# Artificial Intelligence (CSIE 5400)

R13922A22 賴澍雨

## HW-1 Basic MLLM Implementation

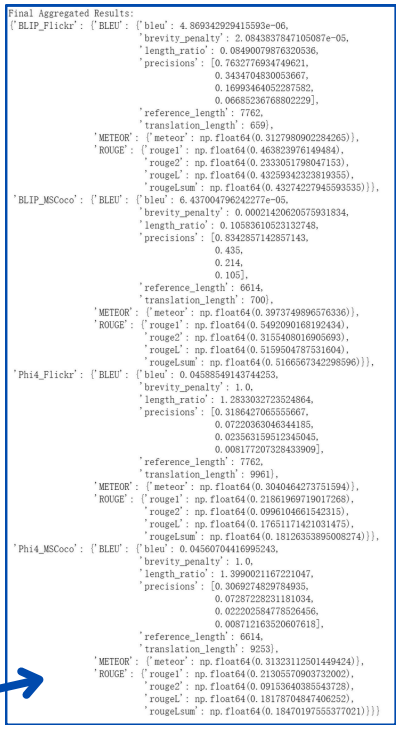
## Task 1: Image Captioning Evaluation

### 1. Briefly describe how you implement the two models (5%)

Complete run of flickr30k (~30k) (in BLIP and Phi-4)

```
[ ] # Show how much data in dataset
    print(len(mscoco_dataset))
    print(len(flickr_dataset))

    5000
    31014
```



```
Final Aggregated Results:
{'BLIP_Flickr': {'BLEU': {'bleu': 4.869342929415593e-06,
                          'brevity_penalty': 2.0843837847105087e-05,
                          'length_ratio': 0.08490079876320536,
                          'precisions': [0.7632776934749621,
                                         0.3434704830053667,
                                         0.1699346405228758Z,
                                         0.06685236768802229],
                          'reference_length': 7762,
                          'translation_length': 659},
                 'METEOR': {'meteor': np.float64(0.3127980902284265)},
                 'ROUGE': {'rouge1': np.float64(0.463823976149484),
                           'rouge2': np.float64(0.2333051798047153),
                           'rougeL': np.float64(0.4325934232319355),
                           'rougeLsum': np.float64(0.4327422794559335)}},
 'BLIP_MSCoco': {'BLEU': {'bleu': 6.437004796242277e-05,
                          'brevity_penalty': 0.00021420620575931834,
                          'length_ratio': 0.10583610523132748,
                          'precisions': [0.8342857142857143,
                                         0.435,
                                         0.214,
                                         0.105],
                          'reference_length': 6614,
                          'translation_length': 700},
                 'METEOR': {'meteor': np.float64(0.3973749896576336)},
                 'ROUGE': {'rouge1': np.float64(0.5492090168192434),
                           'rouge2': np.float64(0.3155408016905693),
                           'rougeL': np.float64(0.5195047875316O4),
                           'rougeLsum': np.float64(0.5166567342298596)}},
 'Phi4_Flickr': {'BLEU': {'bleu': 0.04588549143744253,
                          'brevity_penalty': 1.0,
                          'length_ratio': 1.2833032723524864,
                          'precisions': [0.3186427065555667,
                                         0.07220363046344185,
                                         0.023563159612345045,
                                         0.008177207328433909],
                          'reference_length': 7762,
                          'translation_length': 9961},
                 'METEOR': {'meteor': np.float64(0.3040464273751594)},
                 'ROUGE': {'rouge1': np.float64(0.21861969719017268),
                           'rouge2': np.float64(0.0996104661542315),
                           'rougeL': np.float64(0.17651171421031475),
                           'rougeLsum': np.float64(0.18126353895008274)}},
 'Phi4_MSCoco': {'BLEU': {'bleu': 0.04560704416995243,
                          'brevity_penalty': 1.0,
                          'length_ratio': 1.399002116722104Z,
                          'precisions': [0.3069274829784935,
                                         0.07287228231181034,
                                         0.022202584778526456,
                                         0.008712163520607618],
                          'reference_length': 6614,
                          'translation_length': 9253},
                 'METEOR': {'meteor': np.float64(0.31323112501449424)},
                 'ROUGE': {'rouge1': np.float64(0.21305570903732002),
                           'rouge2': np.float64(0.09153640385543728),
                           'rougeL': np.float64(0.18178704847406252),
                           'rougeLsum': np.float64(0.18470197555377021)}}}
```

