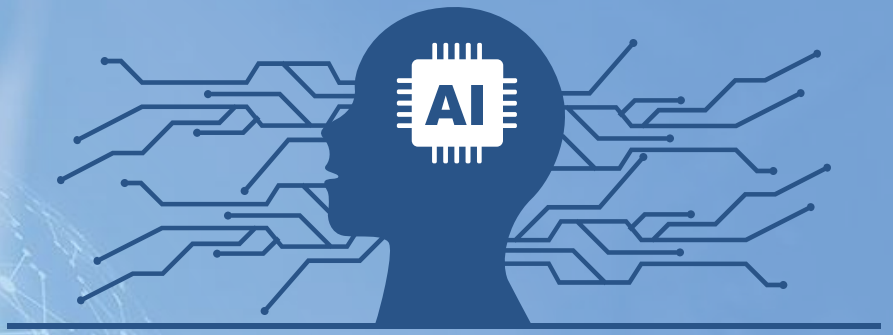
AI: Artificial Intelligence



Homework #1
Basic MLLM Implementation

Wen-Huang Cheng (鄭文皇)

National Taiwan University

wenhuang@csie.ntu.edu.tw

HW1 - Overview

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

Task 1: Image Captioning Evaluation

What is Image Captioning?

A computer screen with a Windows message about Microsoft license terms.

A can of green beans is sitting on a counter in a kitchen.

A photo taken from a residential street in front of some homes with a stormy sky above.

A blue sky with fluffy clouds, taken from a car while driving on the highway.

A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.

A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.

A cup holder in a car holding loose change from Canada.

Task 1: Evaluation Details

- Models (Restricted): BLIP (link), Phi-4 (link)
- Datasets (Restricted): MSCOCO-Test (5k) (link), flickr30k (link)
- Metrics (intro, implementation): BLEU, ROUGE-1, ROUGE-2, METEOR
- Failure to follow the above model, dataset, and metrics will result in a deduction of 10% for each error.

Task 1: Report (20%)

1. Briefly describe how you implement the two models (5%)
2. Experiment table of (2 models) X (2 datasets), for example: (5%)

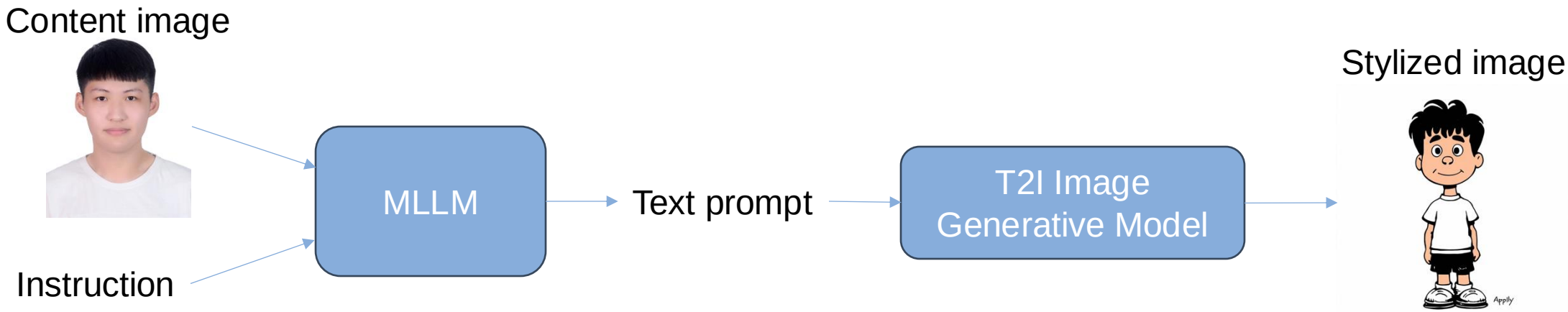
	MSCOCO-Test				flickr30k			
	BLEU	ROUGE-1	ROUGE-2	METEOR	BLEU	ROUGE-1	ROUGE-2	METEOR
BLIP								
Phi-4								

3. Analysis: describe what is observed from the table and what causes the difference in metric between the two models. (5%)
4. Case study: qualitative analysis of interesting samples in both models. (5%)

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

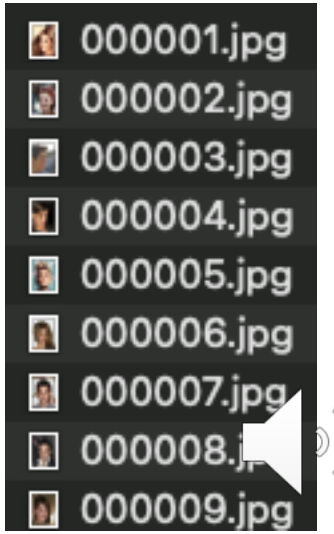
- Style: Snoopy

- Pipeline:



Task 2-1: Implementation Details

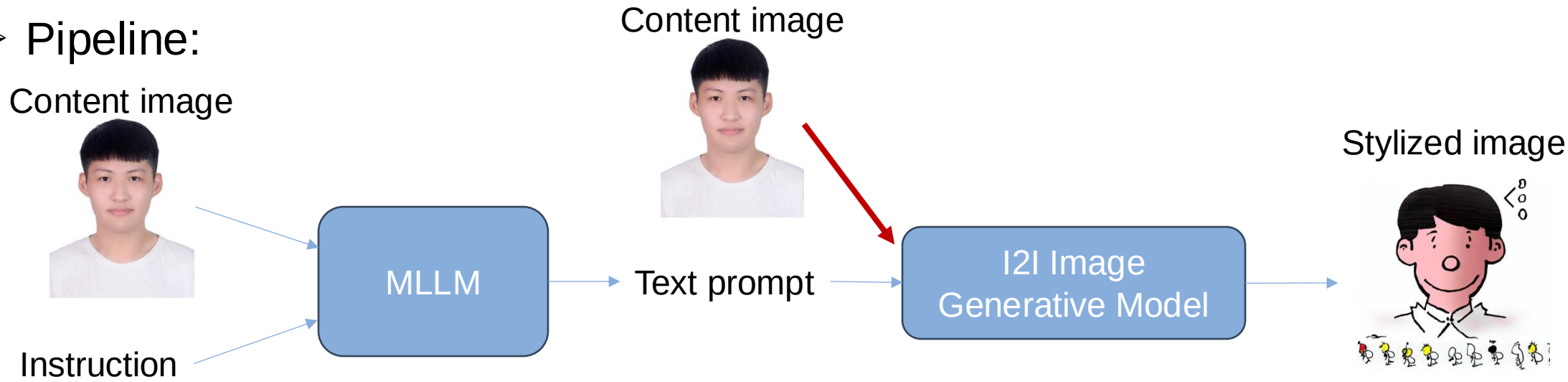
- Style: Snoopy
- Models:
  - MLLM: Phi-4 (link) (Restricted)
  - T2I Image Generative Model: stable-diffusion-3-medium-diffusers(link) (Restricted)
- Input content images : a subset of CeleFaces (100 images) (link)
- Output: 100 stylized images (224 X 224)
- **DO NOT train/fine-tune the model or use additional models**
- How to Lower the VRAM usage



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

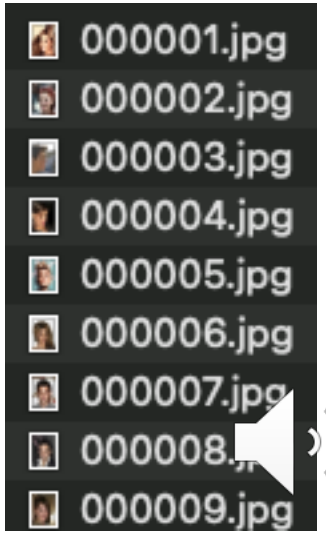
- Style: Snoopy

- Pipeline:



Task 2-2: Implementation Details

- Style: Snoopy
- Models:
  - MLLM: Phi-4 (link) (Restricted)
  - T2I Image Generative Model: stable-diffusion-v1-5 (link) (Restricted)
- Input content images: a subset of CeleFaces (100 images) (link)
- Output: 100 stylized images (224 X 224)
- **DO NOT train/fine-tune the model and use additional models**
- How to Lower the VRAM usage



Task 2: Report (60%)

1. Briefly describe how you implement task 2-1&2-2 (e.g., Instruction strategy) (5% * 2)
2. Visualization on task 2-1&2-2
    1. The style transfer on YOUR PROFILE PHOTO (5% * 2)
    2. 5 success samples and 5 failure samples of CeleFaces and describe (10% * 2)
    3. Compare different instruction strategies (10% * 2)

Task 2: Competition (20%)

- Submit the output stylized images of “Task 2-1” following the format (next page)
- We will use this repo to calculate the ArtFID [1] of the stylized images generated by each person in “Task 2-1”, and rank the scores of the whole class to grade.
- Compute ArtFID on your own (if you want): Download style images from link
- Grading method: Linear grading from 1%-20%

Your output

CeleFaces

ArtFID is computed as (ArtFID = (1 + LPIPS) · (1 + FID)). LPIPS measures content fidelity between the stylized image and the corresponding content image, and FID assesses the style fidelity between the stylized image and the corresponding style image.

[1] Jiwoo Chung, et al., "Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer", CVPR, 2024.

Stylized Images Submission Format

- All the stylized images of Task 2-1 should be resized to 224 X 224 (You don't need to generate 224 X 224 directly, just do resize at the end)
- The filenames should correspond to the content images, e.g., 000001.jpg
- DO NOT include ANY other images or files except for the 100 generated images
- Folder name and structure:

```
hw1_<student_id>_stylized_images
|-- 000001.jpg
|-- 000002.jpg
```

- Zip the folder to hw1_<student_id>_stylized_images.zip
- Violation of the format will result in **0% score for the Task 2 competition**

Submission Rules

- Deadline
  - 2025/03/28 (Fri.) 23:59
- Upload filename and format
  - hw1_<student-id>.zip (e.g. hw1_D12345678.zip)
- Submit to NTU cool

Submission Rules

- Your submission should be a zipped file with the following structure:
  - hw1_<student-id>.zip

    |-- hw1_<student-id> (Should contain this folder, not separate files)

    |-------- hw1_<student-id>.pdf (Your report, including Task 1 / 2-1 / 2-2) (**4-6 pages**)

    |-------- hw1_<student-id>_stylized_images.zip (Your output images of Task 2-1)

    |-------- hw1_<student-id>_code.zip (All tasks, randomly select 10% of the people to re-implement)

    |-------- README.md

    - Your environment details
    - How to run your code

- Incorrect format or exceeding page limitation will result in a deduction of **5%.**
- Failure of re-implementing similar performance will result in **0%.**
- Plagiarism in the report or code will result in **0%.**


Any
Question
ai.ta.2025.spring@gmail.com