CVPDL HW3 Report

鄭恩庭 R12323031

December 15, 2024

1 Selected Papers and Methodologies Used

The relevant papers are BLIP-2 for language-image pre-training and GLIGEN for grounded text-to-image generation. The methodologies used from BLIP-2 include the use of frozen image encoders in combination with large language models to perform efficient image captioning and visual question answering. From GLIGEN, the assignment adopts the approach of grounded generation, where text prompts or object annotations are used to guide and constrain the image synthesis process.

2 Implementation Steps

Image-to-Text

In this stage, I utilized the `Salesforce/blip2-opt-6.7b-coco` and `Salesforce/blip2-opt-6.7b` models to generate captions. For the inputs to these models, the original 2160 images were used without any resizing.
The following prompt was employed for caption generation:

"Please describe the photo in detail."

I found that the captions generated by `Salesforce/blip2-opt-6.7b-coco` are more reasonable.

Text-to-Image

In this stage, I used the GLIGEN model for grounded text-to-image generation. Two prompt templates were designed for this task. Below is an example showcasing these templates:
Generated Text:

"A man is walking down the street with a backpack and a vacuum cleaner."

Template 1:

"This is an image in the Occupational Injury Prevention dataset. A man is walking down the street with a backpack and a vacuum cleaner. There are Person, Face, Glasses, Head, Hands, Tools, Ear, and Shoes in the image."

Template 2:

"This is an image in the Occupational Injury Prevention dataset. A man is walking down the street with a backpack and a vacuum cleaner. There are Person, Face, Glasses, Head, Hands, Tools, Ear, and Shoes in the image. Height: 512, Width: 512, HD quality, highly detailed."

After generating the result images, I resized the original images to 512 × 512 and computed the Fréchet Inception Distance (FID) between the resized original images and the generated images.

	Text grounding		**Layout-to-Image**
prompt	Template #1	Template #2	Template #2
FID	65.4168	63.4713	42.3569

Table 1: Experimental Result