1) **Environment and Library Installation**

Make sure the install the required libraries and set the Transformers version to 4.48.2 to avoid incompatibilities (especially important for Phi-4).

2) **BLIP** (Salesforce/blip-image-captioning-base)

Implementation Steps

1. Load the BLIP processor and model, and move them to GPU (if available).
2. Use the processor to convert an input image to tensors.
3. Call the models generate method to get the predicted caption.
4. Decode the output tokens back into a string.

3) **Phi-4** (microsoft/Phi-4-multimodal-instruct)

Implementation Steps

1. Load the Phi-4 AutoProcessor and AutoModelForCausalLM.
2. Construct the prompt with special tokens (e.g., <|user|>, <|assistant|>, <|image_1|>).
3. Preprocess the input image (and text prompt) with the processor.
4. Generate text with the model's generate method, then decode and post-process the output.

4) **Evaluation**

Define the get_blip_caption(image) and get_phi4_caption(image) functions:

1. Load the desired dataset of images and reference captions.
2. Iterate through each image in the dataset:
   - Generate captions using BLIP or Phi-4.
   - Compare them to ground-truth references or evaluate them using metrics (e.g., BLEU, ROUGE, METEOR).

### 2. Experiment table of (2 models) X (2 datasets), for example: (5%)

| | MSCOCO-Test | | | | flickr30k | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | ROUGE-1 | ROUGE-2 | METEOR | BLEU | ROUGE-1 | ROUGE-2 | METEOR |
| BLIP | 6.437004796242277e-05 | 0.5492090168192434 | 0.3155408016905693 | 0.3973749896576336 | 4.869342929415593e-06 | 0.463823976149484 | 0.2333051798047153 | 0.3127980902284265 |
| Phi-4 | 0.04560704416995243 | 0.21305570903732002 | 0.09153640385543728 | 0.31323112501449424 | 0.04588549143744253 | 0.21861969719017268 | 0.0996104661542315 | 0.3040464273751594 |

### 3. Analysis: describe what is observed from the table and what causes the difference in metric between the two models. (5%)

**Observations from the table**

- **BLIP has very low BLEU scores on both datasets (near zero), yet comparatively higher ROUGE and METEOR scores.**
  - This usually indicates that BLIP's captions are very short but still contain key overlapping words with the references (hence decent ROUGE/METEOR). BLEU is notorious for its brevity penalty, so if a model outputs only a few words, BLEU will drop near zero.
- **Phi-4 shows higher BLEU (especially compared to BLIP) but lower ROUGE/METEOR.**
  - Phi-4 tends to generate longer, more verbose captions, avoiding brevity penalties and thus improving BLEU. However, these longer captions might include more "off-target" or irrelevant words, reducing token-level overlap with reference captions and thus slightly hurting ROUGE/METEOR.

**Causes of the metric differences**

- **Caption Length / Brevity vs. Verbosity:**
  - BLIP's relatively short outputs suffer under BLEU's brevity penalty but still match enough key phrases to perform well in ROUGE and METEOR. Phi-4's longer outputs avoid brevity penalties for BLEU but incorporate extra words that may not match the ground truth as closely under ROUGE/METEOR.
- **Training Objectives and Model Architectures:**
  - BLIP is specialized for concise image-to-text generation; its pre-training encourages succinct, "correct" captions. Phi-4 is instruction-tuned for conversational and descriptive responses, often producing longer, more detailed text.
- **Prompting Style:**
  - Phi-4's prompt explicitly asks, potentially prompting a fuller description. BLIP can generate unconditioned or user-conditioned captions but often defaults to shorter outputs unless prompted otherwise.

