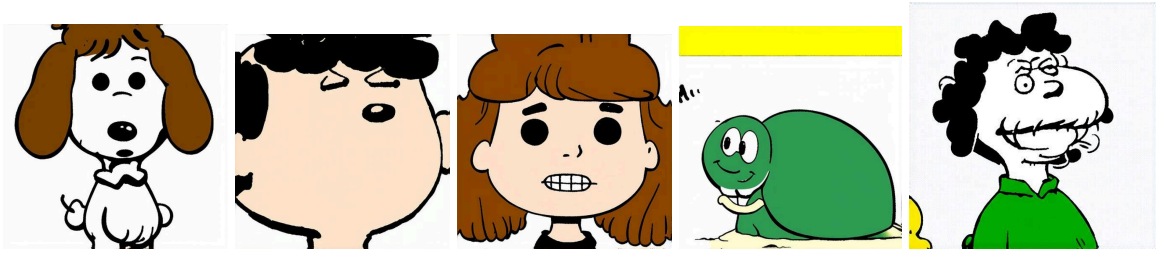


- 5 success samples and 5 failure samples

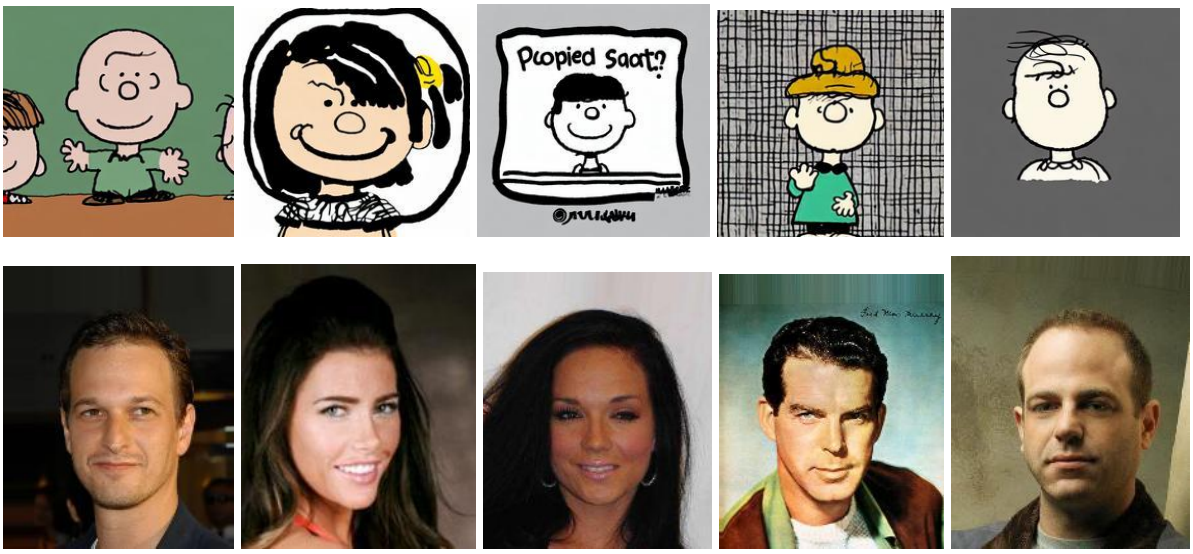
2-1 success part : "Successfully identified and converted the spirit, facial direction, facial expressions, etc. of the original image. Additionally: identified hat and blonde hair surprised me."



2-1 failure part : Unable to successfully identify facial expressions, and generated unnatural facial appearances, even turning into a dog and a turtle.



2-2 success part : The general facial features and spirit were captured, but the facial expressions are more blurred, and it cannot capture precise features like in 2-1



2-2 failure part :Unnatural facial expression,Unable to capture features or accessories, etc.

2-3.Compare different instruction strategies

Prompt Design Strategy Comparison

The two implementations use notably different prompt strategies when instructing the AI models:

TASK 2-1:

Description prompt design : "Describe this person in detail, focusing on facial features, hair style, and expression. Describe as if they were a character in Peanuts/Snoopy cartoon style."

Key feature: Directly incorporates the target style into the description phase

Advantage: Creates descriptions already aligned with the Peanuts aesthetic

Generation prompt design :"A cartoon character in classic Peanuts comic strip style. {description}. Clean, simple line work with minimal detail. Round heads, small bodies with simplified limbs..."

Key feature: Rich contextual details about Peanuts stylistic elements

Advantage: Guides the model with specific artistic direction

Task2-2:

Description prompt design :"Describe this person's appearance in detail. Focus on facial features, hairstyle, expression, and any notable characteristics."

Key feature: Style-agnostic description focused only on appearance

Advantage: Captures more neutral, comprehensive facial details

Generation prompt design :"In the style of Snoopy from Peanuts, cartoon, cute, rounded character designs, simple black line art, bold colors, minimalist backgrounds..."

Key feature: Uses a fixed style prefix concatenated with neutral description

Advantage: Clear separation between content and style guidance

Analysis of Prompt Strategies

The first strategy employs "style-integrated prompting" where the target style influences both stages of generation. The second strategy uses "style-separated prompting" where content description and style direction are handled independently.

Based on the observation that the first approach produces better results, we can conclude that introducing style elements earlier in the pipeline leads to more cohesive stylistic transformations.