

Task1

1. Briefly describe how you implement the two models

ANS : BLIP is a fast visual language model that generates descriptive text by simply typing in an image with inputs = processor(image), outputs = model.generate, and caption = processor.decode. Phi-4 is a multimodal large-scale language model that understands complex commands. Phi-4 is a multimodal large language model that understands complex commands, with inputs = processor(picture) and results from model.generate and processor.decode.

2. Experiment table of (2 models) X (2 datasets), for example

MSCOCO-Test <u>Run all the data.</u>				
	BLEU	ROUGE-1	ROUGE-2	METEOR
BLIP	0.255	0.568	0.335	0.421
Phi-4	0.006	0.157	0.029	0.084

flickr30k				
	BLEU	ROUGE-1	ROUGE-2	METEOR
BLIP(Run all the data.)	0.161	0.478	0.253	0.323
Phi-4(Run 1000 data.)	0.013	0.223	0.048	0.114

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

ANS : BLIP is a model specifically trained for image description tasks with a consistent output format that produces phrases similar to standardized answers, and thus performs better on n-gram comparison scores such as BLEU, ROUGE, and METEOR. In contrast, Phi-4, despite its strong multimodal comprehension ability, is not specifically trained for image annotation, and its output is more subjective and narrative, with lower consistency with standard answers, hence lower scores.

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

ANS :

Description of image marking : ['A man with a red helmet on a small moped on a dirt road. ', 'Man riding a motor bike on a dirt road on the countryside.', 'A man riding on the back of a motorcycle.', 'A dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains. ', 'A man in a red shirt and a red hat is on a motorcycle on a hill side. ']

BLIP : 'a man riding a motorcycle down a dirt road.'

Phi4 : 'The person in front of the metal gate is a female. She has curly brown hair and is dressed in a turquoise shirt and dark pants. Her skin tone appears to be light. '

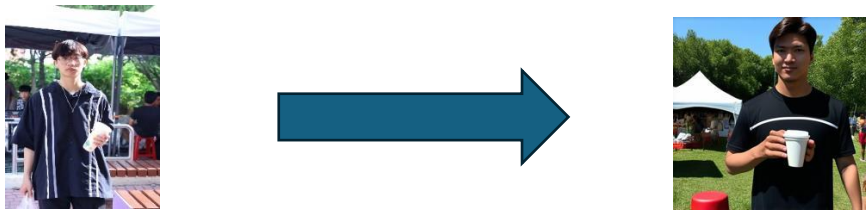
Task 2-1

1. Briefly describe how you implement task 2-1

ANS : (1) First, we import 100 face images from the CelebFaces dataset into Phi-4 and prompt it with the following commands (2) Phi-4 outputs creative, Snoopy-style text descriptions. These text descriptions are then sent as prompts to Stable Diffusion 3 Medium (a powerful text-to-image model) to produce stylized cartoon portraits.(3) The resulting image is finally resized to 224×224 to meet the output requirements. This method utilizes command-following multimodal understanding (Phi-4) and text-directed generative modeling (SD3) for end-to-end cartoon styling.

2. Visualization on task

ANS : (1)The style transfer on YOUR PROFILE PHOTO






(2) 5 success samples and 5 failure samples of CeleFaces and describe

First Transfer prompt : "Please carefully observe the person in the input image and provide a detailed description of their facial features (e.g., face shape, hairstyle, eyes, eyebrows, nose, mouth, expression, and clothing style). When describing, adopt a Snoopy style—that is, use concise, playful, cartoonish, and humorous language while retaining the unique characteristics of the person. The final description should be similar to:Depict this person in a Snoopy style: with a soft, rounded face, lively and playful eyes and eyebrows, minimalist lines outlining the facial features, and a childlike expression that exudes a relaxed and cheerful vibe"

Second Transfer prompt : "A simplified black and white cartoon drawing of a person, with large head and expressive eyes, in Peanuts comic strip style, Snoopy aesthetic."

SUCCESS SAMPLE :

Original	First Prompt	Second Prompt
		



(3) Compare different instruction strategies

By making changes to the prompts provided by Phi-4, such as explicit style prompts, contextualization, enhanced voice, etc., the proper design of the prompts can make the results of the multimodal model + diffusion model more closely match the expected style.

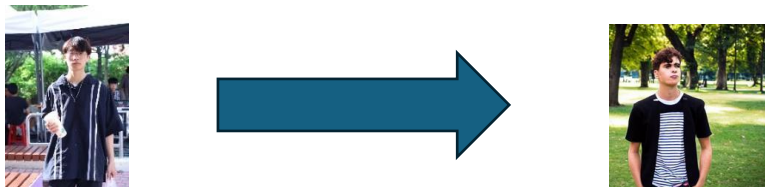
Task 2-2

1. Briefly describe how you implement task 2-2

ANS : (1)Input images (2)Prompt generation (3) Style prompt formatting (4) Styling with SD1.5 (5) Output saving




2. Visualization on task


ANS : (1)The style transfer on YOUR PROFILE PHOTO











(2)5 success samples and 5 failure samples of CeleFaces and describe

SUCCESS SAMPLE :

Original	First Prompt	Second Prompt
		

FAILURE SAMPLE :

Original	First Prompt	Second Prompt
		
		
		
		



(3) Compare different instruction strategies :

Proper prompt engineering plays a key role in multimodal generation. Adding emotion, tone, or cultural references can help models like Phi-4 generate richer captions that better guide the stylization process. In my experiment, I tried a two-stage prompting strategy:

First transfer prompt : A detailed instruction asking for a Snoopy-style description of a person's facial features, using playful and cartoonish language.

Second transfer prompt : A simplified visual prompt: "A black and white cartoon drawing of a person, with large head and expressive eyes, in Peanuts comic strip style, Snoopy aesthetic."

Despite the intention, both stages failed to produce satisfying results. The captions were often too generic or verbose, and the generated images lacked the intended Snoopy/Peanuts cartoon feel. This suggests a mismatch between prompt intent and model training: while Phi-4 may understand stylistic tone, diffusion models like SD3 or SD v1.5 may not recognize abstract cultural references like "Snoopy style."

Moreover, concepts like "playful" or "humorous" are hard to map visually without more concrete visual terms or reference images. In future work, using clear visual descriptors (e.g., "bold outlines, circular eyes, minimal shading") or example-based conditioning could help guide the stylization more effectively.