In short, caption length and content alignment with the ground-truth references largely explain why BLIP scores higher on ROUGE/METEOR but almost zero on BLEU (brevity penalty), whereas Phi-4 produces lengthier captions that benefit BLEU but score lower on the other metrics.

# Task 1: Image Captioning Evaluation

R13922A22 賴澍雨

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

### Qualitative Analysis on **MSCOCO-Test** (5 Random Samples)

Sample 1

**Reference Caption**: ['A man drinking a cup of coffee next to sandwich.', 'A young man drinking a mug of coffee', 'A large sandwich sits on a plate in front of a man sipping coffee. ', 'A guy sits in a cafe sipping coffee and eating a sandwich.', 'A young man is sipping a latte at a cafe.']
**BLIP Caption**: a man sitting at a table with a cup of coffee
**Phi-4 Caption**: The image features a man sitting at a table in a restaurant, enjoying a meal. He is holding a cup of coffee in his hand and appears to be taking a sip. The table is set with a sandwich and a spoon, indicating that he is having a meal.
There are several other people in the scene, with one person sitting close to the man and others scattered around the restaurant. A car can be seen parked outside the restaurant, and a chair is placed nearby. Additionally, there is a bottle on the table, possibly containing a beverage or condiment for the meal.

Sample 2

**Reference Caption**: ['a bathroom with a bath tub a sink and a mirror', 'A bathroom with a tub on the side and a large bowl sink under a large bathroom mirror.', 'A bathroom with a sink that is elevated.', 'A tile bathroom contains a tub and large sink. ', 'A bathroom is outfitted with brown and white.']
**BLIP Caption**: a bathroom with a sink and a mirror
**Phi-4 Caption**: The image displays a bathroom with a sink, a mirror, and a bathtub. The sink is positioned under a mirror, and the bathtub is located next to it. The bathroom appears to be in a state of disrepair, with the sink and bathtub showing signs of wear and tear.

Sample 3

**Reference Caption**: ['The small bathroom has a bathtub without a shower.', 'A bathroom has a tub but no shower.', 'A bathroom with a white bath tub sitting next to a sink.', 'A tidy bathroom decorated with decor and towels.', 'A white and marble tiled bathroom that looks great.']
**BLIP Caption**: a bathroom with a tub, sink, and mirror
**Phi-4 Caption**: The image displays a bathroom with a white bathtub, a sink, and a mirror. The bathroom is well-lit and features a tiled floor. There is a towel hanging on a rack, and a chair is placed near the bathtub. The room also has a window, providing natural light to the space.

Sample 4

**Reference Caption**: ['a toddler wearing a hat sitting in the dry grass ', 'A boy wearing a hat is laying on a grass field', 'A young child in a sun hat crawling in hay at a kite fair', 'A boy plays in the grass with kites flying in the sky. ', 'A young boy in a blue hat crawling in a field.']
**BLIP Caption**: a baby boy is crawling in the grass
**Phi-4 Caption**: The image features a young child crawling on a grassy field, surrounded by a lively atmosphere. The child is wearing a hat and appears to be enjoying the outdoor setting.
In the background, there are several people scattered around the field, possibly watching the child or engaging in their own activities. Additionally, there are multiple kites flying in the sky, adding to the festive and playful ambiance of the scene. The kites come in various shapes and sizes, creating a colorful and dynamic backdrop for the child and the people in the field.

