# CVPDL-HW3

## 1. Image Captioning

- (a) Compare the performance of 2 selected different pre-trained models in generating captions, and use the one you find the most effective for later problems.

  我選擇 **Salesforce/blip2-opt-6.7b-coco** 和 **Salesforce/blip2-opt-6.7b** 這兩個模型來比較，並且選了 6 張圖片來作為模型表現的參考。

| Model | Image1 | Image2 | Image3 | Image4 | Image5 | Image6 |
|---|---|---|---|---|---|---|
| **Salesforce/blip2-opt-6.7b-coco** | a large group of fish swimming in a large tank | a group of jellyfish swimming in a blue ocean with a blue sky | two fish swimming in an aquarium with coral and rocks | a penguin swimming in an aquarium with rocks and a waterfall | a group of starfish are swimming in an aquarium next to a rock | a shark swimming in the water near some rocks and plants |
| **Salesforce/blip2-opt-6.7b** | a large group of fish swimming in a blue ocean | jellyfish in a tank with blue water | a fish in an aquarium with other fish | a penguin swimming in an aquarium | a fish tank with a starfish and a sea anemone | a shark swimming in the water near some rocks |

從上面的結果可以看出，**Salesforce/blip2-opt-6.7b-coco** 對於圖片的描述比較詳細，而 **Salesforce/blip2-opt-6.7b** 對於圖片的描述比較簡單，不過兩個模型都有描述到圖片中的物件，且描述的

內容大部分都是正確的。而因為後續是要生成圖片來訓練物件偵測模型，因此我認為如果使用詳細的描述，會造成容易生成相似的圖片，可能對於物件偵測模型不會提升太大的效果，因此我選擇描述較簡單的 **Salesforce/blip2-opt-6.7b** 這個模型來做後續的實驗，希望較簡單的文字敘述能讓後續模型生成的圖片比較有多樣性。

- (b) Design 2 templates of prompts for later generating comparison.
    - Template #1: `blip2_generated_text`, height: `img_height`, width: `img_width`, underwater background, real word, high quality, 8K Ultra HD, high detailed, Composition: shot with a Canon EOS-1D X Mark III, 50mm lens
    - Template #2: `blip2_generated_text`, `categroy`, height: `img_height`, width: `img_width`, ocean, undersea background, HD quality, high detailed

`blip2_generated_text` 代表模型生成出來的字，`img_height` 和 `img_width` 代表圖片的高和寬，`categroy` 代表圖片的類別。

而因為根據助教所說 BLIP-2 不認識 puffin 這個category，因此我手動調整 puffin 類別的 prompt，主要是將 BLIP-2 生成的文字中描述類別的地方改成 puffin，讓它能夠生成 puffin 的圖片。

## 2. Text-to-Image Generation

- (a) Use 2 kinds of generated prompts from Problem 1(b) to generate images. (text only!)

| Template #1 | Template #2 |
| --- | --- |
|  |  |

- (b) Select the prompts for better-generating results, and perform image grounding generation. (text + image)

| Template #1 + Image |
| --- |
|  |

## 3. Table of your performance based on FID

| | **Text grounding** | **Text grounding** | **Image grounding** |
|---|---|---|---|
| **prompt** | Template #1 | Template #2 | Template #1 |
| **FID** | 143.27 | 147.44 | 135.59 |

## 4. Table of the improvement of your detection model from HW1 after data augmentation

| | **Before Data Augmentation** | **After Data Augmentation (Text grounding)** | **After Data Augmentation (Image grounding)** |
|---|---|---|---|
| **$AP_{[50:5:95]}$** | 0.582 | 0.586 | 0.584 |

### **Detailed settings of experiments**:

- `checkpoint`：跟 hw1 一樣，使用 DINO 本身提供的 pretrained model checkpoint0029_4scale_swin.pth 再訓練 12 個週期，其餘參數也都跟 hw1 一樣。
- `data augmentation`：使用 hw1_dataset 所有的照片，再加上 140 張生成的照片（每個類別 20 張）。
- **雖然沒有實際去平衡每個類別的照片數量，但就結果來說表現還是有提升**。而沒有實際去平衡每個類別的數量原因如下：
    1. 只用 140 張的生成照片，表現是有提升的。
    2. 我認為應該要依照 bounding box label 數量去做平衡，因此有嘗試依照這個想法去生成照片，但換算下來，需要生成大約 7000 多張影像，而又因為生成影像時間過久，不好意思長時間占用實驗室資源，因此就沒有進行實驗。
    3. 若真的去平衡每個類別的照片，可能會使少數類別的影像大部分皆為生成出來的影像，這樣可能會使模型在少數類別上有偏見，變成在辨識生成的影像，感覺生成出來的影像（data augmentation）不應該多於原本的影像。
    4. 若將原本數量較多的類別影像減少，可能會使模型在原本數量較多的類別上表現變差，而且越多該類別的影像，應該會使該類別學得比較好，我認為沒必要減少。
    5. 我有嘗試只加入少數類別的照片，在 penguin 、puffin、starfish、jellyfish 這四個類別上，加入 20 張生成的照片，結果 Text grounding 表現變差，Image grounding 的表現差不多。

