Student ID: R12942159
Name: 邱亮茗

Artificial Intelligence HW1 Report

Task 1: Image Captioning Evaluation

- **Briefly describe how you implement the two models (5%)**

1. **Model Setup:**
   - Loaded BLIP (Salesforce/blip-image-captioning-base) and Phi-4 (microsoft/Phi-4-multimodal-instruct) using Hugging Face Transformers.
   - For the Phi-4 model prompt, give “<|user|><|image_1|>Describe the image in detail.<|end|><|assistant|>”.
   - Used torch.compile to optimize both models for faster inference.
2. **Dataset Preparation:**
   - Loaded MSCOCO-Test (5k) and Flickr30k (~30k) datasets for evaluation.
   - Processed images and captions in batches to improve efficiency.
3. **Inference Pipeline:**
   - Used evaluate_captioning_batch() to generate captions for input images.
   - Completed inference for both BLIP and Phi-4 models on the entire Flickr30k dataset.
4. **Evaluation Metrics:**
   - Computed BLEU, ROUGE-1, ROUGE-2, and METEOR scores to compare model-generated captions with ground truth.
5. **Batch Processing & Performance Optimization:**
   - Processed images in batch size = 8 to balance efficiency and memory usage.

- **Experiment table of (2 models) × (2 datasets), for example: (5%)**

	MSCOCO-Test				Filckr30k			
	BLEU	ROUGE-1	ROUGE-2	METEOR	BLEU	ROUGE-1	ROUGE-2	METEOR
BLIP	0.2052	0.5830	0.3448	0.4207	0.1432	0.4932	0.2672	0.3232
Phi-4	0.0323	0.1982	0.0931	0.3077	0.0293	0.2131	0.0899	0.3034

- **Analysis: describe what is observed from the table and what causes the different in metric between the two models. (5%)**

1. **Observations:**

BLIP outperforms Phi-4 across all metrics on MSCOCO-Test and Flickr30k, achieving significantly higher BLEU, ROUGE, and METEOR scores. For example, BLIP scores 0.2052 in BLEU on MSCOCO-Test, while Phi-4 only reaches 0.0323. The trend persists across both datasets, indicating BLIP's superior n-gram overlap and semantic accuracy. Both models perform better on MSCOCO-Test, possibly due to closer alignment with training data.

2. **Possible Causes of Performance Differences:**

BLIP’s advantage stems from its vision-language architecture and extensive pretraining on multimodal data, enabling better caption generation. Phi-4, as a primarily language-based model, lacks strong visual-text alignment, leading to lower scores. BLIP’s refined feature extraction enhances n-gram and semantic similarity, while Phi-4 struggles with precise image descriptions. However, the smaller METEOR gap suggests Phi-4 still captures some semantic meaning.

- **Case study: qualitative analysis of interesting samples in both models. (5%)**

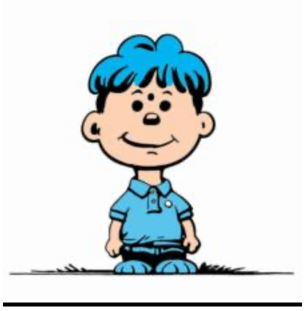
In MSCOCO-Test, BLIP generates a concise, accurate caption—"A man riding a motorcycle down a dirt road."—closely matching the reference. This precision boosts its BLEU and ROUGE scores. In contrast, Phi-4's caption is overly descriptive, adding hallucinated elements like a bridge and emotional context, leading to lower n-gram overlap. This highlights BLIP's strength in factual captioning, while Phi-4 produces more narrative-driven but less precise outputs.