Sample 5

**Reference Caption**: ['A boy in the air as he does a trick on a skateboard from a ramp.', 'A person doing a skateboard trick off a concert platform.', 'A boy doing a jump off a wall with a skateboard.', 'A boy performing a skateboard jump off of a wall.', 'A young person is jumping his skateboarder off of a lodge.']
**BLIP Caption**: a boy is jumping in the air on a skateboard
**Phi-4 Caption**: The image captures a dynamic scene of a skateboarder in mid-air, performing a trick. The skateboarder is suspended above a concrete ledge, with their skateboard beneath them. The skateboarder is dressed in a plaid shirt and jeans, adding to the casual, street-style vibe of the image.
The background provides context to the location of this action. A building can be seen in the distance, suggesting that this might be an urban area or a skate park. The sky above is a clear blue, indicating that the photo was taken during the day.
The skateboarder's position and the skateboard's position relative to them suggest that they are in the middle of a trick, possibly a jump or a flip. The concrete ledge below them adds an element of danger and excitement to the scene.
Overall, this image captures a moment of thrill and skill in the sport of skateboarding.

### Qualitative Analysis on **Flickr30k** (5 Random Samples)

Sample 1

**Reference Caption**: ['A man in a white shirt is metal detecting, while two other men watch.', 'A group of guys walking down the beach using a medal detector.', 'Three men at the beach drink beer and use a metal detector.', 'The young men are having a good time at the beach.', 'Three men on a beach with a metal detector.']
**BLIP Caption**: a man holding a baseball bat
**Phi-4 Caption**: The image features a group of three men standing on a sandy beach, enjoying their time together. One of the men is holding a bottle of beer, while the other two are standing nearby. They appear to be having a good time, possibly engaging in conversation or simply relaxing on the beach.
In the background, there are a few other people scattered around the beach, some of them closer to the water and others further away. A chair can also be seen on the beach, likely for someone to sit and enjoy the view or the company of others.

Sample 2

**Reference Caption**: ['Two men wearing suits are chatting while one is holding a bottle of water.', 'Two men dressed in business casual share a laugh.', 'Two men in business attire make conversation.', 'Two men in suit talking with each other.', 'Two men in suits laugh together.']
**BLIP Caption**: two men standing on a deck
**Phi-4 Caption**: The image features two men standing next to each other outdoors, both wearing suits. They are engaged in a conversation and appear to be enjoying each other's company. One of the men is holding a cell phone, possibly showing something to the other man or taking a photo.
In the background, there is a dining table with a few chairs placed around it. A bottle can be seen on the table, and a book is also present. The scene suggests a casual outdoor gathering or event where the men are socializing and having a good time.

Sample 3

**Reference Caption**: ['Two men one wearing a red shirt, and one wearing a white shirt are standing on a boat and paddling it to shore.', 'Two men, one in a red shirt and one in a white shirt, are on a boat in the beautiful blue-green water.', 'Two people on a floating device in the water on a nice day.', 'Two men stand and row a thin boat in the ocean.', 'Two men are making their way across the ocean.']
**BLIP Caption**: a boat in the water
**Phi-4 Caption**: The image features a man standing on a boat in the middle of a body of water, possibly a lake or a river. He is holding a pole, which might be used for steering or navigating the boat. The boat is relatively small and appears to be a canoe or a small rowboat.
In the background, there are several houses situated near the water's edge, indicating that the area might be a residential community or a vacation spot. The presence of the houses and the man on the boat suggests that this could be a popular location for recreational activities or a place where people live close to the water.

Sample 4

**Reference Caption**: ['A woman rests on the curb of a city street while talking on her cellphone.', 'A woman sits on the edge of a sidewalk with a garbage bin beside her', 'A woman is sitting on a sidewalk with a cellphone at her ear.', 'A woman sits on the curb while talking on her cellphone.', 'Woman sits on the curb talking on a cellphone.']
**BLIP Caption**: a woman sitting on the sidewalk
**Phi-4 Caption**: The image features a woman sitting on the sidewalk, talking on her cell phone. She is positioned near a fruit stand, which is filled with various fruits and vegetables. The woman appears to be enjoying her time outside, possibly at a market or a street fair.
There are several other people in the scene, some of them standing near the fruit stand, while others are scattered around the area. A couple of handbags can be seen in the image, one close to the woman and another further away. The overall atmosphere of the scene suggests a casual, social environment where people are engaging in conversations and enjoying the outdoor setting.