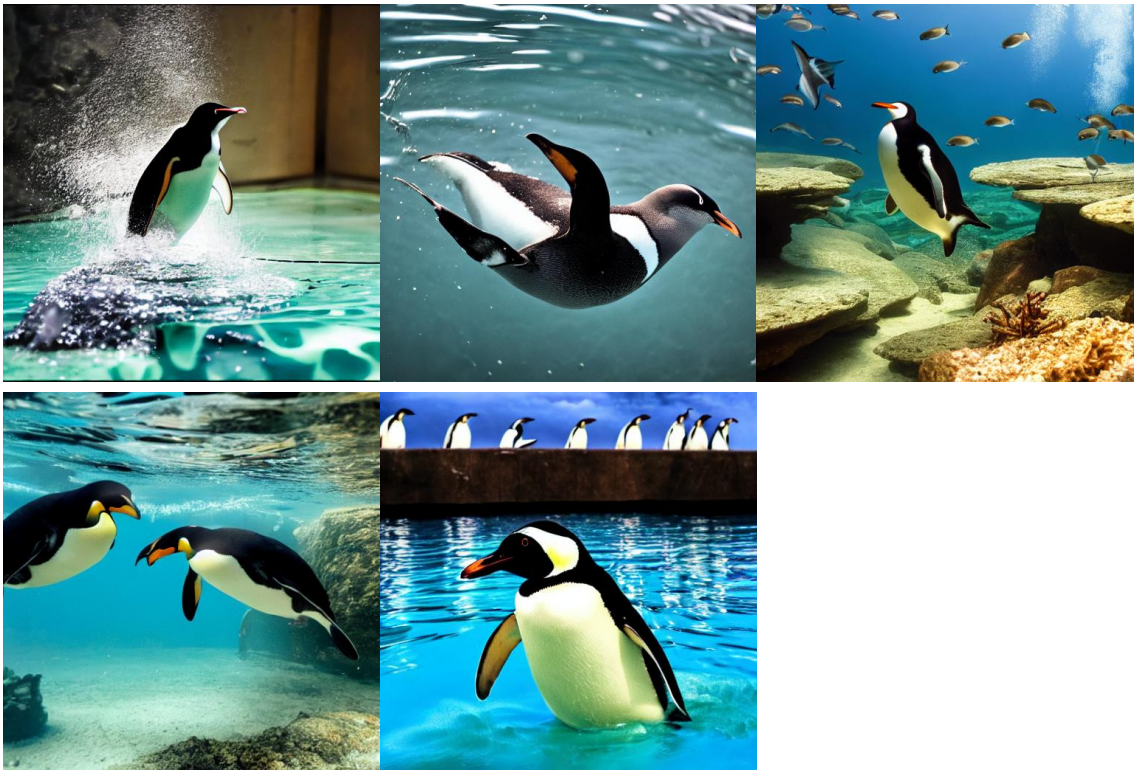
## 5. Visualization

> show the best 5 images for each category (35 images in total!)

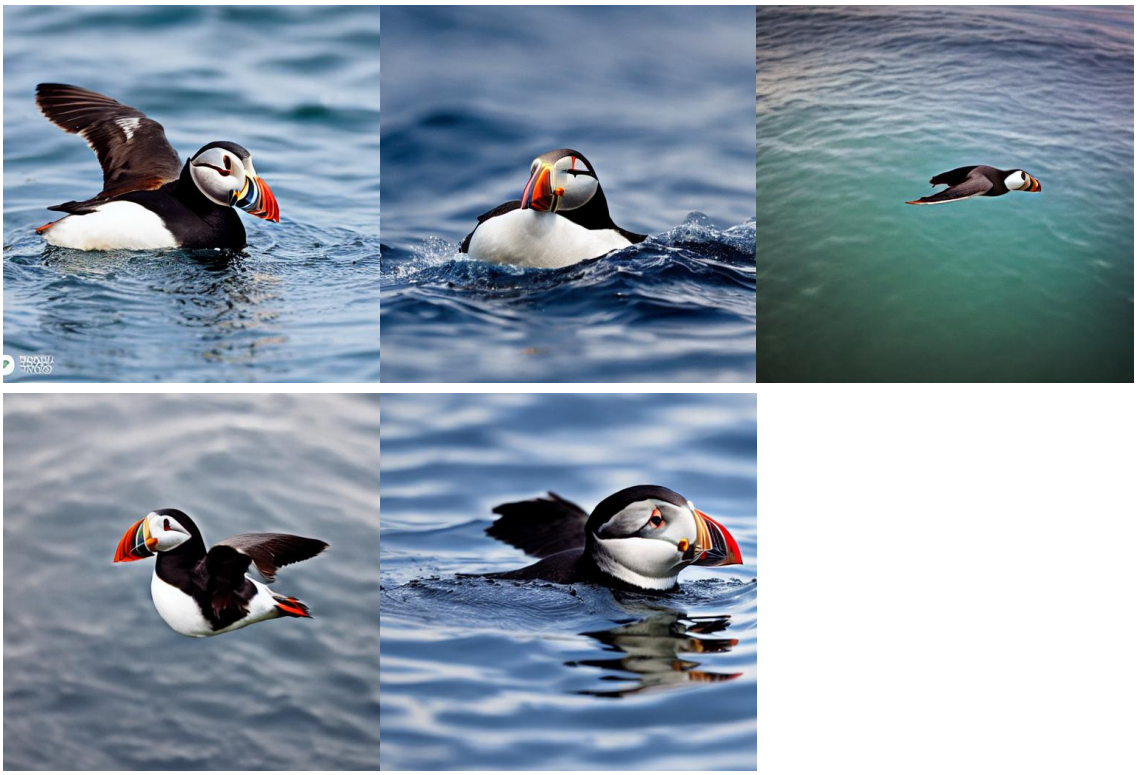
- Fish

- Jellyfish

- Penguin



- Puffin



- Shark

- Starfish

- Stingray