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

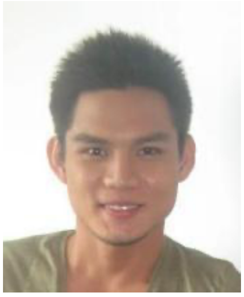
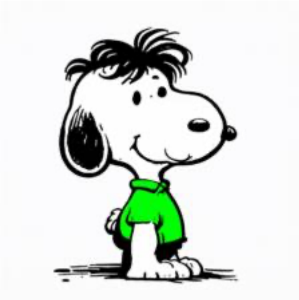
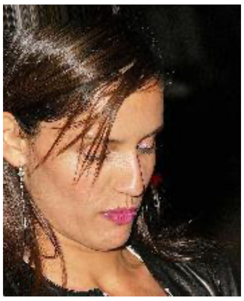
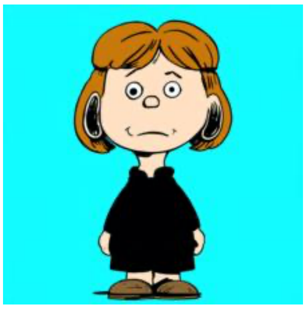
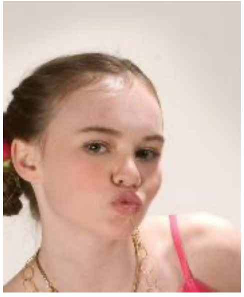
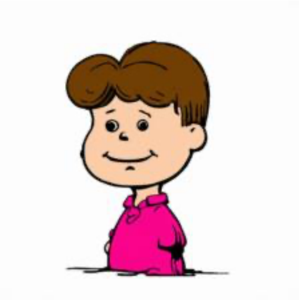
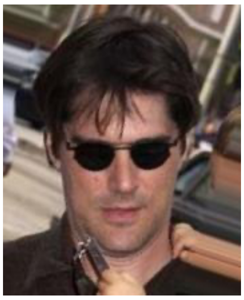
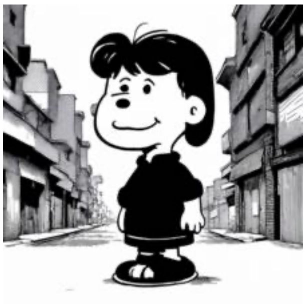
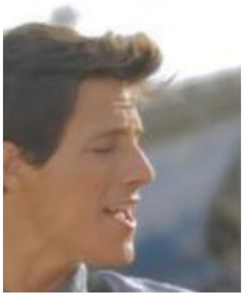
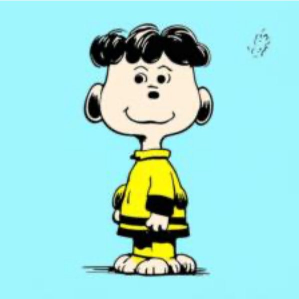
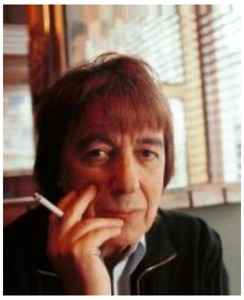
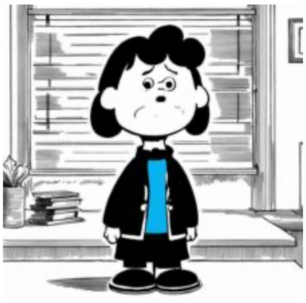
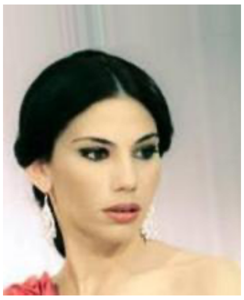
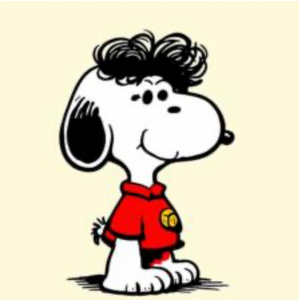
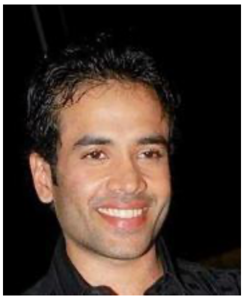
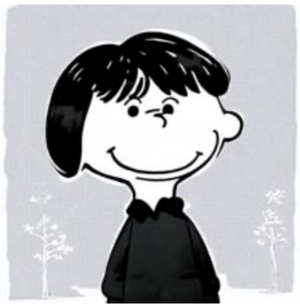
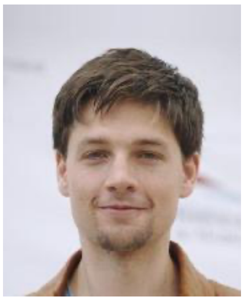
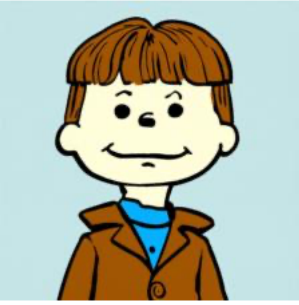
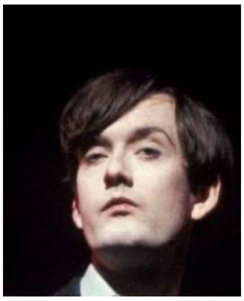
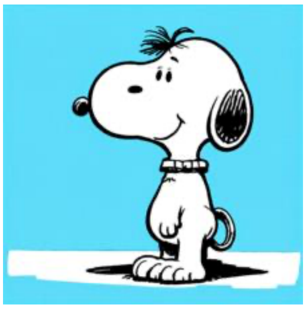
- **Briefly describe how you implement task 2-1. (5%)**