Sample 5

**Reference Caption**: ['A boy in a red shirt, blue jeans and orange socks brushes the teeth of a plush alligator.', 'A boy sitting on a dinosaur chair holding a blue object.', 'A young boy brushes the teeth of a fake alligator.', 'A kid laying on a dinosaur brushing its teeth.', 'A child brushes a play crocodiles teeth.']
**BLIP Caption**: a child in a dinosaur chair
**Phi-4 Caption**: The image features a young boy lying on a bed, with his head inside a large inflatable dinosaur toy. The boy appears to be enjoying his time playing with the toy, as he is lying on his back with his head inside the dinosaur's mouth. The inflatable toy is quite large, taking up a significant portion of the bed.

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

R13922A22 賴澍雨

## 1. Briefly describe how you implement task 2-1 (e.g., Instruction strategy) (5% * 2)

### Instruction Strategy & Implementation Overview

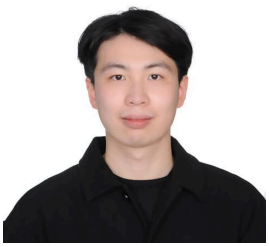
1. **Environment Setup & Data Acquisition:**
   - Install necessary packages (diffusers, accelerate, safetensors, Pillow, gdown).
   - Download 100 content images (a subset of CeleFaces) from a shared Google Drive link and unzip them.
2. **MLLM (Phi-4) for Prompt Generation:**
   - Instruction Strategy ("With Question"):
   - A detailed question is constructed to instruct Phi-4 to describe the image. The prompt asks:
   - "Describe the person's facial features, hair style, clothing, and any notable details. Be as detailed as possible for a Snoopy cartoon-style transformation."
   - This detailed inquiry ensures that Phi-4 outputs a rich, descriptive caption capturing key visual features.
3. **Prompt Refinement:**
   - The generated caption is then refined by appending a style cue (e.g., "A Person in the style of Peanuts Cartoon") to steer the final output toward the desired Snoopy look.
4. **Text-to-Image (T2I) with Stable Diffusion 3 (medium):**
   - The refined prompt is passed to the Stable Diffusion 3 model (a text-to-image pipeline) to generate a stylized image.
   - A negative prompt (e.g., "photorealistic, realistic, 3d, detailed shading, text, watermark, signature") is used to discourage photorealism and enforce a cartoon-like aesthetic.
5. **Post-Processing & Output:**
   - The generated image is resized to 224×224 pixels.
   - The final image is saved as a JPEG with the same filename as its corresponding input, inside a folder named hw1_<student_id>_stylized_images.
6. **VRAM & Memory Management:**
   - Techniques like clearing the CUDA cache (torch.cuda.empty_cache()) are employed to optimize VRAM usage during the batch processing of 100 images.

This two-stage pipeline leverages Phi-4's descriptive capabilities to extract detailed image features and Stable Diffusion 3's generative strength to create stylized outputs—all while adhering to the assignment's constraints (no training/fine-tuning, strict output format, and proper VRAM management).

## 2. Visualization on task 2-1

### (1) The style transfer on YOUR PROFILE PHOTO (5% * 2)

1. **Upload**: Use Colab's file upload to select your profile photo.
2. **Prompt Generation**: Use Phi-4 to generate a descriptive caption from the image.
3. **Refinement**: Append a Snoopy/Peanuts style cue (e.g., "in the style of Peanuts by Charles M. Schulz") to the caption.
4. **Generation**: Pass the refined prompt to Stable Diffusion to create the stylized image.
5. **Post-Processing**: Resize the image to 224×224 pixels, display it, and save it for download.