1. **Model Setup:**
   - The script loads Phi-4-multimodal-instruct from Hugging Face using AutoProcessor and AutoModelForVision2Seq.
   - This model is used to generate a text description of the input image in a Peanuts (Snoopy) cartoon style.
2. **Generate Text Prompt from Image:**
   - A given image is processed and passed to the Phi-4 model along with an instruction: "Describe this person in a simple cartoon-style suitable for Peanuts (Snoopy) characters."
   - The model generates a descriptive text prompt for use in image generation.
3. **Load Stable Diffusion 3 Medium:**
   - The script initializes the StableDiffusionPipeline for stabilityai/stable-diffusion-3-medium-diffusers.
   - The model is configured with the following parameters:
     - Negative Prompt: "blurry, distorted, low quality" to reduce unwanted artifacts.
     - Number of Inference Steps: 32 for a balance between speed and image quality.
     - Guidance Scale: 7.0, which controls the influence of the text prompt on image generation.
4. **Generate Stylized Cartoon Image:**
   - The generated text prompt is used as input to **Stable Diffusion 3 Medium**.
   - The pipeline generates an output image in the requested style.
5. **Resize and Save Image:**
   - The output image is resized to 224×224 pixels using torchvision.transforms.
   - The resized image is saved in an output_images folder.

- **The style transfer on YOUR PROFILE PHOTO. (5%)**



- **5 success samples and 5 failure samples of CeleFaces and describe. (5%)**

Original RGB image	5 Success samples	Original RGB image	5 Failure samples
 (000065.jpg)		 (000004.jpg)	
 (000086.jpg)		 (000053.jpg)	
 (000003.jpg)		 (000067.jpg)	
 (000057.jpg)		 (000081.jpg)	
 (000064.jpg)		 (000048.jpg)	

The analysis of Snoopy-style transformations highlights key factors in preserving facial features and structure. Successful cases (000065, 000086, 000003, 000057, 000064) retain recognizable features with proportionally consistent eyes, nose, and mouth, ensuring smooth, cartoonish stylization without losing identity. In contrast, failed cases (000004, 000053, 000067, 000081, 000048) show extreme distortions, misalignment, or over-simplification, reducing recognizability. Many failures involve dark backgrounds or obstructions (e.g., sunglasses, cigarettes), suggesting the model struggles with complex facial characteristics and challenging visual conditions.

- **Compare different instruction strategies. (5%)**

1. "<|user|><|image_1|>Describe this person in a simple, playful cartoon-style, like Snoopy characters. Keep the person's features but render them in a colorful, exaggerated, and minimalist cartoon style. The character should have the same hair color, expression, and outfit as in the image, but depicted in a Peanuts-like, childlike manner with a simple background. <|end|><|assistant|>"

While this instruction does mention a cartoon-style transformation and references Snoopy characters, it does not strictly enforce the use of the Peanuts artistic style. Consequently, some generated prompts only describe a general cartoon-like appearance without fully adhering to the distinct characteristics of Snoopy-style illustrations.

2. "<|user|><|image_1|>Describe this person in the exact artistic style of Peanuts comics (Snoopy-style). Ensure the description makes the character look like they belong in a Charles Schulz comic strip. The person’s features—such as hair color, expression, and outfit—should remain the same, but they must be transformed into the signature Peanuts cartoon style: simple, bold outlines, flat colors, round heads, dot eyes, and minimal shading. The background should be minimalistic, similar to classic Peanuts comic settings. <|end|><|assistant|>"

This instruction addresses the shortcomings of the previous two approaches by explicitly requiring the Snoopy-style transformation. It provides clear and specific guidance on how the character should be rendered, ensuring that the generated prompts consistently include Snoopy-style elements. As a result, the final outputs align closely with the Peanuts comic aesthetic.

<u>Task 2-2 MLLM Image Style Transfer(Image-to-image)</u>

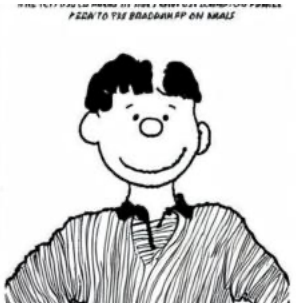
- **Briefly describe how you implement task 2-2. (5%)**

1. **Image Preprocessing:**
   - The read_images function loads 100 images from a specified directory.
   - The resize_image function ensures images are resized to 224×224 before processing.
2. **Text Prompt Generation:**
   - I utilize Phi-4 multimodal (microsoft/Phi-4-multimodal-instruct) to generate descriptive prompts in the style of Peanuts comics.
   - The generate_text_prompt function constructs an instruction to describe a person in a Peanuts/Snoopy-style and feeds it into the Phi-4 model.
   - The generate_text_prompts function loops over all images to generate 100 text prompts.

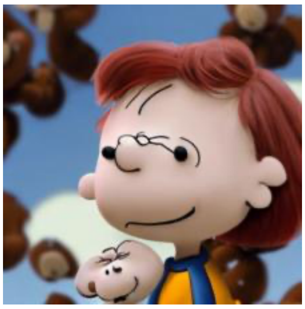
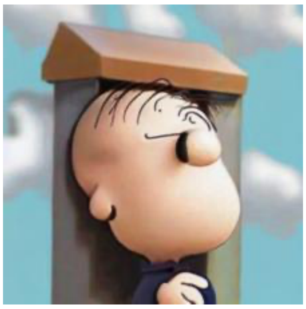
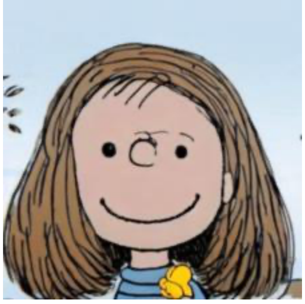
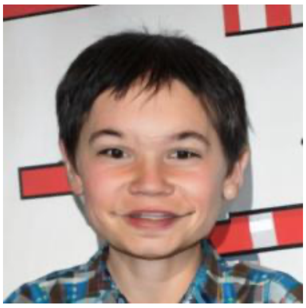
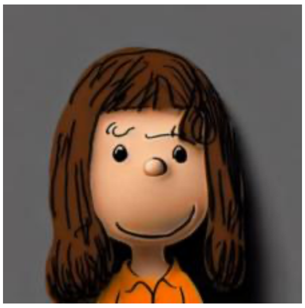
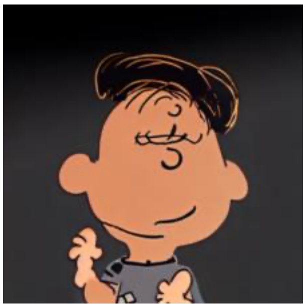
3. Stylized Image Generation (Stable Diffusion Img2Img):