**portrait**.jpg to **portrait_stylized**.jpg

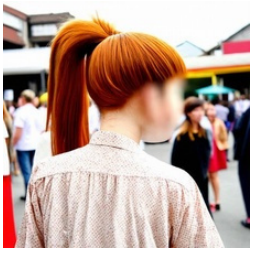
### (2) 5 success samples and 5 failure samples of CeleFaces and describe (10% * 2)

success samples

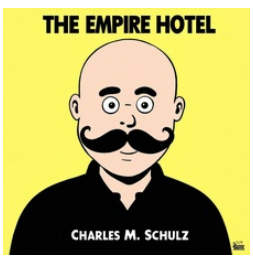

failure samples

Man turns into dog

Too realistic



Blurry




Prompt turns into graphical text

Other style of painting

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

R13922A22 賴澍雨

## 2. Visualization on task 2-1

## (3) Compare different instruction strategies (10% * 2)

### Instruction Strategy 1: With Question and substeps

- **Prompt Construction:**
  - The prompt explicitly asks a question and substeps (e.g., "**What is shown in this image?** Describe the person's facial features, hair style, clothing, and any notable details. Be as detailed as possible for a Snoopy cartoon-style transformation."). This guides Phi-4 to focus on extracting rich, descriptive information.
- **Output Quality**:
  - Because of the detailed inquiry, the generated caption tends to capture more nuanced visual features (e.g., specific facial expressions, hairstyles, clothing details) that can be leveraged to better instruct the T2I model.
- **Advantages:**
  - **Enhanced Detail**: More descriptive output helps in preserving key visual characteristics during style transfer.
  - **Better Stylization**: The enriched caption, when refined with style cues (e.g., "A Person in the style of Peanuts Cartoon"), usually results in a more convincing Snoopy-style transformation.
- **Potential Drawbacks:**
  - **Longer Prompts**: The detailed question may lead to longer prompts that require careful handling (e.g., ensuring the text isn't truncated by CLIP's token limits).

### Instruction Strategy 2: Without Question and substeps

- **Prompt Construction:**
  - The prompt does not include a direct question; it only consists of the basic instruction tokens (e.g., '<|user|><|image_1|>{prompt_suffix}{assistant_prompt}'). This leads Phi-4 to generate a more generic caption without targeted details.
- **Output Quality:**
  - The caption generated might be less descriptive, as Phi-4 isn't specifically instructed to focus on detailed attributes. This means that the subsequent refinement (e.g., appending "A Person in the style of Peanuts by Charles M.Schulz") relies on a less informative base.
- **Advantages**:
  - **Simplicity**: The prompt is shorter and simpler, which might be beneficial in contexts where brevity is preferred.
  - **Efficiency**: Fewer tokens are involved, potentially reducing complications related to token limits.
- **Potential Drawbacks:**
  - **Less Detail**: Without the guiding question, the generated caption may lack essential details needed to drive a convincing style transformation.
  - **Increased Manual Refinement**: You may need to rely more on post-processing (manual addition of style cues) to achieve the desired Snoopy-style output.

### Summary

With Question (Strategy 1) produces a richer, more detailed caption that can lead to better preservation of important visual features during style transfer. This detail is especially useful when transforming images into a highly stylized cartoon look.
Without Question (Strategy 2) yields a more generic caption, which might simplify the prompt but could require additional refinement to match the intended Snoopy style.
Ultimately, the choice between these strategies depends on whether you value detailed descriptive input (and are comfortable handling longer prompts) versus a more streamlined, simpler prompt that may be less informative.

# Task 2-2: MLLM Image Style Transfer (Image-to-Image)

R13922A22 賴澍雨

## 1. Briefly describe how you impInstruction Strategy & Implementation Overview

1. **Download & Data Preparation:**
   - **Download 100 content images:** Use gdown to download a ZIP file of the CeleFaces subset from Google Drive and unzip it into a working directory.
2. **Prompt Generation with Phi-4:**
   - **Load Phi-4 (Multimodal Instruct):** The Phi-4 model is loaded using Hugging Face's Transformers.
   - **Generate Detailed Descriptions:** A detailed question (e.g., "Describe this person's appearance for a Snoopy-style cartoon transformation.") is constructed with special prompt tokens. This instructs Phi-4 to output a descriptive caption covering facial features, hairstyle, and clothing.
   - **Refine the Prompt:** The generated caption is appended with style cues (e.g., "A Person in the style of Peanuts by Charles M. Schulz") to guide the final transformation.
3. **Image-to-Image Generation with Stable Diffusion v1.5:**
   - **Load SD v1.5 Pipeline:** The Stable Diffusion v1.5 Image-to-Image pipeline is loaded and moved to GPU with VRAM-saving optimizations (xformers, attention slicing).
   - **Disable NSFW Filtering:** The safety checker is replaced with a dummy function to bypass NSFW filtering.
   - **Long Prompt Embedding Handling:** Since the combined prompt may exceed CLIP's 77-token limit, the prompt and negative prompt are tokenized without truncation, split into 77-token chunks, encoded separately, and then concatenated. The negative prompt embeddings are padded or truncated to match the positive prompt embeddings' shape.
4. **Image Generation & Post-Processing:**
   - **Resize Input for Stable Diffusion:** Each content image is resized to 512×512 (a resolution that is stable for the SD pipeline).
   - **Generate Stylized Image:** The refined prompt (with its embeddings) is fed into the SD pipeline to produce the stylized output.
   - **Final Resizing & Saving:** The output image is resized to 224×224 pixels and saved as a JPEG with a filename corresponding to the original content image. All images are stored in a folder named according to the required format (e.g., hw1_<student_id>_stylized_images_i2i).

## 2. Visualization on task 2-2

### (1) The style transfer on YOUR PROFILE PHOTO (5% * 2)

1. **Image Upload:** The user uploads a portrait, which is loaded and converted to RGB.
2. **Prompt Generation:**
   - Phi-4 generates a detailed description by asking about facial features for a Snoopy-style transformation.
   - The caption is refined by appending a style cue (e.g., "A Person in the style of Peanuts by Charles M. Schulz").
3. **Embedding & Generation:** Custom tokenization splits long prompts into chunks to bypass CLIP's 77-token limit. The Stable Diffusion v1.5 Image-to-Image pipeline uses these embeddings (with NSFW filtering disabled) to transform the portrait.
4. **Post-Processing:** The resulting image is resized to 224×224 pixels, saved, displayed, and made available for download.


**portrait**.jpg to **portrait_stylized**.jpg

### (2) 5 success samples and 5 failure samples of CeleFaces and describe (10% * 2)

success samples





failure samples





**No mouth, multi eyes** **No nose** **Crooked facial features** **Crooked facial features** **Man turns into dog**

# Task 2-2: MLLM Image Style Transfer (Image-to-Image)

R13922A22 賴澍雨

## 2. Visualization on task 2-2

## (3) Compare different instruction strategies (10% * 2)

**1. Original Strategy (Baseline)**

- **Prompt Generation**: Uses Phi-4 to produce a descriptive caption. The output text prompt is fed directly into the T2I model without extensive manual modifications.
- **Outcome:**
  - The transformation relies solely on the model's interpretation of the generated caption.
  - May yield less consistent stylization if the Phi-4 caption is too generic or lacks strong style cues.

**2. Long Prompt Strategy (strength=0.75, guidance_scale=9.5)**

- **Prompt Details:**
  - **Detailed Style Cue:** The caption from Phi-4 is refined with extended style instructions, e.g.: "{raw_prompt}.A Person in Peanuts style, minimal lines, 2D line art, bright simple colors, in the style of Charles M.Schulz."
- **Rich Detail:** More descriptive prompts aim to preserve nuanced facial features and subtle details.
- **Outcome**: Produces a image where the Snoopy/Peanuts style is evident yet the original background context features are largely maintained, however, the cartoon style generation is unstable and sometime turns to photorealistic image.

**3. Higher Strength Strategy (strength=0.95, guidance_scale=8.5)**

- **Prompt Details**:
  - **More Style Cue**: The refined prompt is shorter, for example: "{raw_prompt}.A Person in the style of Peanuts by Charles M.Schulz", while the style transfer strenght is higher.
- **Outcome**:
  - Results in a more aggressive style transfer where the cartoonish, Peanuts-style features dominate, though at the potential cost of losing finer details of the original image.