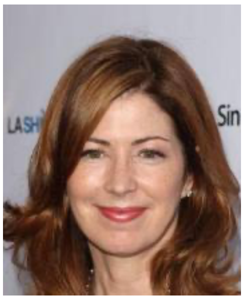
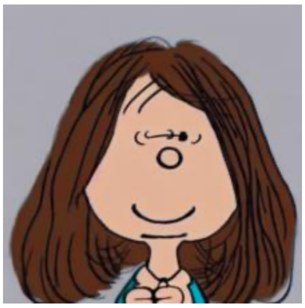
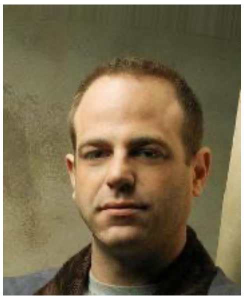
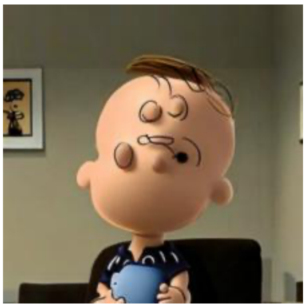
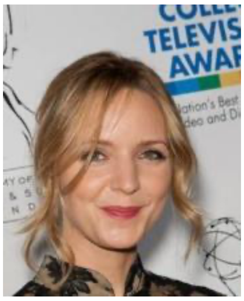
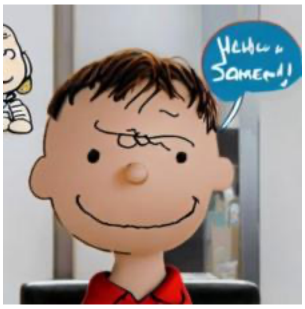
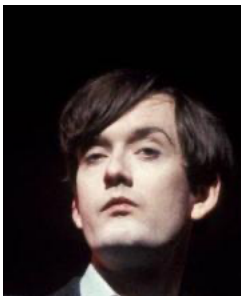
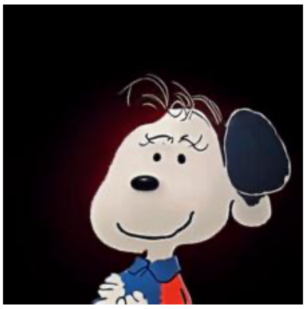
- I load Stable Diffusion v1.5 with the load_stable_diffusion function.
- The generate_stylized_images function takes an image, applies the corresponding Peanuts-style text prompt, and generates a stylized version using Image-to-Image (Img2Img) transformation.
- The transformation is controlled by parameters like strength=0.75 and guidance_scale=7.5 to balance content fidelity and artistic style.

- The style transfer on YOUR PROFILE PHOTO. (5%)



- 5 success samples and 5 failure samples of CeleFaces and describe. (5%)

Original RGB image	5 Success samples	Original RGB image	5 Failure samples
 (000002.jpg)		 (000005.jpg)	
 (000018.jpg)		 (000007.jpg)	
 (000034.jpg)		 (000012.jpg)	

 (000083.jpg)		 (000025.jpg)	
 (000092.jpg)		 (000048.jpg)	

- **Compare different instruction strategies. (5%)**

Strategy1: "Generate a prompt for a drawing in the style of Snoopy comics. The prompt should describe a simple, cartoonish scene with bold outlines and minimal shading, using lighthearted and whimsical language."

Strategy2: "Here are three examples of prompts that describe a Snoopy-style drawing: A happy beagle with big black ears sits on top of a red doghouse, looking at the stars. The style is simple, cartoonish, with clean black outlines and no shading. A small bird with tiny wings and a tuft of feathers on its head flutters near a dog, both smiling in a minimal, newspaper comic strip style. A relaxed dog, lying on his back with a dreamy expression, while a tiny yellow bird perches on his nose. The lines are hand-drawn, expressive, and playful. Now, generate a new prompt in the same style."

The strategy1 is simpler and more flexible, allowing Phi-4 to generate diverse prompts that broadly align with the Snoopy style. However, it may produce inconsistent results if the model does not fully understand Schulz's artistic characteristics. In contrast, the strategy2 provides more structured guidance, leading to more accurate and consistent outputs by demonstrating the expected format. The trade-off is reduced creativity, as Phi-4 may closely mimic the given examples rather than generating novel variations. If precision in style is the priority, strategy2 prompting is superior, whereas Direct Stylistic Description is better for broad experimentation.